DISCUSSION PAPER SERIES

IZA DP No. 8455

Testing for Selection Bias

Joonhwi Joo
Robert LaLonde

September 2014

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Testing for Selection Bias

Joonhwi Joo
University of Chicago

Robert LaLonde
University of Chicago
and IZA

Discussion Paper No. 8455
September 2014

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 8455
September 2014

ABSTRACT

Testing for Selection Bias*

This paper uses the control function to develop a framework for testing for selection bias. The idea behind our framework is if the usual assumptions hold for matching or IV estimators, the control function identifies the presence and magnitude of potential selection bias. Averaging this correction term with respect to appropriate weights yields the degree of selection bias for alternative treatment effects of interest. One advantage of our framework is that it motivates when is appropriate to use more efficient estimators of treatment effects, such as those based on least squares or matching. Another advantage of our approach is that it provides an estimate of the magnitude of the selection bias. We also show how this estimate can help when trying to infer program impacts for program participants not covered by LATE estimates.

JEL Classification: C21, C26, D04

Keywords: selection bias, program evaluation, average treatment effects

Corresponding author:

Joonhwi Joo
Department of Economics
The University of Chicago
1126 East 59th Street
Chicago, Illinois 60637
USA
E-mail: joonhwi@uchicago.edu

* We would like to thank Dan Black and participants at the Federal Reserve Bank of Chicago's Brown Bag Seminar for helpful comments and suggestions. All errors are ours.

1 Introduction

Endogenous selection can make it difficult to identify the impact of a program or intervention. When such selection is present, simply using least squares, matching or even instrumental variables estimators to estimate impacts such as the average effect of the treatment on the treated (ATT) or the average treatment effect (ATE) can lead to biased estimates of these parameters. By contrast, when the underlying model is correctly specified, the control function estimator can account for endogenous selection, and thus yield consistent estimates of alternative treatment effect parameters. But the drawback from this approach is that usually control function estimators are less efficient. Therefore, the question is when can we use other, more efficient estimators (e.g., least squares or propensity score matching) to estimate treatment parameters such as ATE or ATT?

To address the foregoing question, we develop statistical tests for the presence of selection bias. It is often perceived that the magnitude of selection bias in program evaluations is not identified. For this reason, many researchers who estimate ATT or ATE assume away the potential selection biases and proceeded with estimators such as OLS, matching or IV. However, under a certain set of assumptions, it is possible to identify the magnitude of selection biases embedded in the various treatment effects. The control function, which dates back to the seminal work of Heckman (1976, 1979), is the correction term for the selection bias. It therefore plays a critical role in identifying selection bias. Averaging this correction term with respect to appropriate weights yields the degree of selection bias for alternative treatment effects of interest.

Accordingly, in this paper, we use the control function to develop a method to identify and test for the magnitude of selection bias in program evaluations. To be specific, we identify the existence and magnitudes of selection bias by using the control function, where we use alternative methods to consistently estimate the control function. The method we develop is semiparametric, which makes our work different from Olsen (1980); Melino (1982); Lee (1982). For example, we use the two-stage propensity score polynomial estimator to estimate semiparametrically the control function. In our Monte Carlo analysis, we show that our test has a reasonable power with a sample size as small as 1,000 observations. Further, this result is consistent, even when there is specification bias in the econometric model, in addition to selection bias.

The rest of the paper is organized as follows. Section 2 defines the notation used in the paper and summarizes and compares the different assumptions required for identifying the average treatment effects of interest. Section 3 develops a framework for identifying and testing the presence of the selection bias. Section 4 conducts a Monte-Carlo simulation to evaluate the performance of our test. In Section 5, we present an empirical example of our test. The example uses data for the adult women who participated in the classroom training (CT) component of the National JTPA Study (NJS). Section 6 concludes.

2 Self-Selection and Conditional Independence

2.1 Notation Used in the Paper

We begin by defining the terms and notation that we use in this paper. We use the term *assignment* only for the exogenous assignment into a treatment. Random assignment is an example. To indicate

actual participation, we use the term *participation*. Let $D_i \in \{0,1\}$ be a participation indicator, not an assignment indicator. Let X be the set of covariates, Z be the set of of instruments. Subscript i always indicates the random variable/vector of i-th observation.

We express the outcome of interest as follows: $y_i = y_{1,i}D_i + y_{0,i}(1 - D_i)$, where $D_i = 1$ denotes the treatment group of participants, and $D_i = 0$, the comparison group of non-participants. The subscripts $0, 1$ on y_i denote the two statuses of the binary treatment. We also assume additive separability between the observed determinants of the outcome y_i, and the unobserved variable as follows:

$$y_i = f(\mathbf{x}_i, D_i) + u_i \tag{2.1}$$

where $D_i \in \{0,1\}$, $\mathbf{x}_i (\in X)$ is a set of covariates.

Throughout the paper, we focus on the following three commonly estimated parameters of interest discussed in the program evaluation literature:

$$\begin{aligned} \Delta_{ATE} &= E[y_{1,i} - y_{0,i}] \\ \Delta_{ATT} &= E[y_{1,i} - y_{0,i}|D_i = 1] \\ \Delta_{ATN} &= E[y_{1,i} - y_{0,i}|D_i = 0] \end{aligned}$$

We employ a version of the assumptions usually invoked in matching and control function methods. These are not the most general assumptions, but we contend that they work well for most applications.

2.2 Conditional Independence Assumptions

The ability to estimate the models which allows only for the selection-on-observables, such as matching estimators, depends on following two assumptions:

(CIA-1) (Common Support) For all $\mathbf{x}_i \in X$, $0 < \Pr(D_i = 1|\mathbf{x}_i) < 1$.

The common support assumption states one must have the domain of the covariate set $\mathbf{x}_i$ be the same for $D_i = 1$ and $D_i = 0$. This assumption can be easily tested in practice.

The other assumptions are a set of Conditional Independence Assumptions (CIA) that must be invoked. Each of the average treatment effect parameters requires a modestly different version of the CIA.

(CIA-2) (Conditional Independence) $\forall \mathbf{x}_i \in X$, $\forall D_i \in \{0,1\}$,

(i) $CIA_{ATE}: (y_{1,i}, y_{0,i}) \perp\!\!\!\perp D_i|\mathbf{x}_i$

(ii) $CIA_{ATT}: y_{0,i} \perp\!\!\!\perp D_i|\mathbf{x}_i$

(iii) $CIA_{ATN}: y_{i,i} \perp\!\!\!\perp D_i|\mathbf{x}_i$

For consistency of matching estimators, (CIA-2) can be replaced by the (conditional) mean independence assumptions, which is a weaker condition.[1] Formally, the corresponding mean independence assumptions are as below:

[1]For example, the mean independence assumption allows the conditional variance $Var[y_{1,i}|\mathbf{x}_i, D_i]$ to vary across $D_i = \{0,1\}$.

(CIA-2') (Mean Independence) $\forall \mathbf{x}_i \in X$, $\forall D_i \in \{0,1\}$,

(i) MIA_{ATE} : $E[y_{0,i}|\mathbf{x}_i] = E[y_{0,i}|\mathbf{x}_i, D_i]$ and $E[y_{1,i}|\mathbf{x}_i] = E[y_{1,i}|\mathbf{x}_i, D_i]$

(ii) MIA_{ATT}: $E[y_{0,i}|\mathbf{x}_i] = E[y_{0,i}|\mathbf{x}_i, D_i]$

(iii) MIA_{ATN} : $E[y_{1,i}|\mathbf{x}_i] = E[y_{1,i}|\mathbf{x}_i, D_i]$

As Imbens and Wooldridge (2009) summarize in their survey, the appropriateness of these assumptions is questionable in many circumstances. They further argue that the CIA-2 is not testable. However, LaLonde (1986); Heckman and Hotz (1989); Heckman et al. (1998); Rosenbaum (1987) have suggested approaches that under some circumstances that can test whether the CIA holds. The former sets of studies test the the validity of alternative nonexperimental methods by assuming the availability of the corresponding experimental data. When such data are available, estimates using non-experimental methods can be compared directly to experimental estimates. The test of the CIA comes from comparing the comparison group in the non-experimental data to the control group in the experimental data. In the later study, the idea is that it maybe plausible to assume the data on counterfactuals are available.

These assumptions for matching or the least squares estimators imply that conditional on the vector of characteristics, participation is random. This strong condition does not allow for the endogenous selection into participation. In other words, the matching and least squares estimators allow only for selection on observed variables. Under these approaches any noncompliance or sample attrition should be random. Further, some of the matching estimators are consistent and asymptotically normal, whereas others that rely on different matching methods may not be (c.f., See, e.g., Hahn, 1998; Abadie and Imbens, 2006).

For the linear IV estimator, the CIA assumes either full compliance or random attrition from the program. Thus, if CIA is satisfied, a linear IV estimator can estimate the ATE, using, for example, random assignment status as an instrument.[2] In fact, provided that treatment participation is completely random, with perfect compliance or random attrition, even the OLS estimator for δ (the coefficient associated with the participation indicator D_i) is consistent. The simple OLS estimate is more efficient and easier to implement than are other estimators. This point has been analyzed extensively in Heckman and Vytlacil (2005), where they discuss using the marginal treatment effect as a building block for all other treatment effects.

Our discussion so far brings an attention to the tradeoff of invoking CIA: An econometrician can either take the risk of the estimator being biased for the sake of efficiency, or sacrifice efficiency by using a control function estimator that explicitly takes into account program selection.

2.3 The Vytlacil-Imbens-Angrist Assumptions

When the CIA does not hold, estimators based on CIA are no longer consistent, because of selection bias. IV estimation of the Local Average Treatment Effect (LATE) proposed by Imbens and Angrist (1994) gets around the problem of self-selection, by identifying the average treatment effect for a particular subpopulation of the participants. It is the treatment effect for those participants induced to participate in the program by changes in the instrument. Importantly for our purposes, Vytlacil (2002); Heckman and Vytlacil (2005) demonstrate that the assumptions imposed by Imbens and Angrist for identifying LATE are identical to the assumptions required for the semiparametric Control Function estimator.

[2]For a comprehensive treatment on this subject in the perspective of marginal treatment effect, see Heckman and Vytlacil (2005).

Now we consider a general selection model:

$$y_{1,i} = f_1(\mathbf{x}_i) + u_{1,i} \tag{2.2}$$

$$y_{0,i} = f_0(\mathbf{x}_i) + u_{0,i} \tag{2.3}$$

$$D_i^* = g(\mathbf{x}_i, \mathbf{z}_i) + u_{D,i} \tag{2.4}$$

The Vytlacil-Imbens-Angrist assumptions (VIA) for this model are as follows:

(VIA-1) (Common Support) For all $\mathbf{x}_i \in X$, $0 < \Pr(D_i = 1|\mathbf{x}_i) < 1$.

(VIA-2) (Existence of Instruments) $\forall \mathbf{x}_i \in X$, $\mathbf{z}_i$ is a random vector such that $p(\mathbf{z}_i = \mathbf{w}, \mathbf{x}_i) := E[D_i|\mathbf{x}_i, \mathbf{z}_i = \mathbf{w}]$ is a nontrivial function of $\mathbf{w}$.

(VIA-3) (Conditional Independence of Instruments) $(u_{1,i}, u_{D,i}) \perp\!\!\!\perp \mathbf{z}_i|\mathbf{x}_i$ and $(u_{0,i}, u_{D,i}) \perp\!\!\!\perp \mathbf{z}_i|\mathbf{x}_i$.

(VIA-4) (Monotonicity) For all $\mathbf{z}, \mathbf{w} \in Z$, $\forall i$, either $D_i(\mathbf{z}) \geq D_i(\mathbf{w})$ or $D_i(\mathbf{z}) \leq D_i(\mathbf{w})$.

Under the foregoing assumptions, the linear IV estimator identifies the LATE. For our purposes it is important that IV estimation of LATE allows for the heterogeneity in treatment effects, in the sense that it allows for $Cov(D_i, u_i) \neq 0$. One interpretation of the nonzero covariance is that agents observe their own treatment effect and act on it when making their participation decisions.

By virtue of the equivalence result of Vytlacil (2002); Heckman and Vytlacil (2005), it is possible to semiparametrically estimate the parameters of the above model (2.2) to (2.4) under (VIA-1) to (VIA-4)[3] by using $\mathbf{z}_i$ as an instrument for D_i on the selection equation:

$$\Pr(D_i = 1|\mathbf{x}_i, \mathbf{z}_i) = \Pr(D_i^* > 0|\mathbf{x}_i, \mathbf{z}_i)$$

which is the propensity score. Several methods have been suggested for the semiparametric estimation of this model and the asymptotic distributions have been suggested. The parameter estimates of the semiparametric control function method are known to be consistent, and some of them are asymptotically normal.[4]

The underlying assumptions and different parameter estimates of matching and control function approach is presented in, for example, Heckman and Navarro-Lozano (2004). They clarify the sources and the forms of potential biases. Moreover, they find that the matching estimators can be considered as a special case of control function methods under a relatively weak assumption, namely, the exclusion restriction for instruments.

Recent literature on program evaluation focused on using the marginal treatment effects to identify various treatment effects, and their estimation. (e.g. Heckman and Vytlacil, 2005; Carneiro et al., 2010) The suggestion and identification of a parameter of economic interest, such as Policy Relevant Treatment Effect (PRTE), was successful, but the estimation places a lot of demands on the data.[5] In addition, in

[3]Note that we implicitly assumed the additive separability of the structural equations (2.2) and (2.3), as well as the participation equation (2.4).

[4]Two-stage kernel estimator or series estimator is a good example. For a detailed discussion on those estimators and their characteristics, see Ahn and Powell (1993); Newey (2009); Hahn and Ridder (2013) among others.

[5]For example, in order to identify the PRTE, the support of the propensity score $\Pr(D_i = 1|\mathbf{x}_i, \mathbf{z}_i)$ has to be $[0, 1]$ for all possible $\mathbf{x}^0 \in X$. In practice, this condition is difficult to satisfy.

practice, using the local instrumental variables to estimate the marginal treatment effect and taking the weighted average to estimate various parameters of interests appears to be too complicated. Thus, the conventional parameters of interests ATE, ATT, ATN are still widely used in practice. Accordingly, we propose a simple method to identify, quantify and test the presence of selection bias for the conventional parameters under VIA, which are weaker assumptions than CIA.

2.4 Identification of Selection Bias Using Control Functions

Analyzing the CIA in context of the control function estimation reveals the sources of selection biases and a straightforward strategy for its identification. Suppose an econometrician is interested in estimating the model (2.2)~(2.4) using the control function method. Let $K_m(\mathbf{x}_i, D(\mathbf{z}_i), \mathbf{z}_i)$ denote the control function for those whose participation status is $m \in \{0,1\}$.[6] It follows immediately from (VIA-3) that $K_m(\mathbf{x}_i, D(\mathbf{z}_i), \mathbf{z}_i) = K_m(\mathbf{x}_i, D(\mathbf{z}_i))$. When using the method of control functions, the next step of the nonparametric estimation calls for running the following regression:

$$y_{m,i} = f_m(\mathbf{x}_i) + u_{m,i} = f_m(\mathbf{x}_i) + K_m(\mathbf{x}_i, D(\mathbf{z}_i) = m) + \epsilon_{m,i}$$

where $\epsilon_{m,i}$ is a white noise process. The estimation identifies the parameters of $K_m(.)$ up to a constant, as well as those of $f_m(.)$. Using a limit argument, the constants also can be identified.[7]

If we invoke the CIAs, we have that $E[y_{m,i}|\mathbf{x}_i, D(\mathbf{z}_i)] = E[y_{m,i}|\mathbf{x}_i] = f_m(\mathbf{x}_i)$ for each $D(\mathbf{z}_i) \in \{0,1\}$. Under additive separability, the CIAs imply the following:

$$E[K_m(\mathbf{x}_i, D(\mathbf{z}_i))] = 0 \tag{2.5}$$

If the CIA does not hold, then neither does (2.5). Because $E[K_0(\mathbf{x}_i, D(\mathbf{z}_i))] = E[u_{0,i}|\mathbf{x}_i, D(\mathbf{z}_i)]$ by construction, the control functions are the means of the bias corrections. Therefore, the control function motivates the idea of identifying and testing methods of the selection biases in practice.

Before we develop the methods to test the presence of selection bias, we first clarify two questions that one can ask about the nature of selection bias. These questions are closely related to the external validity and internal validity of program evaluation, respectively.

First, an econometrician might be interested in whether the CIA are violated. We term this violation of CIA as indicating the existence of the *mean squared selection bias*. However, mean squared selection bias does not necessarily yield significant biases in the estimates of ATE, ATT, ATN parameters. For example, the bias might "cancel-out" for different values of $\mathbf{x}_i$'s, so it may appear that the bias is tiny even though there is a significant violation of the CIA. The foregoing observation leads us to the second question, the magnitude of *mean effective selection bias* for the average treatment effects. We consider this distinction between the mean squared selection bias and the mean effective selection bias as important.

[6] Notice that $K_0(\mathbf{x}_i, D(\mathbf{z}_i) = 1, \mathbf{z}_i)$ denotes a counterfactual that accounts for the selection bias for those who are in treatment group; $K_1(\mathbf{x}_i, D(\mathbf{z}_i) = 0, \mathbf{z}_i)$ denotes the opposite counterfactual, the selection bias for those in the non-participant (comparison) group.

[7] Suppose we use the propensity score $p(\mathbf{x}_i, \mathbf{z}_i)$ as an instrument, to implement the propensity score polynomial fitting to estimate the control function K_0. When the functional form of $K_0(\mathbf{x}_i, p(\mathbf{x}_i, \mathbf{z}_i))$ is not specified, we should find a small subset for which $p(\mathbf{x}_i, \mathbf{z}_i) \approx 0$ to identify the constant part of $f_0(\mathbf{x}_i)$. This is also the reason why we need an instrument, or to say differently the exclusion restriction of instruments, in the semiparametric estimation of the treatment effects. The argument for f_1 and K_1 is similar.

Suppose, for example, that an average treatment effect from a social experiment turned out to have no mean effective selection bias, but has a large mean squared selection bias for the average treatment effect. This result suggests that the bias is canceled out over the composition of population being studied, denoted by the distribution of $\mathbf{x}_i$. Nonetheless, there is a significant violation of the CIAs. A researcher should be cautious about implementing a similar treatment to other groups with different demographic compositions. For such samples, the coincidence of the selection bias "canceling out" may not happen elsewhere. To summarize, the measurement of mean squared selection bias is concerned with the external validity, while the mean effective selection bias is concerned with the internal validity.

3 Identifying and Testing the Presence of Selection Bias

Our null hypothesis H_0 is that there is no selection bias, while the alternative H_1 is that there is selection bias associated the estimated treatment effects. As explained in the previous section, we test two different types of selection biases as follows:

$$\begin{cases} H_0^B: & \text{There is no mean effective selection bias in the estimated average treatment effects} \\ H_1^B: & \text{There is a mean effective selection bias in the estimated average treatment effects} \end{cases}$$

$$\begin{cases} H_0^A: & \text{There is no violation of the Conditional Independence Assumptions} \\ & \text{(or there is no mean squared selection bias)} \\ H_1^A: & \text{There is a violation of the Conditional Independence Assumptions} \\ & \text{(or there is mean squared selection bias)} \end{cases}$$

In order to test these two different null hypotheses, we develop two different test statistics.

3.1 Identifications of the Parameters of Interest

In this section, we show how to use the control function to separate out the selection bias associated with alternative treatment effects, and to statistically test the validity of CIA under VIA. We begin by recalling the general framework of treatment effects (2.2) and (2.3), which only assumes the additive separability of the errors $u_{0,i}$ and $u_{1,i}$, respectively. Under (VIA-1)~(VIA-4) and additive separability, we express (2.2) and (2.3) in the following form:

$$y_{0,i} = f_0(\mathbf{x}_i) + K_0(\mathbf{x}_i, D(\mathbf{z}_i) = 0) + \epsilon_{0,i} \tag{3.1}$$

$$y_{1,i} = f_1(\mathbf{x}_i) + K_1(\mathbf{x}_i, D(\mathbf{z}_i) = 1) + \epsilon_{1,i}, \tag{3.2}$$

where $E[y_{m,i}|\mathbf{x}_i, D(\mathbf{z}_i)] = f_m(\mathbf{x}_i) + K_m(\mathbf{x}_i, D(\mathbf{z}_i) = m)$ for $m \in \{0,1\}$.[8] Next, observe that under the additive separability and the common support (VIA-4) assumptions, we have the following equation,

[8] Recall that $K_0(\mathbf{x}_i, D(\mathbf{z}_i) = 0)$ and $K_1(\mathbf{x}_i, D(\mathbf{z}_i) = 1)$ are the control functions associated with the untreated (non-participant) and treated (participant) states. We can also define $K_0(\mathbf{x}_i, D(\mathbf{z}_i) = 1)$ and $K_1(\mathbf{x}_i, D(\mathbf{z}_i) = 0)$, which are the control functions for the counterfactual statuses. We also define as before $u_{m,i} := K_m(\mathbf{x}_i, D(\mathbf{z}_i) = m) + \epsilon_{m,i}$.

which is the key to identify the counterfactuals:

$$\begin{aligned} 0 &\equiv E\left[u_{m,i}|\mathbf{x}_i\right] \\ &= \Pr\left(D\left(\mathbf{z}_i\right)=0|\mathbf{x}_i\right) E\left[u_{m,i}|\mathbf{x}_i, D\left(\mathbf{z}_i\right)=0\right] + \Pr\left(D\left(\mathbf{z}_i\right)=1|\mathbf{x}_i\right) E\left[u_{m,i}|\mathbf{x}_i, D\left(\mathbf{z}_i\right)=1\right] \\ &= \Pr\left(D\left(\mathbf{z}_i\right)=0|\mathbf{x}_i\right) K_m\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=0\right) + \Pr\left(D\left(\mathbf{z}_i\right)=1|\mathbf{x}_i\right) K_m\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=1\right) \end{aligned} \tag{3.3}$$

Thus, for a given counterfactual status (i.e., for fixed m), the sign of the control function $K_m\left(.,.\right)$ is the opposite for different values of $D\left(\mathbf{z}_i\right)$. Accordingly, if $K_m\left(.,.\right)$ is positive for those in the "$D_i=0$" group it must be negative for those in the "$D_i=1$" group. Nevertheless, this does not imply that biases are canceled out on average. This point is clearer when we consider the following expressions:

$$E\left[y_{1,i}-y_{0,i}|\mathbf{x}_i\right] = \left\{f_1\left(\mathbf{x}_i\right)-f_0\left(\mathbf{x}_i\right)\right\} + \left\{K_1\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=1\right) - K_0\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=0\right)\right\} \tag{3.4}$$

$$\begin{aligned} &= \left\{f_1\left(\mathbf{x}_i\right)-f_0\left(\mathbf{x}_i\right)+K_1\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=1\right)-K_0\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=1\right)\right\} \\ &\quad + \left\{K_0\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=1\right)-K_0\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=0\right)\right\} \end{aligned} \tag{3.5}$$

$$\begin{aligned} &= \left\{f_1\left(\mathbf{x}_i\right)-f_0\left(\mathbf{x}_i\right)+K_1\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=0\right)-K_0\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=0\right)\right\} \\ &\quad + \left\{K_1\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=1\right)-K_1\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=0\right)\right\} \end{aligned} \tag{3.6}$$

The left-hand side of (3.4) is the estimand of matching estimators under the CIAs. The right-hand sides of (3.4)~(3.6) separate out the treatment effect and the selection bias portions of $E\left[y_{1,i}-y_{0,i}|\mathbf{x}_i\right]$ under the assumptions (VIA-1)~(VIA-4). For example, in (3.4), the average treatment effect is given by $\Delta_{ATE}\left(\mathbf{x}_i\right)=\left\{f_1\left(\mathbf{x}_i\right)-f_0\left(\mathbf{x}_i\right)\right\}$ and the selection bias is given by $Bias_{ATE}\left(\mathbf{x}_i\right)=\left\{K_1\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=1\right)-K_0\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=0\right)\right\}$. Likewise, in (3.5), the average treatment effect on the treated is given by $\Delta_{ATT}\left(\mathbf{x}_i\right)=\left\{f_1\left(\mathbf{x}_i\right)-f_0\left(\mathbf{x}_i\right)+K_1\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=1\right)-K_0\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=1\right)\right\}$ and the selection bias is given by $Bias_{ATT}\left(\mathbf{x}_i\right)=\left\{K_0\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=1\right)-K_0\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=0\right)\right\}$. Similarly, we use equation (3.6) to separate out the average treatment effect on the non-participants, $\Delta_{ATN}\left(\mathbf{x}_i\right)$ from the $Bias_{ATN}\left(\mathbf{x}_i\right)$.

In the equations (3.4)~(3.6), $K_1\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=1\right)$ and $K_0\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=0\right)$ are identified in the control function estimations on (3.1) and (3.2). The counterfactuals $K_1\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=0\right)$ and $K_0\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=0\right)$ are identified using (3.3). This is the way that the control function method identifies $\Delta_{ATE}\left(\mathbf{x}_i\right)$, $\Delta_{ATT}\left(\mathbf{x}_i\right)$, and $\Delta_{ATN}\left(\mathbf{x}_i\right)$. Taking expectations over the covariates $\mathbf{x}_i$ yields the usual treatment parameters of interest Δ_{ATE}, Δ_{ATT}, and Δ_{ATN}. Recall that matching methods identify $E\left[y_{1,i}-y_{0,i}|\mathbf{x}_i\right]$ provided that $K_1=K_0=0$. We show in Section 5 below that (3.4)~(3.6) holds in our empirical example.

The above discussion motivates us to define the squared bias functions $A\left(\mathbf{x}_i\right)$ and net bias functions $B\left(\mathbf{x}_i\right)$ as follows:

$$\begin{aligned} A_{ATE}\left(\mathbf{x}_i\right) &:= K_1\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=1\right)^2 + K_0\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=0\right)^2 \\ A_{ATT}\left(\mathbf{x}_i\right) &:= K_0\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=1\right)^2 + K_0\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=0\right)^2 \\ A_{ATN}\left(\mathbf{x}_i\right) &:= K_1\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=1\right)^2 + K_1\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=0\right)^2 \\ B_{ATE}\left(\mathbf{x}_i\right) &:= K_1\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=1\right) - K_0\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=0\right) \qquad (3.7) \\ B_{ATT}\left(\mathbf{x}_i\right) &:= K_0\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=1\right) - K_0\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=0\right) \qquad (3.8) \\ B_{ATN}\left(\mathbf{x}_i\right) &:= K_1\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=1\right) - K_1\left(\mathbf{x}_i, D\left(\mathbf{z}_i\right)=0\right) \qquad (3.9) \end{aligned}$$

These functions capture different effects of the violation of the CIA. The squared bias functions $A(.)$'s capture the total violation of the CIA, whereas the net bias functions $B(.)$'s capture the effective violation of the CIA. By effective, we mean the violation of CIA which is reflected in the treatment effect of interest. If econometricians or program evaluators want to measure the magnitude of the violation of the CIA, they would want to use the former measure. If they are interested in the effect of selection on various average treatment effects in a particular population, they would use the latter measure.

Although in many cases the effective bias $B(.)$ is of our prime interest, the existence of squared bias $A(.)$ should not be ignored. For example, Suppose an econometrician plans to generalize the result of an experiment, for which there seems to be no selection bias in the average treatment effect, but where there is large mean squared bias.[9] Hence, when the same experiment is implemented to a different population that has different distributions of $\mathbf{x}_i$'s, the magnitude of the mean effective bias could be large for this group. This is the motivation for us recommending both measures of selection bias: the mean effective selection bias and the mean squared bias.

Now, taking expectations on squared bias functions identifies our first set of selection bias parameters of interest as follows:

$$\begin{aligned}
\Psi^A_{ATE} &:= E\left[A_{ATE}\left(\mathbf{x}_i\right)\right] \\
\Psi^A_{ATT} &:= E\left[A_{ATT}\left(\mathbf{x}_i\right)\right] \\
\Psi^A_{ATN} &:= E\left[A_{ATN}\left(\mathbf{x}_i\right)\right]
\end{aligned}$$

Taking expectations on the net bias function, with appropriate weights, identifies our second set of parameters of interest:

$$\begin{aligned}
\Psi^B_{ATE} &:= E\left[B_{ATE}\left(\mathbf{x}_i\right)\right] \\
\Psi^B_{ATT} &:= E\left[B_{ATT}\left(\mathbf{x}_i\right)\right] \\
\Psi^B_{ATN} &:= E\left[B_{ATN}\left(\mathbf{x}_i\right)\right]
\end{aligned}$$

[9] Again, this might happen because simply the biases are canceled out by coincidence over different $\mathbf{x}_i$'s.

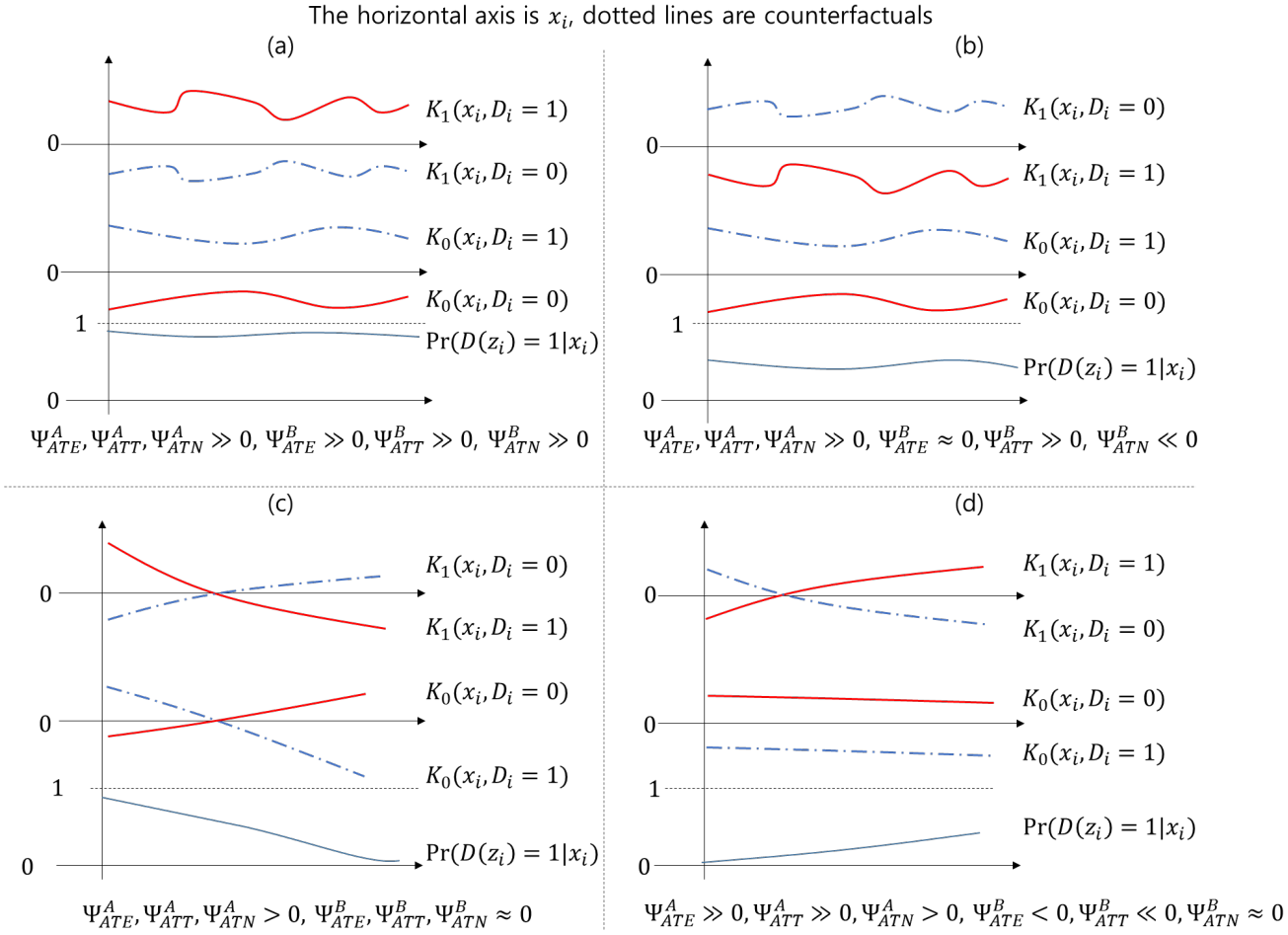
Before we proceed, we define the sample analogues of Ψ^A's and Ψ^B's as below:

$$\begin{aligned}
\hat{\Psi}^A_{ATT} &:= \frac{1}{n}\sum_{i=1}^{n}\hat{A}_{ATT}(\mathbf{x}_i) \\
\hat{\Psi}^A_{ATN} &:= \frac{1}{n}\sum_{i=1}^{n}\hat{A}_{ATN}(\mathbf{x}_i) \\
\hat{\Psi}^A_{ATE} &:= \frac{1}{n}\sum_{i=1}^{n}\hat{A}_{ATE}(\mathbf{x}_i) \\
\hat{\Psi}^B_{ATT} &:= \frac{1}{n}\sum_{i=1}^{n}\hat{B}_{ATT}(\mathbf{x}_i) \\
\hat{\Psi}^B_{ATN} &:= \frac{1}{n}\sum_{i=1}^{n}\hat{B}_{ATN}(\mathbf{x}_i) \\
\hat{\Psi}^B_{ATE} &:= \frac{1}{n}\sum_{i=1}^{n}\hat{B}_{ATE}(\mathbf{x}_i)
\end{aligned}$$

where $\hat{A}(.)$ and $\hat{B}(.)$'s are the estimated bias functions and n is the sample size.

The four sets of graphs in Figure 3.1 plot the control functions as a function of $\mathbf{x}_i$ for four hypothetical cases that display different magnitudes of the Ψ's. As defined above, $K_1(\mathbf{x}_i, D(\mathbf{z}_i) = 1)$ and $K_0(\mathbf{x}_i, D(\mathbf{z}_i) = 0)$ denote the observed control function for participants and nonparticipants, respectively, (shown by the the solid lines in the figures) and $K_1(\mathbf{x}_i, D(\mathbf{z}_i) = 0)$ and $K_0(\mathbf{x}_i, D(\mathbf{z}_i) = 1)$ denote their corresponding counterfactuals (shown by the dashed lines in each of the figures). Each figure has three panels, the top two panels are intended to have the same scale. We could have put all four control functions in the same plane, but broke it out this way to make the figures easier to read. The third panels in each graph plots the propensity scores which range between 0 and 1 as a function of $\mathbf{x}_i$.

Figure 3.1: Magnitudes of the Selection Bias, Ψ, for Some Hypothetical Cases



For now we focus on Figure 3.1 (a) for an illustration. According to the bias equations (3.7) through (3.9), we have the bias for ATE given by $K_1(\mathbf{x}_i, D(\mathbf{z}_i) = 1) - K_0(\mathbf{x}_i, D(\mathbf{z}_i) = 0)$. So in Figure 3.1, each term in this expression is given by the bold lines. Given how this figure is drawn, we observe that $K_1(\mathbf{x}_i, D(\mathbf{z}_i) = 1)$ is uniformly positive and $K_0(\mathbf{x}_i, D(\mathbf{z}_i) = 0)$ is uniformly negative. So the Bias term $\hat{\Psi}^B_{ATE}$ is positive. Next we turn to the bias term $\hat{\Psi}^B_{ATT}$. We first note that this term is given by the difference between the counterfactuals for $D_i = 1$ and the observed control function for $D_i = 0$. In this case $K_0(\mathbf{x}_i, D(\mathbf{z}_i) = 1)$ is uniformly positive and $K_0(\mathbf{x}_i, D(\mathbf{z}_i) = 0)$ is uniformly negative, so the selection bias for ATT is positive.

The remaining Figure 3.1 (b) through (d) display different outcomes followed by different circumstances. For example, in Figure 3.1 (c), the lines are drawn so that the bias cancels out for ATE, ATT and ATN.

3.2 Test Statistics and Their Asymptotic Distributions

In this subsection, we derive our tests of violations of CIA and the presence of mean effective selection bias, by constructing test statistics for which the asymptotic distribution is normal. We begin by laying out additional assumptions. Again, we assume that (VIA-1) through (VIA-4) hold as well the additive separability of the error terms both in the structural equations and in the participation equation. These

additional assumption are as follows:

Assumption 1. *$X \subset \mathbb{R}^p$ is compact.*

Assumption 2. *For $m \in \{0,1\}$, $E[|u_{m,i}|] < \infty$.*

The above two assumptions ensure the existence of all the control functions over X, including the counterfactuals.

Lemma 3.1. *(Existence of Control Functions) $\forall \mathbf{x} \in X$, for $m, l \in \{0,1\}$, $|K_m(\mathbf{x}, l)| < \infty$ a.e. on X.*

Proof. By the law of iterated expectations, we have:

$$E[|u_{m,i}|] = E[E[|u_{m,i}| \,|\mathbf{x}_i, D(\mathbf{z}_i) = m]] < \infty$$

This implies that

$$|K_m(\mathbf{x}, m)| := |E[u_{m,i}|\mathbf{x}_i, D(\mathbf{z}_i) = m]| \leq E[|u_{m,i}| \,|\mathbf{x}_i, D(\mathbf{z}_i) = m] < \infty$$

almost everywhere on $X \times \{0,1\}$. Then, because of (3.3) with (VIA-4), it is also true that $K_m(\mathbf{x}, l)$ exists for $m \neq l$. □

The following two assumptions are needed to ensure the Central Limit Theorem (CLT) and that bootstrap works for the bias estimators.

Assumption 3. *$\mathbf{x}_i (\in X)$ contains at least one continuous variable.*

Assumption 4. *For $m \in \{0,1\}$, $K_m(\mathbf{x}, m)$ is continuously differentiable in the continuous elements of $\mathbf{x}$.*

Assumption 5. *There exists a semiparametric pointwise $\sqrt{n}$-consistent, asymptotically normal estimator $\hat{K}_{m,n}$ for K_m on its domain. That is, $\forall \mathbf{x} \in X$, $\sqrt{n}\left\{\hat{K}_{m,n}(\mathbf{x}, l) - K_m(\mathbf{x}, l)\right\} \to_d \mathcal{N}\left(0, Var\left(\hat{K}_{m,n}(\mathbf{x}, l)\right)\right)$, as $n \to \infty$.*

Assumption 5 is about the existence of a $\sqrt{n}$-consistent and asymptotically normal control function estimator. Semiparametric estimators of the control functions have been studied extensively in the literature. Such estimators may include the two-step series estimators, kernel estimators and spline estimators, (see Das et al., 2003; Newey, 2009; Hahn and Ridder, 2013; Ahn and Powell, 1993 among others) which may require different regularity conditions for the $\sqrt{n}$-consistency and asymptotic normality. Because developing an another estimator is beyond the scope of this paper, we assume that such an estimator exists and the regularity conditions are met.

To make clear that $\hat{K}_m$ is estimated by using the sample with size n, we index the estimated control function by $\hat{K}_{m,n}$. Note that the variance of the control functions are indexed by $\mathbf{x}$. This is because the control function K_m is in general nonlinear, and the variance of the estimate $\hat{K}_{m,n}$ might be different over X. $Var\left(\hat{K}_{m,n}(\mathbf{x}, l)\right)$ is different from $Var(K_m(\mathbf{x}, l))$ as the asymptotic variance involves the variation coming from the estimation as well as from the $Var(K_m(\mathbf{x}, l))$.

Assumption 6. *(i) For each n, $\hat{K}_{m,n}(\mathbf{x}, m)$ is continuous in $\mathbf{x}$ in the continuous element of $\mathbf{x}$.*

(ii) For $k, l \in \{0,1\}$, for some $\delta > 0$, $E\left[\left\{\hat{K}_{m,n}(\mathbf{x}, l)\right\}^{4+\delta}\right] < \infty$ a.e. on X for a large enough n.

Assumption 6 is on the smoothness and on the existence of fourth moment of the estimator. This assumption is required in order to invoke the CLT.

We want to average the control function estimates over X, and then find the asymptotic distribution of the sample average of these control function estimates. To be specific, we want the estimates to follow the normal distribution at least asymptotically. This will be true if the following Lindeberg condition holds.

Lemma 3.2. *(Lindeberg Condition for $\hat{K}_m$) $\forall \eta > 0$, $\forall \epsilon > 0$, $\forall m, l \in \{0,1\}$,*

$$\lim_{n\to\infty} \frac{1}{\hat{s}_n^2} \sum_{i=1}^{n} E\left[\mathbf{1}\left(\left\{\left|\hat{K}_{m,n}(\mathbf{x}_i, l) - K_m(\mathbf{x}, l)\right| > \epsilon \hat{s}_n\right\}\right)\left\{\hat{K}_{m,n}(\mathbf{x}_i, l) - K_m(\mathbf{x}, l)\right\}^2\right] = 0 \tag{3.10}$$

where $\hat{s}_n^2 = \sum_{i=1}^{n} Var\left(\hat{K}_{m,n}(\mathbf{x}_i, l)\right)$.

Proof. Fix $\epsilon > 0$, Let $m, l \in \{0,1\}$. Let $n > 0$. By Assumption 5, we have $\hat{K}_{m,n}(\mathbf{x}_i, l) \to_p K_m(\mathbf{x}_i, l)$ so that $\hat{K}_{m,n}(\mathbf{x}_i, l) \to_d K_m(\mathbf{x}_i, l)$.

Let $\lambda > 0$ such that the distribution of $K_m(\mathbf{x}_i, l)$ is continuous at λ. Consider:

$$\begin{aligned}
& E\left[\mathbf{1}\left(\left\{\left|\hat{K}_{m,n}(\mathbf{x}_i, l) - K_m(\mathbf{x}, l)\right| > \lambda\right\}\right)\left\{\hat{K}_{m,n}(\mathbf{x}_i, l) - K_m(\mathbf{x}, l)\right\}^2\right] \\
= \; & E\left[\left\{\hat{K}_{m,n}(\mathbf{x}_i, l) - K_m(\mathbf{x}, l)\right\}^2\right] - E\left[\mathbf{1}\left(\left\{\left|\hat{K}_{m,n}(\mathbf{x}_i, l) - K_m(\mathbf{x}, l)\right| \le \lambda\right\}\right)\left\{\hat{K}_{m,n}(\mathbf{x}_i, l) - K_m(\mathbf{x}, l)\right\}^2\right] \\
= \; & Var\left(\hat{K}_{m,n}(\mathbf{x}_i, l)\right) - E\left[\mathbf{1}\left(\left\{\left|\hat{K}_{m,n}(\mathbf{x}_i, l) - K_m(\mathbf{x}, l)\right| \le \lambda\right\}\right)\left\{\hat{K}_{m,n}(\mathbf{x}_i, l) - K_m(\mathbf{x}, l)\right\}^2\right] \\
\to \; & Var\left(K_m(\mathbf{x}_i, l)\right) - E\left[\mathbf{1}\left(\left\{\left|K_m(\mathbf{x}_i, l) - K_m(\mathbf{x}, l)\right| \le \lambda\right\}\right)\left\{K_m(\mathbf{x}_i, l) - K_m(\mathbf{x}, l)\right\}^2\right]
\end{aligned} \tag{3.11}$$

Equation (3.11) follows by the following logic. First, we have the fact that $\hat{K}_{m,n}(\mathbf{x}_i, l) \to_d K_m(\mathbf{x}_i, l)$. Next, assumption 6 ensures the uniform integrability of both $\mathbf{1}\left(\left\{\left|\hat{K}_{m,n}(\mathbf{x}_i, l) - K_m(\mathbf{x}, l)\right| \le \lambda\right\}\right)$ and $\left\{\hat{K}_{m,n}(\mathbf{x}_i, l) - K_m(\mathbf{x}, l)\right\}^2$ over n. Furthermore, we have $\mathbf{1}\left(\left\{\left|\hat{K}_{m,n}(\mathbf{x}_i, l) - K_m(\mathbf{x}, l)\right| \le \lambda\right\}\right)$ bounded below by 0 and above by 1 so the multiplications of those two also are uniformly integrable over n. Thus, convergence in distribution implies the convergence of moments.

Lastly, we have that $Var\left(\hat{K}_{m,n}(\mathbf{x}_i, l)\right) \to Var\left(K_m(\mathbf{x}_i, l)\right) > 0$ by our implicit assumption that $Var\left(K_m(\mathbf{x}_i, l)\right) > 0$. This implies $\hat{s}_n^2 = \sum_{i=1}^{n} Var\left(\hat{K}_{m,n}(\mathbf{x}_i, l)\right) \to \infty$ so that $\hat{s}_n \epsilon \to \infty$. Plugging $\lambda = \hat{s}_n \epsilon$ in (3.11) asserts that (3.11) tends to zero as $n \to \infty$. This implies the condition (3.10) holds. □

Lemma 3.2 establishes the condition for applying the Lindeberg-Feller CLT on $\hat{K}_m(\mathbf{x}_i, l)$, which will play a central role in characterizing the asymptotic distributions of our test statistics for Ψ^B's. Because $\hat{B}(\mathbf{x}_i)$'s are simple subtractions of $\hat{K}_m(\mathbf{x}_i, l)$'s, using the same logic, the Lindeberg condition for $\hat{K}_m$'s naturally hold for $\hat{B}(.)$'s. By exactly the same logic, the Lindeberg condition for $\hat{\Psi}^A$'s also holds, because we assumed that the fourth moment of $\hat{K}_m$ exists. One can exactly repeat the proof, replacing $\hat{K}_{m,n}$ by $\hat{K}_{m,n}^2$. We omit the statement and proof here for $\hat{\Psi}^A$.

The variance of the control functions, $Var\left(\hat{K}_m\left(\mathbf{x}_i, l\right)\right)$, comes from two sources: (i) the variation of $\mathbf{x}_i$, and (ii) the variation of the estimates of $\hat{K}_m$. To see this fact more clearly, consider the definition of the variance:

$$Var\left(\hat{K}_m\left(\mathbf{x}_i, l\right)\right) = \int_{\mathbf{x}\in X}\left\{\hat{K}_m\left(\mathbf{x}_i = \mathbf{x}, l\right)\right\}^2 d\Pr\left(\mathbf{x}\right) - \left\{\int_{\mathbf{x}\in X}\hat{K}_m\left(\mathbf{x}_i = \mathbf{x}, l\right) d\Pr\left(\mathbf{x}\right)\right\}^2$$

If $K_m\left(\mathbf{x}_i, l\right)$ is linear in $\mathbf{x}_i$ and the coefficient is estimated as a sample average, the variance is easy and straightforward to compute. However, because we consider the semiparametric estimators which are nonlinear functions of $\mathbf{x}_i$, computing the variances analytically is often infeasible. For example, fitting a polynomial of the estimated propensity score to estimate $K_m\left(.,.\right)$ involves a nonlinear function of $\mathbf{x}_i$, namely, $\Pr\left(D_i = 1|\mathbf{x}_i, \mathbf{z}_i\right)$ or its consistent estimator. Therefore, we want to use the empirical distribution to establish the consistent estimator for the variance. To be specific, we employ bootstrap methods to compute the standard errors.

In order to employ the Lindeberg CLT, we want to estimate $\frac{1}{n^2}s_n^2 := \frac{1}{n^2}\sum_{i=1}^n Var\left(\hat{K}_m\left(\mathbf{x}_i, l\right)\right)$ using resampling methods. The following lemma substantially simplifies the estimation of s_n^2 especially when $\mathbf{x}_i$ is a continuous covariate or the dimension of $\mathbf{x}_i$ is high.

Lemma 3.3. *(Consistent Estimator for $\frac{1}{n^2}s_n^2$) $\widehat{Var}\left(\frac{1}{n}\sum_{i=1}^n \hat{K}_m\left(\mathbf{x}_i, l\right)\right) \to_{a.s.} \frac{1}{n^2}s_n^2$ where $\widehat{Var}$ denotes the sample variance.*

Proof. Fix $l \in \{0, 1\}$. We have:

$$\begin{aligned}
\frac{1}{n^2}s_{B,n}^2 &:= \frac{1}{n^2}\sum_{i=1}^n Var\left(\hat{K}_m\left(\mathbf{x}_i, l\right)\right) \\
&= \frac{1}{n^2}\left[\sum_{i=1}^n Var\left(\hat{K}_m\left(\mathbf{x}_i, l\right)\right) + \sum_{i\neq j} Cov\left(\hat{K}_m\left(\mathbf{x}_i, l\right), \hat{K}_m\left(\mathbf{x}_j, l\right)\right)\right] \\
&= \frac{1}{n^2}Var\left(\sum_{i=1}^n \hat{K}_m\left(\mathbf{x}_i, l\right)\right) \\
&= Var\left(\frac{1}{n}\sum_{i=1}^n \hat{K}_m\left(\mathbf{x}_i, l\right)\right)
\end{aligned}$$

by independence of $\mathbf{x}_i$'s and the measurability of $\hat{K}_m$. Therefore, the strong law of large numbers establishes the result as desired. □

For the sake of notational simplicity, define $\frac{1}{n^2}\hat{s}_{B,n}^2 := \widehat{Var}\left(\frac{1}{n}\sum_{i=1}^n \hat{B}\left(\mathbf{x}_i\right)\right) = \widehat{Var}\left(\hat{\Psi}^B\right)$. In practice, we can use bootstrap methods to calculate consistent estimators of $\frac{1}{n^2}s_{B,n}^2$. Analogously, we define $\hat{s}_{B,ATT,n}, \hat{s}_{B,ATN,n}, \hat{s}_{B,ATE,n}\hat{s}_{A,ATT,n}, \hat{s}_{A,ATN,n}$ and $\hat{s}_{A,ATE,n}$.

We want to test the null hypothesis H_0 that there is no mean effective selection bias: $H_0^B : \Psi^B = 0$ against the alternative hypothesis H_1^B that here is effective selection bias: $H_1^B : \Psi^B \neq 0$. An application of Continuous Mapping theorem and Lindeberg's CLT yields the following theorem.

Theorem 3.1. *(Distributions of Test Statistics for $\hat{B}\left(.\right)$)*

$$\begin{aligned}
\sqrt{n}\frac{n}{\hat{s}_{B,ATT,n}}\left[\hat{\Psi}^B_{ATT}-\Psi^B_{ATT}\right] &\to_d \mathcal{N}(0,1)\\
\sqrt{n}\frac{n}{\hat{s}_{B,ATN,n}}\left[\hat{\Psi}^B_{ATN}-\Psi^B_{ATN}\right] &\to_d \mathcal{N}(0,1)\\
\sqrt{n}\frac{n}{\hat{s}_{B,ATE,n}}\left[\hat{\Psi}^B_{ATE}-\Psi^B_{ATE}\right] &\to_d \mathcal{N}(0,1)
\end{aligned}$$

Similarly, we want to test the null hypothesis H_0^A that there is no mean squared selection bias: $H_0^A : \Psi^A = 0$. against alternative hypothesis H_1^A is there is mean squared selection bias: $H_1^A : \Psi^A \neq 0$. We state the following theorem regarding the distribution for $\hat{\Psi}^A$'s without proof. The logic for deriving the asymptotic distributions is exactly the same, except that we require the existence of the fourth moment of $\hat{K}_m$.

Theorem 3.2. *(Distributions of Test Statistics for $\hat{A}(.)$)*

$$\begin{aligned}
\sqrt{n}\frac{n}{\hat{s}_{A,ATT,n}}\left[\hat{\Psi}^A_{ATT}-\Psi^A_{ATT}\right] &\to_d \mathcal{N}(0,1)\\
\sqrt{n}\frac{n}{\hat{s}_{A,ATN,n}}\left[\hat{\Psi}^A_{ATN}-\Psi^A_{ATN}\right] &\to_d \mathcal{N}(0,1)\\
\sqrt{n}\frac{n}{\hat{s}_{A,ATE,n}}\left[\hat{\Psi}^A_{ATE}-\Psi^A_{ATE}\right] &\to_d \mathcal{N}(0,1)
\end{aligned}$$

Comparing the relative size of $\hat{\Psi}^A$'s and $\hat{\Psi}^B$'s with the computed size of respective average treatment effects reveals the magnitude of selection biases. The size of $\hat{\Psi}$'s relative to their corresponding average treatment effects, as well as statistical significance, also is important. We shall revisit this topic below.

3.3 Bootstrap Algorithm for the Test Statistics

In the previous subsection, we computed the asymptotic distribution of our proposed test statistics. In particular, we found that the studentized test statistics follow the standard normal distribution. Given that the CLT holds for the test statistics, the bootstrap distribution estimates are consistent, and thus the variance estimates using the bootstrap distribution is consistent as well. Furthermore, instead of using the standard error estimates from bootstrap distribution for the normal approximation, we can use the bootstrap distribution of the test statistics $\hat{\Psi}^A$'s and $\hat{\Psi}^B$'s. Bootstrap achieves the second-order accuracy on the confidence intervals of the test statistics under the null hypothesis, because the studentized $\hat{\Psi}^A$'s and $\hat{\Psi}^B$'s are asymptotically pivotal. The errors of bootstrap two-sided equal-tailed rejection probabilities have size $O\left(n^{-\frac{3}{2}}\right)$, while the errors of the first-order asymptotic rejection probabilities have size $O\left(n^{-1}\right)$. We only suggest the bootstrap algorithm and skip the detailed discussion and proof on this subject. The formal proofs on the reductions of error size can be found in standard textbooks, such as Shao and Tu (1995); DasGupta (2008).

The algorithm to construct the bootstrap confidence interval for our proposed test statistics is as follows:

1. Compute the $\hat{\Psi}$ of interest using the original sample.

2. Generate a bootstrap sample of size n randomly, with replacement, where n is the corresponding sample size of interest. Conduct the estimation. Denote the bootstrap analogue of $\hat{\Psi}$ as $\hat{\Psi}^*$.

3. Repeat 2 for enough times.

4. The bootstrap confidence interval of level-α test is $\left(\hat{\Psi} - q_n^*\left(1-\frac{\alpha}{2}\right), \hat{\Psi} - q_n^*\left(\frac{\alpha}{2}\right)\right)$ where $q_n^*\left(\frac{\alpha}{2}\right)$ is the $1-\frac{\alpha}{2}$'th quantile of the bootstrap distribution of $\hat{\Psi}^* - \hat{\Psi}$.

We reject the null hypothesis that there is no selection bias for the corresponding type of the test at significance level α if $0 \notin \left(\hat{\Psi} - q_n^*\left(1-\frac{\alpha}{2}\right), \hat{\Psi} - q_n^*\left(\frac{\alpha}{2}\right)\right)$.

In Section 4, we shall examine the finite sample properties of our proposed test statistics. In particular, we shall compare the finite sample confidence intervals of the normal approximations, bootstrapped confidence intervals, and the confidence intervals from the true data-generating process.

3.4 Economic Significance, Statistical Significance, and Power of the Test

The test suggested in the previous subsections provide a way to systematically test the presence of selection bias. However, a tiny difference caused by a selection bias might turn out to be highly statistically significant as the sample size n gets large. An economically more interesting question is not just the statistical significance of the differences, but the relative size of the differences caused by the selection bias. Thus we propose the following measures of selection biases:

$$\begin{aligned} M_B &:= \frac{\hat{\Psi}^B_{ATE}}{\hat{\Delta}_{ATE}} \\ M_A &:= \frac{\sqrt{\hat{\Psi}^A_{ATE}}}{\hat{\Delta}^C_{ATE}} \end{aligned}$$

An ad-hoc criterion of determination that there exists an economically significant selection bias involved can be suggested that M_B, M_A are larger than, for example, 0.05 and statistically significant in given level α.

In the next section, we conduct Monte-Carlo simulations to examine the performance of the alternative measures that we propose for testing for the presence of selection bias.

4 Monte-Carlo Simulations to Examine the Performances of the Test Statistics

This section examines the finite sample properties of the test statistics that we have proposed in this paper. We first compare the power of our suggested test statistics for selection bias and model specification bias, by sample size n. Next, we employ the two approaches to computing the standard errors, namely, (i) the first-order asymptotic approximation and (ii) the bootstrap, to compute the confidence interval suggested in the previous section. Finally, we repeat these procedures for the cases when the data-generating process is correctly specified and when it is misspecified.

4.1 Setup and Estimation

The data-generating process follows the assumptions (VIA-1) through (VIA-4). The data is assumed to be generated by the generalized Roy model. The joint normality assumptions imposed by Roy model will be weakened when we consider our empirical application in the next section.

4.1.1 No Specification Bias: Setup

We assume that there are two-dimensional covariates $\mathbf{x}_i \in \mathbb{N} \times \mathbb{R}$ and one exogenous random instrument $z_i \in \{0, 1\}$. One of the covariates is a continuous variable; the other is a discrete variable. For example, we can regard the continuous variable as the current wage and the discrete variable as the education level. To be concrete, we assume that $x_{1,i} \sim \text{lognormal}\left(2, 0.5^2\right)$, $x_{2,i} \sim \text{Poisson}\left(5\right)$. z_i's are randomly assigned with probability $\frac{1}{2}$.

We specify the Roy model as follows:

$$
\begin{aligned}
y_{0,i} &= f_0\left(\mathbf{x}_i\right) + \eta_i^0 + e_i && (4.1)\\
y_{1,i} &= f_1\left(\mathbf{x}_i\right) + \eta_i^1 + e_i && (4.2)\\
d_i^* &= f_d\left(\alpha^1 - \alpha^0, \mathbf{x}_i'\left(\beta^1 - \beta^0\right), z_i\right) + \eta_i^1 - \eta_i^0 + \nu_i && (4.3)\\
d_i &= \begin{cases} 1 & \text{if } d_i^* \geq 0 \\ 0 & \text{if } d_i^* < 0 \end{cases} && (4.4)
\end{aligned}
$$

We first characterize $f_0\left(.\right), f_1\left(.\right), f_d\left(.\right)$ and the values of $\alpha^0, \alpha^1, \beta^0, \beta^1$.

$$
\begin{aligned}
f_0\left(\mathbf{x}_i\right) &:= \alpha^0 + \mathbf{x}_i'\beta^0 && (4.5)\\
f_1\left(\mathbf{x}_i\right) &:= \alpha^1 + \mathbf{x}_i'\beta^1 && (4.6)\\
f_d\left(\alpha^1 - \alpha^0, \mathbf{x}_i'\left(\beta^1 - \beta^0\right), z_i\right) &:= f_1\left(\mathbf{x}_i\right) - f_0\left(\mathbf{x}_i\right) + \delta\left(z_i - 0.8\right) && (4.7)
\end{aligned}
$$

where we let $(\beta_1^0, \beta_2^0) = (2, 15)$, $(\beta_1^1, \beta_2^1) = (5, 10)$, $\alpha^0 = 30$, $\alpha^1 = 50$, $\delta = 40$. Because $\beta_2^0 > \beta_2^1$, there is some opportunity cost of being treated, which is increasing in the discrete variable $x_{2,i}$. The adjustment on z_i, $z_i - 0.8$ is made to roughly center the value of d_i^* around zero. Notice that large values of δ can de-facto force treatment status on the individual, as δ can be interpreted as the weight given on z_i in the selection process.

Next, we describe the distributions of the unobservables (or the error terms). The specification NS/NS has only selection-on-observables, while the specification NS/WS has selection-on-unobservables.

NS/NS (No Specification Bias, No Selection Bias) f_0 and f_1 are linear in parameters, $e_i \sim \mathcal{N}\left(0, 10^2\right)$, $\eta_i^0 = \eta_i^1 = 0$, $\nu_i \sim \mathcal{N}\left(0, 12^2\right)$.

NS/WS (No Specification Bias, With Selection Bias) f_0 and f_1 are linear in parameters, $e_i \sim \mathcal{N}\left(0, 10^2\right)$, $\eta_i^0 \sim \mathcal{N}\left(0, 8^2\right)$, $\eta_i^1 \sim \mathcal{N}\left(0, 16^2\right)$, $\nu_i \sim \mathcal{N}\left(0, 12^2\right)$.

4.1.2 Specification Bias: Setup

The no specification bias setup specifies $f_0(.)$ and $f_1(.)$ in the data generating process to be linear in parameters, and we shall assume that the econometrician specifies the data generating process as linear. However, this linear specification is usually thought to approximate some non-linear data generating process. This approximation could cause our estimated treatment effects to be biased, not because of selection bias, but because of model misspecification.

Accordingly, we examine the case in which the true data generating process is concave in covariates and the parameters, but the estimation procedure specifies linear relationship between the covariates, the parameters and the outcome. The baseline specification of Roy model (4.1)~(4.4) is same in this setup, but (4.5)~(4.7) are modified as following:

$$f_0(\mathbf{x}_i) := \sqrt{\alpha^0 + \mathbf{x}_i\beta^0} \tag{4.8}$$

$$f_1(\mathbf{x}_i) := \sqrt{\alpha^1 + \mathbf{x}_i\beta^1} \tag{4.9}$$

$$f_d\left(\alpha^1 - \alpha^0, \mathbf{x}_i\left(\beta^1 - \beta^0\right), z_i\right) := f_1(\mathbf{x}_i) - f_0(\mathbf{x}_i) + \delta(z_i - 0.8) \tag{4.10}$$

In order for the treatment effects to be estimated in a similar range with those in NS/WS and NS/NS, we squared the values of $\alpha^0, \alpha^1, \beta^0, \beta^1$ so that we have $(\beta_1^0, \beta_2^0) = (2^2, 15^2)$, $(\beta_1^1, \beta_2^1) = (5^2, 10^2)$, $\alpha^0 = 30^2$, $\alpha^1 = 50^2$. We leave δ unchanged so that $\delta = 40$. We continue to assume that the unobservables (or the error terms) are separable and their variances are also left unchanged and identical with those in NS/WS and NS/NS, respectively. The following two lines summarize the setup WS/NS and WS/WS in which the data generating process is concave in covariates.

WS/NS (With Specification Bias, No Selection Bias) f_0 and f_1 is concave in parameters, $e_i \sim \mathcal{N}(0, 10^2)$, $\eta_i^0 = \eta_i^1 = 0$, $\nu_i \sim \mathcal{N}(0, 12^2)$.

WS/WS (With Specification Bias and Selection Bias) f_0 and f_1 is concave in parameters, $e_i \sim \mathcal{N}(0, 10^2)$, $\eta_i^0 \sim \mathcal{N}(0, 8^2)$, $\eta_i^1 \sim \mathcal{N}(0, 16^2)$, $\nu_i \sim \mathcal{N}(0, 12^2)$.

4.1.3 Estimation

To estimate the parameters of the control function, we use the two-step propensity score series estimator.[10] In our estimations, we use the logit or probit to estimate the propensity score $\Pr(d_i = 1|\mathbf{x}_i, z_i)$, and fit a polynomial of the estimated propensity scores to estimate the control function. The functions subject to estimation are:

$$y_{0,i} = \alpha^0 + \mathbf{x}_i'\beta^0 + \epsilon_{0,i} \tag{4.11}$$

$$y_{1,i} = \alpha^1 + \mathbf{x}_i'\beta^1 + \epsilon_{1,i} \tag{4.12}$$

$$d_i^* = \alpha^d + \mathbf{x}_i'\beta^d + \delta^d z_i + \epsilon_{d,i} \tag{4.13}$$

$$d_i = \begin{cases} 1 & \text{if } d_i^* \geq 0 \\ 0 & \text{if } d_i^* < 0 \end{cases} \tag{4.14}$$

[10] For the detailed discussions on the properties of this estimator, see Newey (2009) among others.

We estimated the control functions using a second order polynomial of the propensity scores. Degrees higher than a second order polynomial did not have any gain on estimation. To estimate the bias terms $\hat{\Psi}$'s, we followed the estimation strategy described in Section 3.

After the estimates of $\hat{\Psi}$'s are obtained for the whole sample, we bootstrapped the estimation procedure $3,000$ times. We repeated the full sample estimation and bootstrap process for samples of $n = 10^3, 10^4, 10^5, 10^6$, respectively. We also ran the same procedure for the whole data-generating process with $n = 10^6$ observations in order to obtain references for the true behavior of the quantities of interest. To obtain the DGP rows in Table 1, we repeated the data-generating process for $3,000$ times for $n = 10^6$ observations. Accordingly, we regard these values as very close approximation to the true values of these parameters and quantities.

4.2 Results of the Monte-Carlo Simulations

Table 1 summarizes the finite sample behaviors of our proposed test statistics. For each of our four specifications, we report (i) the bootstrapped values of $\hat{\Psi}$'s and (ii) their bootstrap standard errors in the rows labeled "10^3" through "10^6" .

We find that the performance of our test statistics is robust with respect to the sample size, n. To be specific, $\hat{\Psi}^B$'s obtain reasonable power when n is as small as 10^3 observations, while $\hat{\Psi}^A$'s obtain reasonable power when n is as small as 10^4 observations. We suspect such discrepancies arise because $\hat{\Psi}^A$'s are the mean of the squared biases. Taking the squares of the bias makes $\hat{\Psi}^A$'s more sensitive to the extreme values. Overall, our results indicate that our tests for selection bias do not require unreasonably large samples and can be implemented using sample sizes commonly available in many datasets.

The standard error estimates in Table 1 also provide evidence of the consistency of bootstrap standard error estimates. In particular, the standard error shrinks by roughly $1/\sqrt{10}$ when the sample size is increased by 10 times, and coincides with that of DGP when $n = 10^6$. This result coincides with what is expected in Theorems 3.1 and 3.2. Thus, here we find a heuristic justification for using the first-order asymptotic approximation for our hypothesis tests.[11]

The magnitude of $\hat{\Psi}^B$'s is approximately the same as the average size of selection bias associated with matching estimates of corresponding average treatment effects. We also show below in our empirical example that the magnitude of $\hat{\Psi}^B$s is similar to the difference between matching and control function estimates of the average treatment effects.

[11] For an ad-hoc explanation for this result, notice that the variance of $\hat{\Psi}$ are similar to $Var\left(\frac{1}{n}\sum_{i=1}^{n}\hat{B}\left(\mathbf{x}_i\right)\right)$. Assuming the independence of $\mathbf{x}_i$'s and $\hat{B}\left(\mathbf{x}_i\right)$'s do not vary much with respect to i, the variance becomes around $1/10$ when n gets 10 times larger. Thus the standard error will shrink by around $1/\sqrt{10}$ as the sample size is increased by 10 times.

Table 1: Performances of Bias Estimators by Sample Size n

	n	$\hat{\Psi}^B_{ATE}$	$\hat{\Psi}^B_{ATT}$	$\hat{\Psi}^B_{ATN}$	$\hat{\Psi}^A_{ATE}$	$\hat{\Psi}^A_{ATT}$	$\hat{\Psi}^A_{ATN}$
No Specification Bias No Selection Bias (NS/NS)	10^3	0.195	5.856	−7.355	1.720	32.213	53.731
		(1.702)	(6.070)	(5.198)	(7.234)	(82.019)	(86.556)
	10^4	0.498	1.139	0.422	0.236	1.100	0.165
		(0.511)	(1.901)	(1.599)	(0.841)	(6.474)	(3.897)
	10^5	−0.300	0.178	−1.172	0.168	0.075	1.404
		(0.159)	(0.569)	(0.503)	(0.150)	(0.490)	(1.675)
	10^6	0.018	−0.130	−0.237	0.005	0.075	1.404
		(0.051)	(0.191)	(0.165)	(0.009)	(0.072)	(1.116)
	DGP	−0.001	−0.002	0.001	0.002	0.002	2.384
		(0.050)	(0.188)	(0.163)	(0.005)	(0.048)	(1.047)
With Specification Bias No Selection Bias (WS/NS)	10^3	−0.020	−5.435	−1.682	1.466	29.873	2.512
		(2.334)	(17.212)	(29.720)	(12.592)	(801.801)	(1736.122)
	10^4	1.914	12.026	8.573	2.718	139.787	67.099
		(0.726)	(7.470)	(10.807)	(2.048)	(192.997)	(231.965)
	10^5	0.475	9.190	−0.021	0.291	80.961	0.007
		(0.205)	(2.352)	(2.627)	(0.155)	(42.041)	(41.041)
	10^6	0.495	12.300	−0.293	0.456	145.288	0.080
		(0.067)	(0.754)	(0.896)	(0.058)	(17.881)	(1.166)
	DGP	0.439	12.738	−1.339	0.483	156.379	2.443
		(0.067)	(0.751)	(0.892)	(0.059)	(18.411)	(2.473)
No Specification Bias With Selection Bias (NS/WS)	10^3	7.041	5.931	9.997	106.867	37.045	77.134
		(2.511)	(6.814)	(7.580)	(35.780)	(87.613)	(126.273)
	10^4	7.029	−0.487	19.948	138.986	2.120	296.432
		(0.764)	(2.178)	(2.506)	(12.443)	(5.777)	(79.039)
	10^5	5.240	−2.984	16.305	121.837	6.411	196.299
		(0.246)	(0.718)	(0.810)	(3.765)	(2.618)	(20.093)
	10^6	5.182	−3.273	15.867	121.629	7.495	186.459
		(0.078)	(0.228)	(0.258)	(1.222)	(0.936)	(6.202)
	DGP	5.104	−3.439	15.940	122.594	8.271	188.237
		(0.078)	(0.230)	(0.255)	(1.192)	(1.003)	(6.171)
With Specification Bias With Selection Bias (WS/WS)	10^3	3.282	−8.516	−0.600	93.612	53.209	44.022
		(6.462)	(13.996)	(25.349)	(107.437)	(318.034)	(1631.588)
	10^4	11.276	1.789	24.685	206.337	5.518	429.000
		(2.033)	(5.510)	(7.844)	(42.117)	(42.662)	(287.558)
	10^5	6.699	1.234	8.237	119.545	5.285	62.571
		(0.623)	(1.731)	(2.408)	(8.528)	(6.166)	(23.244)
	10^6	6.717	−1.048	10.047	126.443	2.305	81.643
		(0.203)	(0.579)	(0.800)	(2.927)	(0.564)	(9.791)
	DGP	6.540	−1.570	9.913	125.916	3.203	80.403
		(0.198)	(0.569)	(0.779)	(2.889)	(0.952)	(9.378)

Note. Bootstrap standard errors are in parenthesis for $n = 1,000 \sim 1,000,000$. For the DGP rows, we repeated the true DGP with $n = 1,000,000$ for 3,000 times, computed the corresponding bias statistics, and took the mean and standard deviations.

In table 2, we present (i) the normality test results, (ii) the 95% confidence interval using first-order asymptotic approximation, and (iii) 95% bootstrap confidence interval for different sample sizes. To simplify the presentation, we only focus on $\hat{\Psi}^B_{ATE}$ and $\hat{\Psi}^A_{ATE}$. The results for other statistics are qualitatively similar.

We employed Shapiro-Wilk test to test for the normality of our test statistics. For the bootstrap confidence interval, we used the confidence interval of the form $\left(\hat{\Psi} - q^*_n\left(1-\frac{\alpha}{2}\right), \hat{\Psi} - q^*_n\left(\frac{\alpha}{2}\right)\right)$ where q_n is the quantile of the centered bootstrap estimator $\hat{\Psi}^* - \hat{\Psi}$. For the DGP confidence interval column, we repeated the data generating process and estimation for 3,000 times, to calculate the $\frac{\alpha}{2}$'th and $1-\frac{\alpha}{2}$'th quantiles for corresponding sample size n. It turns out that there is not much gain in the bootstrap

confidence intervals, and the finite sample properties of the bootstrap confidence interval can be quite poor. Thus, we can conclude that we can simply employ the usual normal approximation confidence intervals, rather than computing the bootstrap confidence intervals.

As we noted before, $\hat{\Psi}^A_{ATE}$ has a quite poor finite sample properties, so that in all four specifications we can reject the normality hypothesis for $n = 10^4$ at the 5% significance level. Again, we get this result, because the $\hat{\Psi}^A$'s are the sample averages of the squared biases so that the finite sample distribution is likely to be skewed.

Table 3 compares the estimates of Δ_{ATE} by different estimation methods. Once again, to simplify our presentation, we only report the estimates of $\hat{\Delta}_{ATE}$, and omit the results for $\hat{\Delta}_{ATT}$ and $\hat{\Delta}_{ATN}$. We compare our estimates from the control function method, matching, IV (LATE), and OLS. The column labeled "True" is the true Δ_{ATE} for the given generated sample. Thus the performance of the estimators should be evaluated by comparing the estimates in each of the other columns with the True column. Note that it is possible to figure out the true Δ_{ATE} in our Monte-Carlo simulations where we observe all the counterfactuals. But, of course, this is not possible in real-world applications, because the counterfactuals are not observed.

Table 3 shows how it can be misleading when one ignores the presence of selection bias in estimating the average treatment effects. As the sample size gets large, the control function estimates always converge to the true average treatment effects, even when there exists specification bias. (WS/WS and WS/NS) The OLS column in NS/WS and the IV column in WS/WS seem to perform very well, but there is no good reason to believe that they work better. Instead, it is more likely that this result is just a coincidence. For instance, the OLS column for WS/WS and the IV column in NS/WS shows poor performance as an estimator for Δ_{ATE}.

For each specification in Table 1 the differences between the estimates of the matching estimator and the true average treatment effect are roughly the same as $\hat{\Psi}^B_{ATE}$. Hence, we verify that the decompositions (3.4) through (3.6) hold.

Table 2: Normality Test Results and Alternative 95% Confidence Intervals for the Estimated Selection Bias Parameters

	n	$\hat{\Psi}^B_{ATE}$				$\hat{\Psi}^A_{ATE}$			
		S-W p-value	Normal Approx. 95% CI	Bootstrap 95% CI	DGP 95% CI	S-W p-value	Normal Approx. 95% CI	Bootstrap 95% CI	DGP 95% CI
No Specification Bias No Selection Bias (NS/NS)	10^3	0.141	$(-3.109, 3.564)$	$(-3.140, 3.573)$	$(-3.340, 3.174)$	0.000	$(-5.070, 23.286)$	$(-24.708, 2.279)$	$(0.567, 21.676)$
	10^4	0.610	$(-0.505, 1.496)$	$(-0.514, 1.478)$	$(-1.000, 1.202)$	0.000	$(-0.753, 2.543)$	$(-2.691, 0.393)$	$(0.052, 2.181)$
	10^5	0.833	$(-0.617, 0.017)$	$(-0.613, 0.022)$	$(-0.315, 0.323)$	0.000	$(-0.072, 0.532)$	$(-0.285, 0.301)$	$(0.005, 0.215)$
	10^6	0.964	$(-0.080, 0.118)$	$(-0.082, 0.117)$	$(-0.099, 0.098)$	0.000	$(-0.007, 0.029)$	$(-0.027, 0.008)$	$(0.001, 0.021)$
With Specification Bias No Selection Bias (WS/NS)	10^3	0.000	$(-4.540, 2.906)$	$(-4.190, 3.116)$	$(-4.346, 5.321)$	0.000	$(-5.364, 25.993)$	$(-26.873, 3.232)$	$(1.059, 39.406)$
	10^4	0.728	$(-0.742, 1.986)$	$(-0.713, 2.022)$	$(-0.903, 1.792)$	0.000	$(-0.392, 4.077)$	$(-2.775, 1.555)$	$(0.173, 3.439)$
	10^5	0.915	$(-0.574, 0.264)$	$(-0.572, 0.271)$	$(0.023, 0.832)$	0.000	$(1.512, 0.366)$	$(0.133, 1.259)$	$(0.229, 1.000)$
	10^6	0.877	$(0.365, 0.627)$	$(0.361, 0.622)$	$(0.309, 0.570)$	0.000	$(0.351, 0.580)$	$(0.329, 0.552)$	$(0.374, 0.602)$
No Specification Bias With Selection Bias (NS/WS)	10^3	0.908	$(2.178, 11.923)$	$(2.023, 11.932)$	$(0.352, 9.843)$	0.000	$(48.556, 185.515)$	$(21.698, 156.072)$	$(64.671, 221.648)$
	10^4	0.757	$(5.531, 8.518)$	$(5.547, 8.532)$	$(3.534, 6.597)$	0.012	$(115.698, 164.474)$	$(112.934, 161.278)$	$(101.664, 148.553)$
	10^5	0.398	$(4.734, 5.731)$	$(4.746, 5.758)$	$(4.634, 5.585)$	0.454	$(114.386, 129.376)$	$(114.203, 129.290)$	$(115.396, 130.395)$
	10^6	0.665	$(5.029, 5.334)$	$(5.030, 5.334)$	$(4.948, 5.255)$	0.543	$(119.178, 123.970)$	$(119.241, 124.054)$	$(120.311, 124.940)$
With Specification Bias With Selection Bias (WS/WS)	10^3	0.000	$(-9.131, 15.596)$	$(-7.997, 16.628)$	$(-7.894, 20.978)$	0.000	$(-56.080, 324.696)$	$(-151.200, 143.695)$	$(50.879, 519.973)$
	10^4	0.148	$(7.304, 15.274)$	$(7.172, 15.209)$	$(2.445, 10.623)$	0.000	$(128.452, 293.552)$	$(109.596, 274.464)$	$(80.467, 195.146)$
	10^5	0.994	$(5.484, 7.926)$	$(5.484, 7.940)$	$(5.291, 7.748)$	0.000	$(103.383, 136.814)$	$(100.900, 135.172)$	$(108.921, 145.384)$
	10^6	0.593	$(6.317, 7.113)$	$(6.328, 7.122)$	$(6.148, 6.927)$	0.425	$(120.751, 132.224)$	$(120.552, 132.213)$	$(120.456, 131.669)$

Note. The Shapiro-Wilk test rejects the null hypothesis that the statistics are normally distributed when the p-value is smaller than the predetermined level.

Table 3: The Performance of Alternative Estimators of Δ_{ATE}

	n	$\hat{\Delta}_{ATE}$				
		Control	Matching	IV(LATE)	OLS	True
No Specification Bias No Selection Bias (NS/NS)	10^3	21.116	20.063	16.940	17.347	20.595
		(1.336)	(1.064)	(1.221)	(0.895)	(−)
	10^4	19.675	19.711	15.402	16.060	20.124
		(0.441)	(0.343)	(0.372)	(0.280)	(−)
	10^5	20.275	20.043	15.795	16.038	20.127
		(0.136)	(0.159)	(0.119)	(0.092)	(−)
With Specification Bias No Selection Bias (WS/NS)	10^3	13.148	12.285	12.243	12.209	11.635
		(2.086)	(0.695)	(0.696)	(0.631)	(−)
	10^4	11.282	11.337	11.195	11.330	11.601
		(0.739)	(0.226)	(0.235)	(0.209)	(−)
	10^5	11.433	11.663	11.720	11.641	11.594
		(0.224)	(0.070)	(0.072)	(0.064)	(−)
No Specification Bias With Selection Bias (NS/WS)	10^3	18.173	22.593	12.687	20.108	20.365
		(2.619)	(1.277)	(2.220)	(1.277)	(−)
	10^4	18.314	23.410	13.440	20.755	19.948
		(0.833)	(0.401)	(0.667)	(0.366)	(−)
	10^5	20.343	23.583	14.189	20.807	20.151
		(0.271)	(0.135)	(0.211)	(0.115)	(−)
With Specification Bias With Selection Bias (WS/WS)	10^3	16.124	15.875	10.787	15.841	11.406
		(6.432)	(1.076)	(1.603)	(0.976)	(−)
	10^4	8.577	15.906	10.192	15.917	11.425
		(2.071)	(0.339)	(0.500)	(0.315)	(−)
	10^5	13.008	16.408	11.731	16.404	11.618
		(0.632)	(0.109)	(0.159)	(0.097)	(−)

Note. The standard errors are in parentheses and are the bootstrap standard errors.

Based on the results of this section, we conclude that our test works quite well with reasonable sample sizes under various circumstances. In particular, our test works well even when there exists specification bias, which is likely to be common in many econometric applications.

5 An Empirical Example Using Data From a U.S. Job Training

5.1 Data and Specification

Here we apply our test of selection bias to data from a U.S. job training program. Our application involves a sample of about 2,200 adult women selected to the classroom training component of the National JTPA Study (NJS).[12] The JTPA program, the predecessor of the current Workforce Investment Act (WIA) program, provided a range of employment and training services to economically disadvantaged persons. The NJS was a social experiment conducted in 16 sites throughout the United States that tested the efficacy of JTPA services (for a detailed discussion on the JTPA program, see Doolittle and Traeger, 1990).[13] In this example, we focus on the post-program employment rates of adult NJS women, who upon applying for the program, were determined by program officials to likely benefit from classroom training. These women were then randomly assigned into a treatment or a control group. We present the baseline summary statistics in Table 4, separately for public aid recipients and for women who were

[12] A description of our data is contained in the appendix.

[13] Throughout this section, we did not include the site dummies in the covariates. We also tried to include the 15 site dummies in the covariates, which only resulted in a similar coefficient estimates but larger standard error estimates due to a larger number of covariates.

not receiving public aid at the baseline. The baseline is defined as the date of random assignment, which differs across calendar times for different participants.[14]

Table 4: Baseline Summary Statistics

	Public Aid Recipient at Baseline	No Public Aid Recipient at Baseline	Full Sample
Age	31.81	36.37	33.71
	(6.65)	(11.50)	(9.44)
Kids under 4 years old	0.43	0.20	0.34
	(0.50)	(0.40)	(0.47)
Married	0.13	0.33	0.22
	(0.34)	(0.47)	(0.41)
Never married	0.43	0.24	0.35
	(0.50)	(0.43)	(0.48)
Black	0.41	0.20	0.34
	(0.49)	(0.40)	(0.47)
Hispanic	0.14	0.10	0.12
	(0.35)	(0.29)	(0.34)
High school dropout	0.51	0.44	0.48
	(0.50)	(0.50)	(0.50)
Years of schooling	11.27	11.52	11.42
	(1.63)	(1.78)	(1.70)
Experiment Status	0.66	0.69	0.67
	(0.47)	(0.46)	(0.47)
Employment in Month 18	0.48	0.65	0.54
	(0.50)	(0.48)	(0.50)
n	1207	809	2216

(a) Mean Demographic Characteristics at the Baseline, Employment Rates During Month 18, and Experimental Impacts at Month 18

	Public Aid=1	Public Aid=0	Full Sample
$\Pr(D_i = 1 \mid z_i = 1)$	0.71	0.69	0.69
$\Pr(D_i = 1 \mid z_i = 0)$	0.41	0.42	0.41
$\Pr(y_i = 1 \mid z_i = 1)$	0.50	0.64	0.56
$\Pr(y_i = 1 \mid z_i = 0)$	0.43	0.66	0.52
$\Pr(z_i = 1)$	0.66	0.69	0.67

(b) Observed Conditional Probabilities of Receiving Training by Month 18, and Being Employed during Month 18

Note. z_i is the experimental status indicator at baseline, D_i is the training status indicator in month 18, y_i is the employment indicator in month 18.

Our analysis focuses on (i) women in the entire sample, and on (ii) the subsample of women who were receiving public aid at the baseline. The women we study here are from a very economically disadvantaged adult population. At the baseline, the mean age of these women is nearly 34, about 50 percent are high school dropouts, with a mean years of schooling equal to 11.5, and only about 10 percent were employed at the baseline. Even during the 18th month following the baseline, these employment rates, especially for those women who were on public aid (about 60% of the sample) continue to be relatively low (or about 43% compared to 66% for those who were not on public aid at the baseline).

[14]Pubic aid recipients are defined as women who reporting receiving either Food Stamps (now SNAP) or AFDC (now TANF) benefits at the baseline.

The Experimental Impacts and Control Group Substitution As Bloom et al. (1993) note, the experimental design appears to have "worked" as the baseline characteristics of the treatments and controls were balanced. The experimental evaluation indicated that adult women benefited from the opportunity to participate in such services through modestly increased earnings and employment rates (Bloom et al., 1993). As shown by Table 4b, the experimental impact of training was equal at about 4 percent, but there was about an 10 percentage point difference in the experimental impact for women who were receiving public aid at the baseline compared with those who were not.[15]

Because of the decentralized structure of federally supported training programs, there was substantial amount of control group substitution as well as many treatment no-shows or early dropouts. As training providers were subcontracted locally, it was not only possible for the control group members to get similar training as the treatments, but even to be enrolled in the same classroom under a different source of funding than the funding source for the treatments. As a result, there are substantial differences between the program's experimental impact (the intent-to-treat effect), the impact that takes account of the no-shows and early dropouts (the Bloom estimate), and the impact that takes account of both the no-shows and early dropouts and the control group substitution (Generalized Bloom estimate), (Heckman et al., 1999).

Turning to Table 4b, we observe that about 41% of those assigned to the treatment group were "always takers;" This means that 41% of the treatment would have participated in training even if they had been assigned to the control group. About 31% were "never takers;" this means that 31% of the treatments were no-shows or early dropouts. Invoking the monotonicity/uniformity assumption, we assume that these treatments would not have participated had they been assigned into the control group. Finally, only about 28% of the treatments were "compliers." Such people were induced by the experiment to take training and would otherwise not have participated.

We estimate the foregoing percentages by virtue of the random assignment indicator, z_i. The fraction of women who participate in training, whether they are assigned to the treatment or control groups, the "always takers," is given by $\Pr(D_i = 1|z_i = 0)$; $\Pr(D_i = 1|z_i = 1)$ is sum of the probabilities of "always takers" and the "compliers." Similarly, $\Pr(D_i = 0|z_i = 1) = 1 - \Pr(D_i = 1|z_i = 1)$ is the probability that a treatment is a "never taker."[16] These patterns are important below when we show that, among women in the public aid subsample, our ability to estimate the magnitude of the selection bias can be used to infer quite different program impacts on the "always takers" compared with the "compliers."

We begin our analysis of the training effect by presenting estimates based on four different estimators: (i) Matching; (ii) OLS; (iii) Linear IV and (iv) the semiparametric control function estimator. We use as the left-hand side variable $y_i \in \{0, 1\}$, employment status during the 18th month following the baseline. This corresponds to the linear probability model.

Calculating the Propensity Scores With our matching estimator we estimated several different versions of this model. We did this because as Heckman and Navarro-Lozano (2004) observe that "...

[15]This can be confirmed either (i) by subtracting conditional probabilities $\Pr(y_i = 1|z_i = 0)$ from $\Pr(y_i = 1|z_i = 1)$, or (ii) by investigating the OLS estimator $\hat{\beta}$ of the equation $y_i = \alpha + z_i\beta + u_i$ for each of the subsamples and the full sample, respectively. Here, $\hat{\beta}_{fullsample} = 0.04$, $\hat{\beta}_{pubaid=1} = 0.07$, and $\hat{\beta}_{pubaid=0} = -0.03$.

[16]Note that under the VIA, we assume that there were no defiers (i.e., the Monotonicity/Uniformity assumption holds).

matching offers no guidance as to which control variables to include ..." in the model.[17] Before estimating the alternative matching estimators, we have to check whether there is common support in the estimated propensity scores of the treatments and controls. Without common support, we cannot use the propensity score matching to estimate the treatment parameters.

We chose to estimate the propensity score $\Pr(D_i = 1|\mathbf{x}_i, z_i)$ using the covariate vector:

$$\begin{aligned}\mathbf{x}_i &= \{\text{age}_i,\ \text{squared age}_i,\ \text{kids under 4 years old}_i,\ \text{married}_i,\ \text{never married}_i,\\ &\quad \text{black}_i,\ \text{hispanic}_i,\ \text{high school dropout}_i,\ \text{years of schooling}_i\} \qquad (5.1)\end{aligned}$$

and the experimental status variable, $z_i \in \{0, 1\}$, where $D_i \in \{0, 1\}$ denotes program participation status. Because of the relatively small sample size, we used probit or logit to estimate the propensity score, instead of using nonparametric methods.

Table 5 presents the result of the propensity score estimation using Probit and Logit model, which gives a description what determines people's participation in the traning. Column (i) and (iii) used Probit model, column (ii) and (iv) used Logit model. Column (i) and (ii) are the result for the full sample, (iii) and (iv) are for the public aid subsample. It turns out that the variables Age, Squared age, Experimental status are significant at 1% significance level throughout (i)~(iv). The variable Black is negative and significant at 1% significance level for the full sample, while the variable Hispanic is positive and significant at 5% significance level for the public aid subsample. During the estimation of treatment effects using the control function and the matching estimators, we used the Probit model estimates of the propensity scores (column (i) and (iii) of Table 5). While the coefficient estimates of the probit and logit model estimates show some difference, the method that we used to estimate the propensity score had little effect on the estimated treatment effects both in the matching estimators and the control function estimators.

As shown by Figure 5.1, the propensity scores for program participants and program non-participants approximately overlap. The support of the propensity scores for program participants and program non-

[17] For this, we tried the following combinations for the independent variables and experimental status z_i:

$$\begin{aligned}\mathbf{x}_i &= \{\text{age}_i,\ \text{squared age}_i, \#\text{ kids under 4 years old}_i,\ \text{married}_i,\ \text{never married}_i,\\ &\quad \text{black}_i,\ \text{hispanic}_i,\ \text{high school dropout}_i,\ \text{years of schooling}_i\}\end{aligned}$$

$$\begin{aligned}(\mathbf{x}_i, z_i) &= \{\text{age}_i,\ \text{squared age}_i, \#\text{ kids under 4 years old}_i,\ \text{married}_i,\ \text{never married}_i,\\ &\quad \text{black}_i,\ \text{hispanic}_i,\ \text{high school dropout}_i,\ \text{years of schooling}_i,\ \text{experimental status}_i\}\end{aligned}$$

$$\begin{aligned}\mathbf{x}_i &= \{\text{age}_i,\ \text{squared age}_i, \#\text{ kids under 4 years old}_i,\ \text{married}_i,\ \text{never married}_i,\\ &\quad \text{black}_i,\ \text{hispanic}_i,\ \text{high school dropout}_i,\ \text{years of schooling}_i,\\ &\quad \text{and the interaction terms of the variables above}_i\}\end{aligned}$$

$$\begin{aligned}(\mathbf{x}_i, z_i) &= \{\text{age}_i,\ \text{squared age}_i, \#\text{ kids under 4 years old}_i,\ \text{married}_i,\ \text{never married}_i,\\ &\quad \text{black}_i,\ \text{hispanic}_i,\ \text{high school dropout}_i,\ \text{years of schooling}_i,\\ &\quad \text{and the interaction terms of the variables above}_i\ \text{experimental status}_i\}\end{aligned}$$

In our third matching estimator, we include higher order terms and their interactions, without experimental status, to estimate the propensity score. Finally, for our last matching estimators follows the third estimator, but also includes experimental status. We found that the second specification worked the best among the four specifications that we tried.

Table 5: Result of Propensity Score Estimation

$\Pr(D_i = 1 \mid \mathbf{x}_i, z_i)$	(i)	(ii)	(iii)	(iv)
Age	0.077	0.129	0.106	0.174
	(0.020)	(0.034)	(0.030)	(0.051)
Squared age	−0.001	−0.002	−0.002	−0.002
	(0.000)	(0.000)	(0.000)	(0.001)
Kids under 4 years old	0.007	0.011	−0.052	−0.086
	(0.068)	(0.111)	(0.129)	(0.213)
Married	0.031	0.047	0.116	0.192
	(0.075)	(0.123)	(0.112)	(0.186)
Never married	−0.085	−0.143	−0.070	−0.122
	(0.072)	(0.118)	(0.136)	(0.224)
Black	−0.299	−0.487	−0.110	−0.173
	(0.068)	(0.112)	(0.124)	(0.204)
Hispanic	0.129	0.215	0.350	0.598
	(0.093)	(0.155)	(0.174)	(0.297)
Other minor race	0.187	0.324	0.168	0.302
	(0.166)	(0.280)	(0.225)	(0.382)
High schol dropout	−0.088	0.009	0.025	0.035
	(0.074)	(0.037)	(0.124)	(0.206)
Years of schooling	0.007	0.009	0.018	0.030
	(0.022)	(0.037)	(0.035)	(0.059)
Experimental status	0.702	1.143	0.658	1.081
	(0.059)	(0.096)	(0.100)	(0.164)
(Intercept)	−1.322	−2.182	−2.065	−3.390
	(0.476)	(0.792)	(0.745)	(1.251)
n	2216	2216	1207	1207

Note. Standard errors are in parentheses. $\mathbf{x}_i$ is the covariate vector, z_i is the experimental status indicator at baseline, D_i is the participation status indicator by month 18. Column (i) is full sample with Probit model, (ii) is full sample with Logit model estimation. (iii) is public aid subsample with Probit model, (iv) is public aid subsample with Logit model estimation.

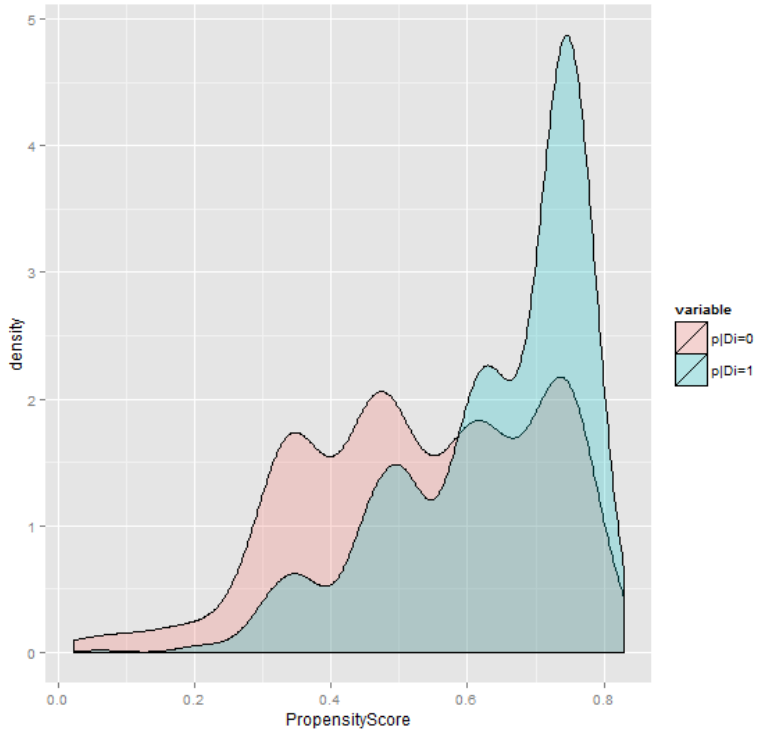
participants overlap on the interval $[0.2, 0.83]$.[18] Thus, there is a common support for those women in the the participant and nonparticipant groups. We find that all four approaches generate propensity scores of which the supports overlap quite well. This indicates that the nonparticipant group is comparable to our participant group. For all four approaches we tried, the mean propensity score for the participants is about 12 percentage points greater than the mean propensity score for the nonparticipant group.

Alternative Estimators

Matching specification As shown by Table 6a and 6b, after we adjust for these differences in propensity scores, we find that the impact of training (ATE) estimated using the matching estimator equals -0.01 in the full sample and -0.05 in the public aid subsample. The other matching impacts for ATT and ATN are similar. These results are insensitive to which of the four approaches we described in the above footnote to calculate the propensity score.

[18] We find only about 3% of the nonparticipants and 0.3% of the participants have propensity scores between [0, 0.2]. These percentages are negligible. Also, for both participants and nonparticipants, the maximum values of the propensity scores are 0.83.

Figure 5.1: Density of Propensity Scores by Participation Status D_i for the Full Sample of JTPA-CT Adult Women



OLS specification We specify the OLS specification as:

$$y_i = \mathbf{x}_i'\beta + D_i\delta + \epsilon_i \tag{5.2}$$

where the covariates $\mathbf{x}_i$ is the same as specified in (5.1), $y_i \in \{0, 1\}$ is the employment status in month 18. We used the linear probability model, which is consistent if conditional independence assumptions hold.

The least squares estimates are very similar to the propensity score matching estimates. As shown by Table 6a and 6b, after we adjust for the covariates, we find that the impact of training equals -0.01 in the full sample and -0.03 in the public aid subsample. Note that although least squares estimates appear to be estimates of ATT, they differ from the matching and control function estimates of ATT, because of different weights that they place on individuals' marginal treatment effects (See Heckman and Vytlacil, 2005).

IV specification The linear IV estimator we used is given by estimating (5.2) using the experimental assignment status z_i as the instrument for D_i. Notice that we use a woman's experimental status as an instrumental variable. Because we don't have perfect compliance, this IV estimator consistently estimates the local average treatment effect for those women who complied with the treatment (cf., Imbens and Angrist, 1994). As shown in Table 6a and Table 6b, when we run the IV regression with using the experimental status as the instruments for the training status, we find that the impact on the compliers is 0.148 in the full sample and 0.226 in the public aid subsample. That is, for the complier group in the NJS, training was effective, because it raised the probability of being employed, during the 18 month

following the baseline, by 14.8% and 22.6%, respectively.

Control Function Estimation Specification Finally, our control function specification is given by the following system of equations:

$$\begin{aligned} D_i &= \mathbf{1}\left(\mathbf{x}_i'\gamma + z_i\zeta + \eta_i\right) \\ y_{0,i} &= \mathbf{x}_i'\beta^0 + \epsilon_{0,i} \\ y_{1,i} &= \mathbf{x}_i'\beta^1 + \epsilon_{1,i} \end{aligned}$$

where the treatment status is $D_i \in \{0,1\}$, the experimental status is $z_i \in \{0,1\}$, the employment status during month 18 for the people who are not treated is given by $y_{0,i} \in \{0,1\}$, for who are treated is given by $y_{1,i} \in \{0,1\}$ and $\mathbf{1}(.)$ is the indicator function. Notice again, we use the women's experimental status z_i as an instrumental variable in constructing the control function estimates of the impact of training. In the estimation, we used the same method as in the Section 4.1, except for taking only the first degree polynomial of the propensity score for the control function. We tried to use the quadratic version of propensity score to estimate the control function, but found that efficiency loss is too large, so we only used the linear version. This is because of the relatively small sample size of the data.

5.2 Results of the National JTPA Study Estimation

Table 6 presents the results of the alternative estimations discussed in the previous subsection. First, as expected, the standard errors associated with the OLS, matching, and even the IV estimators are substantially smaller than those associated with the control function estimators. But, because the OLS, matching, and IV estimators may be biased, we should first check that whether the corresponding selection bias estimates are statistically close to zero. In the table we observe that the point estimates of the ATT are each close to zero for OLS, matching, and the control function estimates. The bias terms for the ATT also are close to zero, These facts indicate that we can use the more efficient matching or the OLS estimates of the impact of training on participants.

We summarize the key results of Table 6 in Figure 5.2. This figure also illustrates how the decomposition in (3.4) through (3.6) presented in Section 3 works in our example. The first bars in each group are the matching estimator, which consistently estimate $E[y_{1,i} - y_{0,i}]$ when the CIA holds. When this happens, this would imply that the first set of bars in the figure would coincide with the second bars for each grouping. However, this is not the case in our example. If we use matching to estimate the ATE, ATT, and ATN, we find that all three of these treatment effects are weakly negative.

The control function estimators tell a different story about the impact of training. We find that $\hat{\Delta}_{ATE}$ and $\hat{\Delta}_{ATN}$ are both positive in the public aid subsample, and $\hat{\Delta}_{ATN}$ is large, positive, and statistically significant at the 5% level in both the public aid subsample and in the full sample. The third bars in each group are the biases associated with each of the estimated training effects. Even though the treatment itself has a positive effect on women's employment rates on average (ATE) and on the people who are not treated (ATN), selection makes it appear that the training effects are weakly negative when the selection biases are ignored.

The fourth bars in each group is simply the sum of the second and third bars. We confirm that the

sign and the magnitude of the fourth bars coincide with the first bars. We present this fourth group of bars in order to demonstrate that the decomposition (3.4)~(3.6) works in "real-world" example. The first bars in Figure 5.2 correspond to the left-hand side of (3.4) through (3.6), or to $E\left[y_{1,i} - y_{0,i}\right]$, whereas the fourth set of bars correspond to the right-hand side of these expressions, or to $\Delta + \Psi$.

Finally, we observe that the magnitudes of $\sqrt{\hat{\Psi}^A}$'s are quite similar to that of $\left|\hat{\Psi}^B\right|$'s.[19] However, the t-values for the $\hat{\Psi}^A$'s turn out to be smaller than those of the corresponding $\hat{\Psi}^B$'s. We attribute the higher standard errors associated with the $\hat{\Psi}^A$'s to the squared terms in this expression, which is a convex function.

The result that $\hat{\Delta}_{ATN}$ is highly positive while $\hat{\Delta}_{ATT}$ is close to zero suggests that the wrong people received training. We arrive at this conjecture, because we are able to infer the effect of training on the "always takers." If we think of the linear IV estimate as an unbiased estimate of the effect of training on the "compliers," and our control function ATT estimate as an unbiased estimate of the effect of training on all trainees, we can infer that training harmed the employment prospects of the "always takers." By using the proportion of "compliers" among all trainees, we infer that the effect of training on the "always takers" was equal to about -0.1. That is, training appears to lower the probability of being employed at month 18 by about 10 percentage points for the "always takers." The training impacts for the "compliers", the "never takers", and the "always takers" are equal to 0.23, 0.42, and -0.12 respectively, in the public aid subsample. This finding makes more concrete our contention that the wrong people received training.[20]

[19] $\left(\sqrt{\hat{\Psi}^A_{ATE}}, \sqrt{\hat{\Psi}^A_{ATT}}, \sqrt{\hat{\Psi}^A_{ATN}}\right) = (0.2086, 0.0529, 0.3586)$ for the public aid subsample, and $\left(\sqrt{\hat{\Psi}^A_{ATE}}, \sqrt{\hat{\Psi}^A_{ATT}}, \sqrt{\hat{\Psi}^A_{ATN}}\right) = (0.1204, 0.01, 0.2093)$ for the full sample.

[20] We arrive at this inference by solving for Δ_{Always} in: $\Delta^{Control}_{ATT} = \Delta^{IV}_{LATE}\frac{s_C}{s_C+s_A} + \Delta_{Always}\left(1 - \frac{s_C}{s_C+s_A}\right)$, where Δ_{Always} is the impact of training for the always takers, s_C is the share of compliers, s_A is the share of the always takers. We observe from Table 5b that $s_C = 0.28$, $s_A = 0.41$. This result is robust in both the public aid subsample, which is equal to -0.12, and the full sample, with both public aid statuses, which is equal to -0.10.

Table 6: Alternative Estimates of the Impact of Training on Employment Rates and Estimates of Selection Bias

(a) Impacts on Employment Rates During Month 18, Public Aid Subsample, $n = 1207$

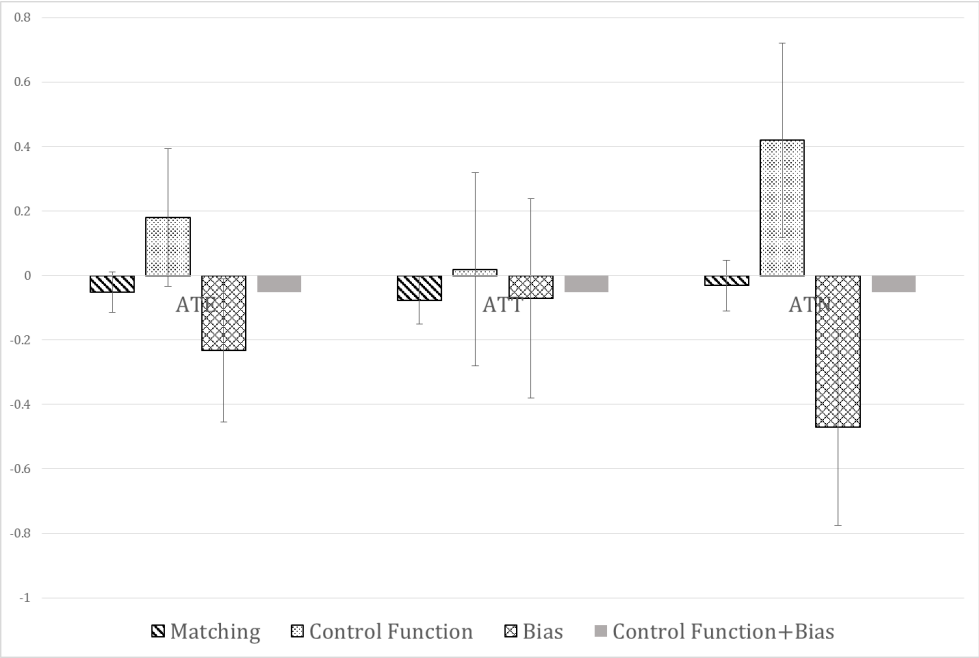
	$\hat{\Psi}^B_{ATE}$	$\hat{\Psi}^B_{ATT}$	$\hat{\Psi}^B_{ATN}$	$\hat{\Psi}^A_{ATE}$	$\hat{\Psi}^A_{ATT}$	$\hat{\Psi}^A_{ATN}$	$\hat{\Delta}^{IV}_{LATE}$	$\hat{\Delta}^{OLS}$	$\hat{\Delta}^{Matching}_{ATE}$	$\hat{\Delta}^{Matching}_{ATT}$	$\hat{\Delta}^{Matching}_{ATN}$	$\hat{\Delta}^{Control}_{ATE}$	$\hat{\Delta}^{Control}_{ATT}$	$\hat{\Delta}^{Control}_{ATN}$
Full Subsample	−0.2324	−0.0714	−0.4713	0.0435	0.0028	0.1286	0.2256	−0.0337	−0.0539	−0.0777	−0.0309	0.1799	0.0189	0.4189
Bootstrap Mean	−0.2355	−0.0735	−0.4764	0.0591	0.0168	0.1469	0.2289	−0.0334	−0.0513	−0.0663	−0.0300	0.1832	0.0260	0.4281
Bootstrap SE	0.1167	0.1576	0.1584	0.0350	0.0252	0.0917	0.1118	0.0292	0.0323	0.0375	0.0402	0.1119	0.1529	0.1536
t-value	−1.9904	−0.4529	−2.9755	1.2440	0.1122	1.4029	2.0178	−1.1525	−1.6649	−2.0700	−0.7686	1.6079	0.1239	2.7269

(b) Impacts on Employment Rates During Month 18, Full Sample, $n = 2213$

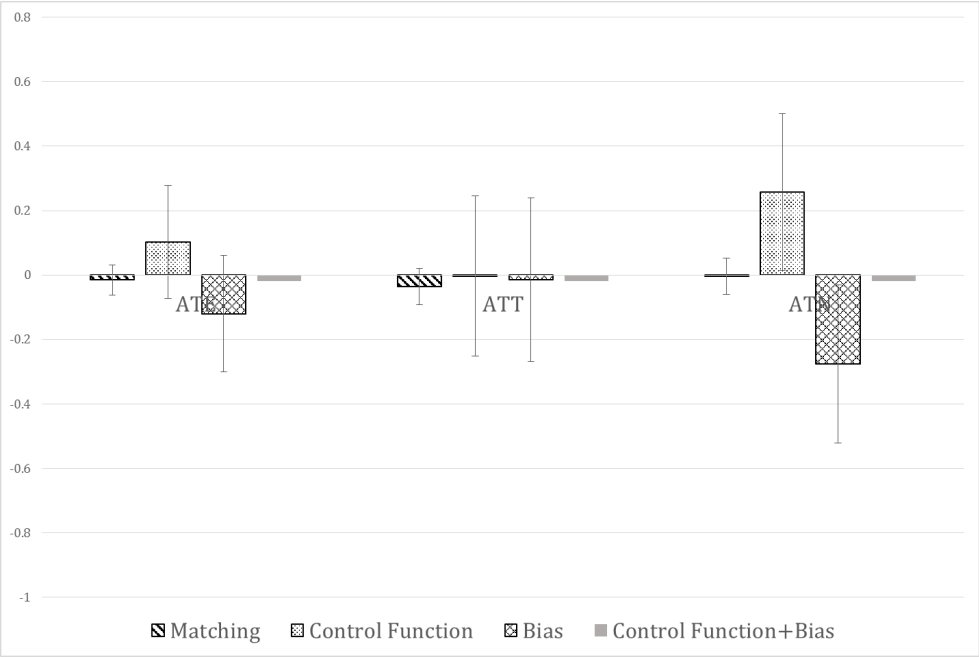
	$\hat{\Psi}^B_{ATE}$	$\hat{\Psi}^B_{ATT}$	$\hat{\Psi}^B_{ATN}$	$\hat{\Psi}^A_{ATE}$	$\hat{\Psi}^A_{ATT}$	$\hat{\Psi}^A_{ATN}$	$\hat{\Delta}^{IV}_{LATE}$	$\hat{\Delta}^{OLS}$	$\hat{\Delta}^{Matching}_{ATE}$	$\hat{\Delta}^{Matching}_{ATT}$	$\hat{\Delta}^{Matching}_{ATN}$	$\hat{\Delta}^{Control}_{ATE}$	$\hat{\Delta}^{Control}_{ATT}$	$\hat{\Delta}^{Control}_{ATN}$
Full Sample	−0.1207	−0.0153	−0.2759	0.0145	0.0001	0.0438	0.1485	−0.0097	−0.0122	−0.0357	−0.0045	0.1023	−0.0030	0.2576
Bootstrap Mean	−0.1202	−0.0121	−0.2794	0.0242	0.0094	0.0541	0.1499	−0.0097	−0.0120	−0.0218	0.0017	0.1021	−0.0062	0.2600
Bootstrap SE	0.0918	0.1287	0.1247	0.0167	0.0130	0.0430	0.0871	0.0219	0.0232	0.0287	0.0288	0.0890	0.1267	0.1244
t-value	−1.3138	−0.1188	−2.2128	0.8682	0.0101	1.0183	1.7038	−0.4437	−0.5248	−1.2425	−0.1552	1.1495	−0.0240	2.0709

Figure 5.2: Estimates of the Effects of Training on Employment Rates 18 months After the Baseline: Matching and Control Function Estimates, Selection Bias Estimates, and 95% Normal Confidence Intervals

(a) Estimates of the Impact of Training on Employment Rates During Month 18, Public Aid Subsample



(b) Estimates of the Impact of Training on Employment Rates During Month 18, Full Sample



6 Concluding Remarks

This paper uses the control function to construct estimates of the magnitude of the selection bias, and to test for the presence of selection bias in program evaluations. The idea is that given the VIA assumptions hold, the control function can separately identify alternative treatment effects from the magnitude of the selection bias. Such selection bias estimates allow us to test whether matching estimates of the ATE, ATT, and ATN are biased. These tests allow evaluators to consider more efficient estimators of the treatment effect, rather than relying on those based on the control function.

Our Monte-Carlo simulations indicate that our tests for selection bias do not require unreasonably large samples and can be implemented using sample sizes commonly available in many datasets. This is especially true of our bias parameter, the mean effective selection bias parameter, which performed well even in samples as small as 1,000 observations. In order to perform as well, the other bias parameter, the mean squared selection bias parameter, which can be used to test for violations of the CIA, appears to require larger samples - on the order of 10,000 individuals. We believe this discrepancy between our two bias estimators likely results from greater weight given to outliers, because of the squaring of the individual bias terms. We also find that (i) our test for selection bias was robust to the model specification error and that (ii) the standard error estimates provide evidence that bootstrapped standard errors are consistent.

The power of our test for selection bias depends on the efficiency of the control function estimators. More efficient estimation yields higher power. In our empirical example that used data for women assigned to the classroom training component of the National JTPA study, we show that the test was powerful enough to be implemented with a relatively modest sample size. If the null hypothesis of no selection bias is not rejected, an evaluation can be improved by using a more efficient estimator, such as one based on matching or least squares.

Although this National JTPA study was a randomized social experiment, the intent-to-treat effects are likely far off from the ATT or ATE, because of the high rate of control group substitution and treatment no-shows and early dropouts. Accordingly, it is of interest to know the impact of training on the participants, instead of just the impact of the treatment assignment on the assignees. We find that for the public aid subsample, the control function estimates of ATT, matching estimates of ATT, and the OLS estimates produce similar and small estimates of the training effect. None of them is statistically significantly different from zero at conventional levels of statistical insignificance. Moreover our measures of bias indicate that there is little evidence that our ATT estimates are biased. However, our LATE estimate for the compliers is large and positive. Because the participants consist of the compliers and the always takers, these findings suggest another benefit from systematically comparing the different estimates using our method: namely to infer the program impacts for the always takers. We find evidence here that many of the people who received the training were in fact harmed by their training experiences.

The assumptions justifying the use of the control function are the same as those which justify the use of IV, and weaker than the assumptions used to justify matching or OLS to consistently estimate the average treatment effect. Given this fact, we suggest a strategy for conducting a non-experimental evaluation as follows. Firstly, check to ensure that the support for the participants' and nonparticipants' propensity scores for the participants overlap. Secondly, use the control function to estimate alternative program impacts and the selection bias functions. Next, this then will tell us whether the evaluation can

use more efficient estimators such as matching or OLS. Finally, the evaluator can then use the instrument used to identify the control functions to estimate the LATE for the program compliers, and then use the ATT estimate to infer the program's impact on the always takers.

A Data Appendix

A.1 Design of the Social Experiment

The National JTPA Study (NJS) was a social experiment that examined how the availability of JTPA services affected individuals' labor market outcomes and their use of the social welfare system (Kemple et al., 1993). (For a discussion of the problems associated with implementing the NJS in 16 sites of nearly 600 SDAs nationwide, See Doolittle and Traeger (1990)). Because JTPA often provides multi-stage services to its participants, the study did not evaluate the impact of a specific JTPA service such as classroom training or job search assistance. Instead, it examined the impact of JTPA services on the employment histories of three different groups. The definition of these groups was based on the training plans that the SDA's administrators devised when potential participants applied for the program. Persons in the first group studied were recommended to received CT and possibly other services, but not OJT. Those in the second group studied were recommended to receive OJT and other services, but not CT. Finally, those in the third group studied were recommended to receive other combinations of services including both CT and OJT. However, most persons in this latter category received only job search assistance. The women in the first group, those recommended for CT, are those who are used in our example in Section 5 of the paper.

Applicants to the program were randomly assigned to the treatment group or the control group after program officials verified their eligibility and assigned them to one of their training plans described in the previous paragraph. The treatments were eligible to receive JTPA-sponsored training and services. Although those assigned to the control group were prevented from receiving JTPA sponsored training, they could receive other training services available in their communities. Members of the control group sometimes received a few hours of job search assistance as part of the JTPA screening process. In addition, they were given a list of non-JTPA programs operating in their communities, but they received no referrals. This list of other training alternatives included the Employment Service, local community colleges, technical institutes, and community agencies that provide social services. Because the controls sometimes received other non-JTPA training services, the NJS study examined the "marginal" impact of access to JTPA in the sites included in the study.

A.2 Description of Our Dataset

The JTPA data used in this study comes from participants' Background Information Form (BIF) and from two follow-up surveys. The BIF provides information about participants' characteristics at the baseline: the time when they were randomly assigned into either the treatment or control groups. The BIF provided information on participants' age, race, month of random assignment, schooling, work and social welfare histories, children in the household, marital status, site, experimental status, and recommended service strategy. The follow-up surveys provided information on participants' self-reported monthly employment and training status. Our local labor market data is from the U.S. Department of Labor and various monthly issues of the U.S. DOL Employment and Earnings.

Our sample includes all adult women recommended to receive classroom training and other services, but not on-the-job training. We excluded from our sample, women whose employment status was missing on the BIF or who did not have a continuous monthly employment history covering at least the first

18 months after the baseline. We also excluded women who reached age 60 anytime during the sample period. In cases in which the information about employment status at the baseline was not the same on the BIF as on the follow-up surveys, we used the information from the BIF. Our research sample included 2,671 women. When we exclude women with missing information on marital status, schooling, and number of children under 5 years old, we are left with a sample of 2,213 women.

A.3 Training Received by Treatments and Controls

As indicated by Table 4b in the text, both treatments and controls received some training: About two-third of the treatments participated in training, compared with about 40 percent of the controls. According to the follow-up data, the average time spent in training was about 4 months and the time spent in training given that a women participated was about 20 hours per week. By the 18th month after the baseline, the typical women who participated in training had been out of training for more than 10 months.

A.4 The Explanatory Variables

The following are the definitions of the explanatory variables used in the least squares, IV, control function and propensity score estimations:

- Experimental Status: =1 if the woman had access to JTPA-CT, =0 otherwise.
- Schooling: =highest grade of schooling completed.
- Nodegree: =1 if no high-school degree, =0 otherwise.
- Kids Under 4: =1 if children under 4 years old live with the woman at the baseline, =0 otherwise.
- Never Married: =1 if the woman had never been married at the baseline, =0 otherwise.
- Married: =1 if the woman is married with spouse present at the baseline, =0 otherwise.
- Black: =1 if the woman is black/nonhispanic, =0 otherwise.
- Hispanic: =1 if the woman is Hispanic, =0 otherwise.
- Age: = woman's age in years.

References

Aakvik, Arlid, James J. Heckman, and Edward J. Vytlacil, "Estimating treatment effects for discrete outcomes when responses to treatment vary: an application to Norwegian vocational rehabilitation programs," *Journal of Econometrics*, 2005, *Vol. 125*, 15–51.

Abadie, Alberto and Guido W. Imbens, "Large sample properties of matching estimators for average treatment effects," *Econometrica*, 2006, *Vol. 74, No. 1*, 235–267.

Ahn, Hyungtaik and James L. Powell, "Semiparametric estimation of censored selection models with a nonparametric selection mechanism," *Journal of Econometrics*, 1993, *Vol. 58*, 3–29.

Basu, Anirman, James J. Heckman, Salvador Navarro-Lozano, and Sergio Urzua, "Use of instrumental variables in the presence of heterogeneity and self-selection: An application to the treatment of breast cancer patients," *Health Economics*, 2007, *Vol. 16*, 1133–1157.

Bifulco, Robert, "Addressing self-selection bias in quasi-experimental evaluations of whole-school reform: A commonparison of methods," *Evaluation Review*, 2002, *Vol. 26, No. 5*, 545–572.

Bloom, Howard S., Larry L. Orr, Stephen H. Bell, and Fred Doolittle, *The National JTPA Study: Title II-A Impacts on Earnings and Employment at 18 Months*, Bethesda, MD.: Abt Associates, 1993.

Carneiro, Pedro, James J. Heckman, and Edward Vytlacil, "Evaluating marginal policy changes and the average effect of treatment for individuals at the margin," *Econometrica*, 2010, *Vol. 78, No.1*, 377–394.

_ , Karsten T. Hansen, and James J. Heckman, "Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice," *International Economic Review*, 2003, *Vol. 44, No. 2*, 361–422.

Cuddeback, Gary, Elizabeth Wilson, John G. Orme, and Terri Combs-Orme, "Detecting and statistically correcting sample selection bias," *Journal of Social Service Research*, 2004, *Vol. 30, No. 3*, 19–33.

Das, Mitali, Whitney K. Newey, and Francis Vella, "Nonparametric estimation of sample selection models," *The Review of Economic Studies*, 2003, *70*, 33–58.

DasGupta, Anirban, *Asymptotic theory of statistics and probability*, Springer, 2008.

Doolittle, Fred and Linda Traeger, *Implementing the National JTPA Study*, New York, NY: Manpower Demonstration Research Corporation, 1990.

Florens, J. P., C. Meghir J. J. Heckman, and E. Vytlacil, "Identification of treatment effects using control functions in models with continuous, endogenous treatment and hetrogeneous effects," *Econometrica*, 2008, *Vol. 76, No. 5*, 1191–1206.

Frolich, Markus, "Nonparametric IV estimation of local average treatment effects with covariates," *Journal of Econometrics*, 2007, *Vol. 139*, 35–75.

Hahn, Jinyong, "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 1998, *Vol. 66, No. 2*, 315–331.

_ **and Geert Ridder**, "Asymptotic variance of semiparametric estimators with generated regressors," *Econometrica*, 2013, *81*, 315–340.

Heckman, James and Salvador Navarro-Lozano, "Using matching, instrumental variables, and control functions to estimate economic choice models," *The Review of Econoics and Statistics*, 2004, *Vol. 88, No. 1*, 30–57.

_ **, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd**, "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 1998, *Vol. 66, No. 5*, 1017–1098.

Heckman, James J., "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models," *Annals of Economic and Social Measurement*, 1976, *5*, 475–492.

_ , "Sample Selection Bias as a Specification Error," *Econometrica*, 1979, *Vol. 47, No. 1*, 153–161.

_ **and Bo E. Honore**, "The Empirical Content of the Roy Model," *Econometrica*, 1990, *Vol. 58, No. 5*, 1121–1149.

_ **and Edward J. Vytlacil**, "Local instrumental variables and latent variable models for identifying and bounding treatment effects," *Proceedings of the National Academy of Sciences*, 1999, *Vol. 96*, 4730–4734.

_ **and** _ , "The relationship between treatment parameters within a latent variable framework," *Economics Letters*, 2000, *Vol. 66*, 33–39.

_ **and Edward Vytlacil**, "Policy-relevant treatment effects," *The American Economic Review Papers and Proceedings*, 2001, *Vol. 91, No. 2*, 101–111.

_ **and** _ , "Structural equations, Treatment effects, and Econometric Policy Evaluation," *Econometrica*, 2005, *Vol. 73, No.3*, 669–738.

_ **and V. Joseph Hotz**, "Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training," *Journal of the American Statistical Association*, 1989, *Vol. 84, Issue 408*, 862–874.

_ **, Robert J. LaLonde, and Jeffrey A. Smith**, *The economics and econometrics of active labor market programs*, Vol. 3 of *The Handbook of Labor Economics*, Amsterdam: Elsevier, 1999.

Heckman, James, Justin L. Tobias, and Edward Vytlacil, "Simple Estimators for Treatment Parameters in a Latent-Variable Framework," *The Review of Economics and Statistics*, 2003, *Vol. 85, No. 3*, 748–755.

Imbens, Guido W. and Jeffrey M. Wooldridge, "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 2009, *Vol 47, No. 1*, 5–86.

_ **and Joshua D. Angrist**, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 1994, *Vol. 62, No. 2*, 467–475.

Imbens, Keisuke Hirano Guido W. and Geert Ridder, "Efficient estimation of average treatment effects using the eestimate propensity score," *Econometrica*, 2003, *Vol. 71, No. 4*, 1161–1189.

Kemple, James J., Fred Doolittle, and John W. Wallace, *The National JTPA Study: Site Characteristics and Participation Patterns*, New York, NY: Manpower Demonstration Research Corporation, 1993.

Krueger, Alan B., "Experimental estimate of education production functions," *The Quarterly Journal of Economics*, 1999, *Vol. 114, No. 2*, 497–532.

LaLonde, Robert J., "Evaluating the econometric evaluations of training programs with experimental data," *The American Economic Review*, 1986, *Vol. 76, No. 4*, 604–620.

Lee, Lung-Fei, "Some approaches to the correction of selectivity bias," *Review of Economic Studies*, 1982, *49*, 355–372.

Melino, Angelo, "Testing for Sample Selection Bias," *Review of Economic Studies*, 1982, *49* (1), 151–153.

Newey, Whitney K., "Two-step series estimation of sample selection models," *Econometrics Journal*, 2009, *Vol. 12*, 217–219.

_ **, James L. Powell, and James R. Walker**, "Semiparametric estimation of selection models: Some empirical results," *The American Economic Review Papers and Proceedings*, 1990, *Vol. 80, No. 2*, 324–328.

Olsen, Randall J., "A Least Squares Correction for Selectivity Bias," *Econometrica*, 1980, *48* (7), 1815–1820.

Rosenbaum, Paul R., "Model-Based Direct Adjustment," *Journal of the American Statistical Association*, 1987, *Vol. 82, No. 398*, 387–394.

_ **and Rubin Donald B**, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 1983, *Vol. 70, No. 1*, 41–55.

Shao, Jun and Dongsheng Tu, *The Jackknife and bootstrap*, Springer, 1995.

Vella, Francis, "Estimating Models with Sample Selection Bias: A Survey," *The Journal of Human Resources*, 1998, *Vol. 33, No. 1*, 127–169.

Vytlacil, Edward, "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 2002, *Vol. 70, No. 1*, 331–341.