Black, Dan A.; Joo, Joonhwi; LaLonde, Robert J.; Smith, Jeffrey A.; Taylor, Evan J.

Working Paper

Simple Tests for Selection Bias: Learning More from Instrumental Variables

IZA Discussion Papers, No. 9346

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Black, Dan A.; Joo, Joonhwi; LaLonde, Robert J.; Smith, Jeffrey A.; Taylor, Evan J. (2015) : Simple Tests for Selection Bias: Learning More from Instrumental Variables, IZA Discussion Papers, No. 9346, Institute for the Study of Labor (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/120996

DISCUSSION PAPER SERIES

IZA DP No. 9346

Simple Tests for Selection Bias: Learning More from Instrumental Variables

Dan A. Black
Joonhwi Joo
Robert LaLonde
Jeffrey A. Smith
Evan J. Taylor

September 2015

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Simple Tests for Selection Bias: Learning More from Instrumental Variables

Dan A. Black
University of Chicago, IZA and NORC

Jeffrey A. Smith
University of Michigan, NBER and IZA

Joonhwi Joo
University of Chicago

Evan J. Taylor
University of Michigan

Robert LaLonde
University of Chicago and IZA

Discussion Paper No. 9346
September 2015

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. 9346
September 2015

ABSTRACT

Simple Tests for Selection Bias: Learning More from Instrumental Variables*

We provide simple tests for selection on unobserved variables in the Vytlacil-Imbens-Angrist framework for Local Average Treatment Effects. The tests allow researchers not only to test for selection on either or both of the treated and untreated outcomes, but also to assess the magnitude of the selection effect. The tests are quite simple; undergraduates after an introductory econometrics class should be able to implement these tests. We illustrate our tests with two empirical applications: the impact of children on female labor supply from Angrist and Evans (1998) and the training of adult women from the Job Training Partnership Act (JTPA) experiment.

JEL Classification: C10, C18, J01, J08

Keywords: local average treatment effects, selection, instrumental variables

Corresponding author:

Dan A. Black
Harris School
University of Chicago
1155 East 60th Street
Chicago, IL 60637
USA
E-mail: danblack@uchicago.edu

* We thank seminar participants at the University of Chicago, the University of Illinois at Chicago, and the Hitotsubashi Institute for Advanced Study, along with Josh Angrist, Ben Feigenberg, and Darren Lubotsky and Josh Angrist, for helpful comments.

1. Introduction

In the 20 years since the publication of Imbens and Angrist (1994), applied researchers have embraced the interpretation of Instrumental Variables (IV) estimators, particularly with binary instruments, as measuring the impact of treatment on the subset of respondents who comply with the instrument, which Imbens and Angrist term a Local Average Treatment Effect, or LATE. The LATE framework allows researchers to consistently estimate models in which individuals may differ in the effects of treatment. But the LATE framework comes with some costs. First, the LATE approach requires that we assume that instruments have a "monotonic" impact on behavior. Informally, instruments must induce all agents to behave in a weakly uniform manner when subjected to a change in the value of the instrument. Thus, if the instrument induces some agents to enter the treatment, then the instrument must not induce any agent to leave the treatment. Second, because the impact of treatment may be heterogeneous across agents, the traditional Durbin-Hausman-Wu test for the equivalence of IV and OLS estimates is not valid in a LATE framework. Indeed, the relationship between the Ordinary Least Squares (OLS) and IV estimates is completely uninformative about the existence of selection within the LATE framework. Thus, researchers face the paradox of using IV estimation to correct for selection on untreated outcomes, but with no clear evidence to demonstrate that such selection exists.

To see why, consider the framework of Angrist, Imbens, and Rubin (1996) in which there is a binary instrument, $Z_i \in \{0,1\}$. Without loss of generality, let $Z_i = 1$ increase the likelihood of treatment. They show that we may divide agents into three mutually exclusive sets: the "Always takers," the "Never takers," and the "Compliers." These are defined as:

$$A = \{i : D_i(z_i = 1) = D_i(z_i = 0) = 1\}$$

$$N = \{i : D_i(z_i = 1) = D_i(Z_i = 0) = 0\}$$

$$C = \{i \mid D_i(z_i = 1) = 1; D_i(z_i = 0) = 0\} \ .$$

In this framework, the Wald estimator defines a LATE estimator where

$$\Delta^W = \frac{E(Y_i \mid z_i = 1) - E(Y_i \mid z_i = 0)}{E(D_i \mid z_i = 1) - E(D_i \mid z_i = 0)} = E(Y_{1i} - Y_{0i} \mid C)$$

where Y_{1i} is the potential outcome of the i^{th} agent if treated, Y_{0i} is the potential outcome of the i^{th} agent if not treated, and $Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$ is the observed outcome.

We would say that there is selection on unobserved variables if either of these two conditions fail:

$$E(Y_0 \mid N) = E(Y_0 \mid C) = E(Y_0 \mid A)$$

$$E(Y_1 \mid N) = E(Y_1 \mid C) = E(Y_1 \mid A)$$

That these three conditions are unrelated to the equivalence OLS and IV estimates may be easily demonstrated with the following example. Suppose $P(C) = P(A) = P(N) = \frac{1}{3}$. Further, let $E(Y_1 \mid A) = 1$, $E(Y_1 \mid C) = E(Y_0 \mid C) = 0$, and $E(Y_0 \mid N) = 1$. In this case, the OLS estimate of the impact of treatment is exactly zero. The IV estimate, however, is also exactly zero, but $E(Y_1 \mid A) > E(Y_1 \mid C)$ and $E(Y_0 \mid N) > E(Y_0 \mid C)$, so clearly there is selection on Y_1 and Y_0. How then do we test for such selection?

In this paper, we provide a set of simple tests for the presence of selection bias. Drawing on the work of Black, Sanders, Taylor, and Taylor (2015), the tests come in two forms. First, conditional on covariates, we compare the outcomes of the set of agents who comply with the instruments when not treated to the set of agents who never take treatment. Second, we compare the mean outcomes of agents who comply with the instrument when treated to the set of agents who always take treatment. Mechanically, these tests are implemented by estimating outcome

equations for those who are untreated, or treated, as a function of the covariates and the instruments (or the probability of selection). With a simple Wald-like adjustment, our tests allow researchers to assess the economic magnitude of the selection bias as well.

Our tests are similar to those in Heckman's (1979) seminal paper on the normal selection model. In the two-step estimator for the normal selection model with a common treatment effect, the inverse Mill's ratio represents the control function, and the coefficient on the inverse Mill's ratio identifies the correlation between the errors of the outcome equation and the selection equation. Under the null hypothesis of no selection on unobserved variables, a simple test for selection is to see if the coefficient on the inverse Mill's ratio differs from zero. In more general selection models, the exact form of the control function is unknown, and the control function is estimated semiparametrically.

Not surprisingly given their close relationship to Heckit, aficionados of latent index models have recognized the utility of testing for the existence of selection bias. For instance, Blundell, Dearden, and Sianesi (2005) compare estimates from OLS, matching, IV, and latent index models. They note that the coefficients on their control functions constitute a test for selection, or, in the nomenclature of matching, violation of conditional independence.

Our paper is closely related to Heckman, Schmierer, and Urzua (HSU, 2010) who derive both parametric and nonparametric tests for the correlated random effects model. Formally, HSU develop a test for whether the idiosyncratic return to treatment is independent of treatment status conditional on the covariates. Drawing on the work of Heckman and Vytlacil (2005, 2007a,b), HSU (2010) propose parametric and nonparametric tests that regress the realizations of the dependent variable against the estimated propensity score (which includes the instruments) to see

if the realizations of the outcome variables are linear functions of the propensity score.[1] But as HSU note, the nonparametric tests suffer from low power in sample sizes common in empirical studies. In addition, our tests are considerably easier to implement than their nonparametric tests, which generally require the use of the bootstrap procedures of Romano and Wolf (2005) and Romano and Shaikh (2006) for the step-down method of multiple hypothesis testing. Our tests also provide more insight into the precise nature of the selection problem because we allow for selection on one or both of Y_1 and Y_0.

Bertanha and Imbens (2014) consider similar tests in the context of fuzzy regression discontinuity designs. They do not, however, relate their discussion to general tests for selection on unobserved variables for IV. Similarly, Huber (2013) provides an analysis of noncompliance in experiments virtually identical to ours, but does not extend the analysis to other IV settings. Angrist (2004) proposes a test that compares the estimated treatment effect for compliers to an estimate obtained from using the always takers and never takers. His test does not distinguish among selection on one or both of Y_1 and Y_0 and assumes the magnitude of the treatment effect does not vary with covariates.

While the LATE revolution has led to a more sophisticated interpretation of IV estimates, there has not been much emphasis on testing whether the use of IV methods is necessary. This is peculiar. Heckman, Ichimura, Smith, and Todd (1998) find that most of the difference between simple nonexperimental and experimental estimates of the treatment effect in the Job Training Partnership Act (JTPA) data results from the lack of common support and the distributional differences in covariates, with selection on unobserved variables accounting for

[1] Heckman and Vytlacil (2005) show that if (Y_0, Y_1) are independent conditional on X then the marginal treatment effects (MTEs) are constant.

only about seven percent of the difference. Blundell, Dearden, and Sianesi (2005) find, when estimating their "single treatment" model using the very rich National Child Development Survey data, that there is little evidence that their matching estimates suffer from any selection bias. Similarly, when their outcomes are measured identically, Diaz and Handa (2006) report that propensity score matching estimates matched the experimental evidence from the famous PROGRESA experiment. While by no means conclusive, matching on a rich set of covariates motivated by theory and the institutional context, and limiting the samples to where one has common support between treated and nontreated agents, appears to reduce dramatically any biases in many substantive contexts. Indeed, given the well-documented increase in the variance of estimates when using IV estimation, one may be willing to live with a modest bias if the resulting variance is drastically reduced, much as in nonparametric estimation, one is willing to live with a larger bias in return for variance reduction.

In the next section of our paper, we outline the necessary restrictions to implement matching and OLS. In section three, we outline the necessary assumptions for Imbens and Angrist's IV estimation and the latent index approach of Vytlacil (2002). In section four, we outline a simple test for violation of the conditional independence assumption. In section five, we provide our empirical applications, and in section six we offer concluding remarks.

2. Matching and Ordinary Least Squares and Selection on Observed Variables

In this section, we briefly present the standard evaluation framework for thinking about estimating the causal impact of treatment. The presentation here is largely based on Heckman, Ichimura, and Todd (1997, 1998), Heckman and Smith (1998), and Heckman, LaLonde, and

Smith (1999).

Let $D_i \in \{0,1\}$ index whether the i^{th} agent receives treatment or not. Each agent has two potential outcomes (Y_{0i}, Y_{1i}), where Y_{0i} is the agent's outcome if not treated, and Y_{1i} is the agent's outcome if treated. We define the causal impact of the treatment on the i^{th} agent as

$$\delta_i = Y_{1i} - Y_{0i}. \tag{1}$$

The fundamental problem of evaluation is that we observe only one of the two potential outcomes; researchers must estimate the other, which is often referred to as the "missing counterfactual."

One intuitive class of estimators for generating the missing counterfactuals is matching estimators. Matching estimators rely on the assumption that researchers have sufficiently rich covariates that any differences in the decision of the agents to take treatment are independent of the agents' potential outcomes conditional on the covariates. Let X denote those covariates. Formally, matching estimators rely on two assumptions. The first is the Common Support Assumption (CSA) or

$$0 < \Pr(D = 1 \mid X) < 1. \tag{CSA}$$

The CSA simply requires that if, for example, you are going to match a treated agent to someone who did not take treatment, there must be someone in the set of untreated agents with approximately the same realization of the covariates. Of course, the CSA is a testable assumption. When it is violated, as is often the case, researchers generally change their definition of the relevant population to the population over which the CSA holds, reflecting the limited variation that the data provide.[2]

[2] Crump, Hotz, Imbens, and Mitnik (2009) argue that the efficiency bounds of semiparametric and parametric estimators imply that estimation using propensity score matching estimators

More vexing, however, is the next assumption necessary for the use of matching estimation: the Conditional Independence Assumption (CIA). Indeed, there are various "flavors" of the CIA depending on the parameter that the researcher wishes to estimate. The strongest of the flavors is:

$$(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i \mid X_i \qquad \text{(CIA)}$$

This version of the CIA allows researchers to estimate the Average Treatment Effect, or

$$\Delta = E\left(Y_1 - Y_0\right), \qquad (2)$$

using matching methods. The CIA for Y_0 assumes

$$Y_{0i} \perp\!\!\!\perp D_i \mid X \ . \qquad (\text{CIA}^0)$$

This version of the CIA allows the estimation of

$$\Delta^{ATET} = E\left(Y_1 - Y_0 \mid D = 1\right). \qquad (3)$$

Because researchers observe Y_{1i} for those who are treated, estimation of the Δ^{ATET} only requires the weaker (CIA^0) rather than the (CIA^{ATE}). Similarly, when estimating the average treatment effect for the nontreated, researchers need only assume

$$Y_{1i} \perp\!\!\!\perp D_i \mid X\ , \qquad (\text{CIA}^1)$$

which allows the estimation of

$$\Delta^{ATEN} = E\left(Y_1 - Y_0 \mid D = 0\right). \qquad (4)$$

Of course, the (CIA) is the union of the (CIA^1) and (CIA^0).

---

should be limited to the regions of "thick" overlap. They suggest the range [0.1, 0.9] is an adequate approximation of the optimal range for many applications. Black and Smith (2004) consider an even more limited interval of overlap.

For semiparametric or nonparametric matching models, it is common to specify the functions that determine the potential outcomes (Y_{0i}, Y_{1i}) as

$$Y_{1i} = g_1(X_i) + \varepsilon_{1i} \quad (5)$$
$$Y_{0i} = g_0(X_i) + \varepsilon_{0i} \quad (6)$$

where $(g_0(\cdot), g_1(\cdot))$ are unknown conditional mean functions and $(\varepsilon_{0i}, \varepsilon_{1i})$ summarize the residual uncertainty associated with the unobserved variables. With the CSA and appropriate version of the CIA, researchers may use a variety of methods to estimate the unknown conditional mean functions; see for instance Imbens (2004), Heckman, LaLonde, and Smith (1999), Smith and Todd (2005), Huber, Lechner, and Wunsch (2013), and Busso, Dinardo, and McCrary (2014).

A commonly used alternative to matching methods is to use Ordinary Least Squares (OLS) to estimate parametric linear models. Researchers specify the functional form of the conditional mean function as

$$g_1(X_i) = X_i'\beta_1 \quad (7)$$
$$g_0(X_i) = X_i'\beta_0 . \quad (8)$$

In these models, the researcher may avoid making the CSA by instead making assumptions about the functional form.

The common criticism of matching and OLS estimates is that they rely on the overly strong CIA. To avoid making the CIA, applied researchers have turned to Instrumental Variables (IV) estimation. While traditional IV methods require that the treatment effect be the same for everyone, Imbens and Angrist (1994) demonstrate that with stronger assumptions IV estimation allows for heterogeneous treatment effects. Researchers now routinely invoke the LATE framework when using IV analysis.

It is difficult to overemphasize the importance of this advance. Models without heterogeneous treatment effects seem incapable of explaining the complexity of human behavior. With heterogeneous treatment effects, much more plausible and interesting models may be considered and estimated, including the justifiably famous Roy (1951) model. Indeed, Heckman, Urzua, and Vytlacil (2006) term these heterogeneous impacts "essential heterogeneity."

3. The IV and Control Function Approach to Selection on Unobserved Variables

Imbens and Angrist (1994) provide assumptions that allow for the estimation of heterogeneous treatment effects with IV methods. We wish to consider the possible decisions of the i^{th} agent for any value of Z_i, which is the set $\{D(z) \mid z \in \mathbb{Z}\}$. We may now state the assumptions of the LATE estimator as the Existence of Instruments (EI) and Monotonicity (M). Formally,

$$(Y_{0i}, Y_{1i}, \{D(z) \mid z \in \mathbb{Z}\}) \perp\!\!\!\perp Z \mid X \text{ and } \Pr(D = 1 \mid X, Z = z) \text{ is a nontrivial function of } z \qquad \text{(EI)}$$

$$\forall z^0, z^1 \in \mathbb{Z} \text{ either } D_i(z^0) \geq D_i(z^1) \forall i \text{ or } D_i(z^0) \leq D_i(z^1) \forall i \,. \qquad \text{(M)}$$

The M assumption requires that all agents respond to the instrument in the same direction, not that the function $\Pr(D = 1 \mid X, Z = z)$ be monotone in z; this led Heckman, Urzua, and Vytlacil (2006) to rename the condition uniformity, although the somewhat confusing monotonicity was too well-established to be displaced. The M assumption is of course restrictive. Should the M assumption fail while the EI assumption holds, IV estimation provides of mixture of treatment effects associated with agents who both enter and leave the treatment as the instrument varies.

To keep the notation simple, for the remainder of the paper we will let $Z_i \in \{0,1\}$; our arguments, however, generalize to continuous instruments.

Imbens and Angrist noted that the latent index models pioneered by Heckman and various co-authors imply the (EI) and (M) conditions. In an important paper, Vytlacil (2002) shows the equivalence of the two approaches. Latent index models may be used to circumvent the problems associated with selection on unobserved variables. In our notation, one may define the expectations of the errors in our equations (5) and (6) as zero, or $E(\varepsilon_{1i}) = E(\varepsilon_{0i}) = 0$. This is, of course, a convenient normalization with any nonzero mean being absorbed into the conditional mean functions. When we observe only a portion of our potential outcomes, we no longer know that the conditional expectations $E(\varepsilon_{1i} \mid D_i = 1)$ and $E(\varepsilon_{0i} \mid D_i = 0)$ equal zero. To see why, we follow Vytlacil (2002) and let

$$D_i = 1\left(h(Z_i, X_i) + U_i \geq 0\right) \text{ and } h(Z_i, X_i) \text{ be a nontrivial function of } z \qquad \text{(V1)}$$

$$Z_i \perp\!\!\!\perp (Y_{1i}, Y_{0i}, U_i) \mid X_i \qquad \text{(V2)}$$

where $1(\cdot)$ is an indicator function for the condition holding, U_i is a random variable, and $h(Z_i, X_i)$ is the index function.

With assumptions (EI) and (M) (or the equivalent assumptions (V1) and (V2) for latent index models), we may write $E(\varepsilon_{1i} \mid D_i = 1)$ and $E(\varepsilon_{0i} \mid D_i = 0)$ as

$$E(\varepsilon_{1i} \mid D_i = 1) = c_1\left(X_i, P(X_i, Z_i)\right) + e_{1i} \qquad (9)$$

$$E(\varepsilon_{0i} \mid D_i = 0) = c_0\left(X_i, P(X_i, Z_i)\right) + e_{0i} \qquad (10)$$

where $P(Z_i, X_i) \equiv \Pr(D_i = 1 | Z_i, X_i)$ is the probability of selecting treatment and $c_1(\cdot)$ and $c_0(\cdot)$ are the control functions that model the conditional mean so that we now have $E(e_{1i}) = E(e_{0i}) = 0$.[3]

With the control function approach, it is a bit easier to interpret the independence assumption, $(Y_{0i}, Y_{1i}, \{D(z) | z \in \mathbb{Z}\}) \perp\!\!\!\perp Z | X$, that is embedded in the assumption (EI). The independence assumption simply requires that Z_i be independent of (U_i, Y_{1i}, Y_{0i}) conditional on X_i. Given the equivalence of the LATE and control function assumptions, we refer to the (EI) and (M) assumptions, or (V1) and (V2), as the Vytlacil-Imbens-Angrist (VIA) assumptions.

4. Testing for Conditional Independence under the VIA Assumptions

A. Instrumental Variables

In this section, we develop a simple, easily applied test for selection on unobserved variables. As noted above, the various (CIA) assumptions allow researchers to ignore the possibility of selection on unobserved variables, although they are generally invoked without examining whether there is evidence for selection on unobserved variables. In contrast, the VIA assumptions allow us to consistently estimate LATEs for those individuals who comply with the instruments. In the case of a parametric model with a single instrument that is linear in parameters we would augment equations (7) and (8) to obtain

$$E(Y_{1i} | D_i = 1) = X_i \beta_1 + \alpha_1 Z_i \quad (11)$$

$$E(Y_{0i} | D_i = 0) = X_i \beta_0 + \alpha_0 Z_i . \quad (12)$$

[3] The probability of treatment $P(X_i, Z_i) \equiv \Pr(D_i = 1 | X_i, Z_i)$ is not the propensity score used in matching estimators as it depends on the instruments Z; see Heckman and Navarro (2004) for a discussion of the information sets necessary for the use of matching.

With nonbinary instruments researchers may wish to add higher order terms – i.e. replace Z_i with $f(\mathrm{Z}_i)$. With multiple instruments, researchers would probably want to replace Z_i with the estimated propensity score, $\hat{p}(X_i, Z_i)$ and adjust the standard errors for generated regressors as in Murphy and Topel (2002).[4] In the case of matching estimators, we would augment equations (5) and (6) and specify the conditional mean functions as

$$E(Y_{1i} \mid X_i, D_i = 1) = g_1(X_i) + \alpha_1 Z_i \tag{13}$$
$$E(Y_{0i} \mid X_i, D_i = 0) = g_0(X_i) + \alpha_0 Z_i . \tag{14}$$

To clarify the relationship among the various forms of the CIA and the test we are implementing, it is useful to outline the samples being used and hypotheses involved when estimating these auxiliary regressions. Formally, we estimate (12) or (14) using the sample of untreated observations to test

H^0: CIA^0 holds, or $\alpha_0 = 0$

H^A: CIA^0 does not hold, or $\alpha_0 \neq 0$.

Similarly, we estimate equations (11) and (13) using the sample of treated observations to test

H^0: CIA^1 holds, or $\alpha_1 = 0$

H^A: CIA^1 does not hold, or $\alpha_1 \neq 0$.

To develop some intuition for the tests, assume that $D_i(z_i = 1) \geq D_i(z_i = 0)$. As in Angrist, Imbens, and Rubin (1996), we may divide agents under the VIA into three types: the "always takers", the "never takers", and the "compliers" that we defined in the introduction. The test given in either equation (11) or equation (13) simply compares $E(Y_1 \mid x, A)$ to $E(Y_1 \mid x, C)$. As Black, Sanders, Taylor, and Taylor (2015) note, this is easily done because

[4] See Joo and LaLonde (2014) for more discussion and a test that relies on the control functions.

$E(Y_1 \mid x, z=0) = E(Y_1 \mid x, A)$ and $E(Y_1 \mid x, z=1) = E(Y_1 \mid x, A \cup C)$. Thus, at $X = x$, we have that $\alpha_1(x) = \dfrac{\Pr(C \mid x)(E(Y_1 \mid x, C) - E(Y_1 \mid x, A))}{\Pr(C \mid x) + \Pr(A \mid x)}$. The regression coefficient in either equation (11) or equation (13) then simply integrates over the realizations of X, or $\alpha_1 = \int \alpha_1(x)\, dF(x)$ for some function $F(x)$. Put differently, the tests look for evidence of a nonconstant control function in equation (9): i.e., evidence that unobserved variables are affecting the outcomes. A parallel argument applies to α_0.

The finding that either $\alpha_0 \neq 0$ or $\alpha_1 \neq 0$ constitutes evidence either of selection or of violation of the exclusion restrictions (i.e., the failure of EI). If one is willing to maintain the assumption that Z_i is an exclusion restriction, however, then there is simple, compelling evidence for violation of the CIAs when rejecting the null hypotheses. Indeed, we view the simplicity of our tests as their greatest virtue.

The tests also allow researchers to assess whether any selection arises on Y_0, which represents a violation of CIA^0, or whether any selection arises on Y_1, which represents a violation of CIA^1, or both. In addition, as with the tests for selection in Heckman (1979), our tests allow researchers to determine the sign of the relevant selection biases and their magnitudes. This allows researchers to provide a much more nuanced discussion of the nature of the agents' choice problems. Given the equivalence that Vytlacil (2000, 2002) demonstrates, it is perhaps not surprising that we may learn much more about the selection problem using IV methods than we learn from current practices.

In the next two subsections, we show how these tests may be adapted to two other common settings in applied research: fuzzy regression discontinuity designs, and experiments with imperfect compliance.

B. Fuzzy Regression Discontinuity

In regression discontinuity designs, a running variable *S* results in a discrete jump in the probability of getting treatment at a given point. We assume that the jump in the probability of treatment occurs at $S = 0$. For our test, we need to assume that the regression discontinuity is a fuzzy design so that the probability of treatment is such that

$$\left| \lim_{S \uparrow 0} \{\Pr(D = 1 \mid X, S)\} - \lim_{S \downarrow 0} \{\Pr(D = 1 \mid X, S)\} \right| = k < 1.$$

As emphasized by Imbens and Lemieux (2008) and Lee and Lemieux (2010), when faced with a fuzzy RD, researchers estimate a LATE at $S = 0$ using IV.

Because there is a discrete change in the probability of treatment, if there is selection on unobserved variables we would expect the control function to jump as well. Hence, if there is selection on unobserved variables, we should see a jump in the value of $E(Y_0 \mid S, D = 0)$ as S crosses zero, while the (CIA$^{\text{ATET}}$) assumption would require $E(Y_0 \mid S, D = 0)$ to be a smooth function as S crosses zero. This suggest a simple test of the form

$$E(Y_0 \mid D = 0) = g_0(X_i) + \alpha_0 1[S_i \geq 0] \tag{15}$$

with the null hypothesis being that $\alpha_0 = 0$ or the corresponding

$$E(Y_1 \mid D = 1) = g_1(X_i) + \alpha_1 1[S_i \geq 0] \tag{16}$$

with the null hypothesis being that $\alpha_1 = 0$. For estimating equation (15) only observations that are not treated are used, which implies that the sample is a mixture of compliers and never

takers. Similarly, for estimating equation (16) only observations that are treated are used, which implies that the sample is a mixture of compliers and always takers.

In an important paper, Bertanha and Imbens (2014) consider similar tests for fuzzy regression discontinuity designs. Indeed, they state, "As a matter of routine, we recommend that researchers present graphs with estimates of these two conditional expectations in addition to graphs with estimates of the expected outcome conditional on the forcing variable alone." We concur.

C. Experiments with Imperfect Compliance

As Heckman (1996) emphasizes, experimental assignment of treatment may be thought of as an instrument for treatment. Because many social experiments have imperfect compliance (e.g., Heckman, Hohmann, Smith, and Khoo (HHSK), 2000), with both treatment group dropout and control group substitution into similar treatments provided elsewhere, one could easily implement our tests to see if there is selection on Y_1 or Y_0 in experiments. For instance, HHSK report that only between 49 and 59 percent of the Job Training Partnership Act (JTPA) treatment groups received services while between 27 and 40 percent of the control groups received services.

With this much dropout and substitution, applied researchers will often rely on the Bloom (1984) estimator. To use the Bloom estimator, the researcher need only use the random assignment to the treatment group as an instrument for the receipt of treatment. This results in an instrumental variables estimator that uses a binary instrument and hence uses a Wald estimator that recovers the LATE for those who comply with the experimental protocol. Huber (2013)

provides an analysis of noncompliance in experiments that is virtually identical to our analysis, but does not extend it to other IV settings.

D. Recovering Estimates of the Magnitude of the Selection Bias

To recover estimates of the magnitude of the selection bias, continue to assume that $Z = 1$ encourages treatment, and ignore covariates for notational simplicity. We have

$$E(Y_0 \mid z_i = 0) = \frac{\Pr(C)}{\Pr(C) + \Pr(N)} E(Y_0 \mid C) + \frac{\Pr(N)}{\Pr(C) + \Pr(N)} E(Y_0 \mid N) \tag{17}$$

while

$$E(Y_0 \mid z = 1) = E(Y_0 \mid N) \tag{18}$$

so that

$$\alpha_0 \equiv E(Y_0 \mid z = 1) - E(Y_0 \mid z = 0) = \frac{\Pr(C)}{\Pr(C) + \Pr(N)} \left(E(Y_0 \mid N) - E(Y_0 \mid C) \right). \tag{19}$$

Thus, a measure of the selection bias for Y_0, which we denote B_0, is simply

$$B_0 = \frac{\Pr(C) + \Pr(N)}{\Pr(C)} \alpha_0 \ . \tag{20}$$

Similarly, we have

$$E(Y_1 \mid z_i = 1) = \frac{\Pr(C)}{\Pr(C) + \Pr(A)} E(Y_1 \mid C) + \frac{\Pr(A)}{\Pr(C) + \Pr(A)} E(Y_1 \mid A) \tag{21}$$

while

$$E(Y_1 \mid z = 0) = E(Y_1 \mid A) \tag{22}$$

so that

$$\alpha_1 \equiv E(Y_1 \mid z = 1) - E(Y_1 \mid z = 0) = \frac{\Pr(C)}{\Pr(C) + \Pr(A)} \left(E(Y_1 \mid C) - E(Y_1 \mid A) \right). \tag{23}$$

A measure of the selection bias for Y_1, which we denote B_1, is simply

$$B_1 = \frac{\Pr(C) + \Pr(A)}{\Pr(C)} \alpha_1 \tag{24}$$

To implement these measures empirically, we may use the OLS estimates of (α_0, α_1). We know that $\Pr(A) = \Pr(D = 1 \mid z = 0)$ and $\Pr(N) = \Pr(D = 0 \mid z = 1)$ so we have sample analogues of all the terms on the right-hand side of equations (20) and (24).

5. Empirical Applications

A. Angrist and Evans (1998) data

Our first application is taken from Angrist and Evans (1998). This paper uses data from the 1980 and 1990 US Censuses to measure the causal impact of children on a family's labor supply. Because fertility is likely to be endogenous with respect to women's labor supply decisions, Angrist and Evans devised an ingenuous instrumental variables strategy. Limiting their sample to women who have at least two children, Angrist and Evans noticed that women whose first two children are the same sex are more likely to have additional children than women whose first two children are of opposite sexes. For instance, in the 1980 Census, married women whose first two children are of the same sex are about 6 percentage points more likely to have additional children than women whose first two children are of opposite sexes. For our analysis, we focus on the labor supply decisions of women in the 1980 Census, corresponding to the estimates in column (2) of their Table 7.[5]

In many ways, this design is ideal. Because of the random nature of child sex determination, the sample is split approximately equally between families whose first two

[5] We thank Bill Evans for providing us with the data.

children are of the same sex and those whose children are of opposite sexes. In these data, 51.1% of the children born are male, and 50.6% of families' first two children are of the same sex. Formally, the system that Angrist and Evans estimate is:

$$y_i = x_i'\beta + \delta\, morekids_i + \varepsilon_i \quad (25)$$

$$morekids_i = x_i'b + \gamma\, samesex_i + u_i \quad (26)$$

where the covariates x_i include the age of the mother, the age of the mother at first birth, indicators for whether the mother is black or whether the mother is nonblack and nonwhite (white is the omitted category), an indicator for whether the mother is Hispanic, an indicator for whether the first child was a boy, and an indicator for whether the second child was a boy. The instrument, *samesex*, is an indicator for whether the first two children were either both boys or both girls. For dependent variables, we use a subset of those explored by Angrist and Evans: whether the mother worked in the previous year, the number of weeks worked in that year, typical hours worked in that year, and her income from working. All variables include zeros when the women did not work in the previous year. The sample is limited to women 21 to 35 years of age; see Angrist and Evans (1998) for more details.

In Table 1, we replicate the Angrist and Evans results in the 1980 Census; see their Table 7, columns (1) and (2). We also use a semiparametric approach and estimate

$$y_i = g(x_i) + \delta morekids_i + \varepsilon_i \quad (27)$$

$$morekids_i = h(x_i) + \gamma samesex_i + u_i \quad (28)$$

where $\{g(\cdot), h(\cdot)\}$ are unknown functions. Because our data are discrete, we estimate $g(x_i)$ by fully interacting X. Our parametric results – both the OLS and Two-Stage-Least-Squares (TSLS) estimates – exactly match the Angrist and Evans findings. Moreover, the semiparametric estimates are virtually identical to the parametric estimates of Angrist and Evans, which is not

too surprising given that the sex of women's offspring is independent of all of our observed characteristics.

For this IV approach to be interpretable as a LATE, of course, we need to assume the VIA conditions. Angrist and Evans documented that the instrument does indeed raise fertility. In addition, we need to assume that the instrument is an exclusion restriction in the sense that having the first two children be the same sex does not directly affect the women's labor supply decisions, and we need to assume the monotonicity (or uniformity) condition so that having two children of the same sex reduces no one's fertility. With these assumptions, we may now implement our parametric tests of the CIAs using:

$$y_{0i} = x_i'\beta_0 + \alpha_0 samesex_i + \varepsilon_i \quad (29)$$

$$y_{1i} = x_i'\beta_1 + \alpha_1 samesex_i + v_i \quad (30)$$

and semiparametric tests using

$$y_{0i} = g_0(x_i) + \alpha_0 samesex_i + \varepsilon_i \quad (31)$$

$$y_{1i} = g_1(x_i) + \alpha_1 samesex_i + v_i \ . \quad (32)$$

where for our semiparametric analysis we need to drop the indicator for the second child being a boy to avoid making the *samesex* variable perfectly collinear with the x_i vector. Equations (29) and (31) are estimated on the sample of 236,092 women who have two children, and equations (30) and (32) are estimated on the sample of 158,743 who have three or more children.

Our results are presented in Table 2. For the case (CIA^{ATET}), the data strongly reject the null that $\alpha_0 = 0$. For each of the four outcomes, we reject the null hypothesis at a five-percent confidence level. Indeed, for two of the four outcomes we reject the null in excess of a one-in-a-thousand confidence level. In each case, the estimated α_0 is positive, indicating that the compliers (those who limit their fertility when having a boy and a girl) have higher earnings,

hours worked, weeks worked, and are more likely to work conditional on our covariates relative to the never takers.

In contrast, we find very little evidence against the (CIA^{ATEN}). Relative to our estimates of α_0, our estimates of α_1 are statistically insignificant and economically very small. Thus, we find no evidence of a selection effect when estimating the missing counterfactual Y_1. Frankly, we find this result stunning. The US Census data have large sample sizes but suffer a paucity of covariates, with the data including only broad demographic controls. Before undertaking this analysis, we fully expected to show a two-sided selection problem. The data disagreed.

Angrist (2004) tests for selection on weeks worked (as well as other outcomes we do not consider) using his test and finds evidence of selection. His test considers the joint hypothesis of no selection on (Y_{0i}, Y_{1i}), however, so our test allows for a bit more nuanced understanding of the impact of selection.

In terms of the magnitude of the selection effects we use the nonparametric estimates in column (3) of Table 2. We see that never takers are five percentage points more likely to have worked last year compared with the compliers, they worked three weeks more than the compliers, worked two hours per week more, and earned $1,965 more per year than the compliers. Comparing the compliers to the always takers, we find that compliers were one percent more likely to work last year, they worked 0.4 extra week per year, they worked a tenth of an hour more per week, and earned $38 dollars less per year than the never takers. Obviously, the compliers represent a poor comparison group for the never takers, but the compliers do not seem substantially different than the always takers.

B. Eberwein, Ham, and LaLonde (1997) data

When there is control group substitution and treatment group dropout in an experiment, researchers will often estimate two treatment parameters: the intent-to-treat parameter, estimated as the differences in the dependent variables between the treatment and control groups, and the impact of treatment for those who comply with the treatment protocol, estimated using Bloom's (1984) estimator. Bloom's estimator corresponds to TSLS where assignment to the treatment group is an instrument for the receipt of treatment. Because assignment to the treatment group is independent of the potential outcomes (Y_0, Y_1), the assignment represents an exclusion restriction that functions as an instrument under the VIA assumptions.

We examine the impact of training for a sample of adult women who took part in the Job Training Partnership Act (JTPA) experiment; see e.g. Bloom *et al.* (1997) for a discussion of the experiment and analysis of the results. Our sample, the same one used by Eberwein, Ham, and LaLonde (1997), is restricted to the set of women recommended for classroom training (the CT-OS treatment stream in the jargon of the experiment). We measure treatment as the onset of classroom training within nine months of randomization.[6]

For our outcome variable, we use an indicator for whether the participant was employed in the eighteenth month after random assignment. We also know whether the individual was assigned to the treatment group or the control group. There was much non-compliance in this sample. Only about 65 percent of the treatment group reported received classroom training in the first nine months after random assignment. There was much substitution as well; about 34

[6] We rely on self-reported training and ignore other services, such as job search assistance, received by some members of both the treatment group and the control group. See Smith and Whalley (2015) for a depressing exploration of the concordance of administrative and self-reported measures of training in the JTPA study data.

percent of the control group receive classroom training in the first nine months after random assignment.

In Table 3, we provide two sets of estimates of the intent-to-treat parameter. In column (1) we provide the simple difference estimates given by

$$y_i = \beta + \delta R_i + \varepsilon_i, \tag{33}$$

where y_i is an indicator for whether the participant is employed, R_i is an indicator for whether the participant was assigned to the treatment group during random assignment, ε_i is the error term, and (β, δ) are parameters to be estimated. The intent-to-treat parameter, δ, equals 0.041 and is statistically significant at the five-percent level. This relatively modest effect, however, hides a larger impact of treatment for people who complied with the treatment protocol, which equals 0.136 and again is statistically significant at the five-percent level. The reason for the differential, of course, is that random assignment only increases the rate of treatment about 0.305, which is the coefficient on random assignment from the first-stage of our TSLS Bloom estimator.

Nothing in this analysis, however, informs whether there is selection on unobserved variables into treatment. Toward that end, we next estimate the following equations

$$y_{0i} = \beta_0 + \alpha_0 R_i + \varepsilon_{0i} \tag{34}$$

$$y_{1i} = \beta_1 + \alpha_1 R_i + \varepsilon_{1i} \tag{35}$$

where (y_{0i}, y_{1i}) are the outcomes of those not receiving training and receiving training. Equation (34) is estimated on the 1,233 adult women who do not receive training, while equation (35) is estimated on the 1,501 who do receive training. We find little evidence that the compliers with the protocol have different Y_0 than those who never take training. The coefficient on being

assigned to the treatment group in equation (34) is small, 0.006, and statistically insignificant at the five-percent level. In contrast, the coefficient on being assigned to the treatment group in equation (35) is large, 0.068, and statistically significant. These estimates imply that while the always takers have a mean employment rate of 0.50, the compliers when treated have a mean employment rate of 0.65. Thus, the always takers are adversely selected with respect to the likelihood of employment.

Showing large selection on unobserved variables without covariates, however, is hardly surprising. Thus, we augment our equations with a vector of education, demographic, and pre-random assignment labor market and transfer payment recipiency variables; see the footnotes to Table 3 for a complete list of the covariates. Their inclusion (as expected) has only modest effects on the intent-to-treat and LATE estimates, although some may be dismayed that the estimates no longer clear the five-percent hurdle. For our tests for the presence of selection on unobserved variables, there is also little change in our estimates. In the Y_1 regression, the coefficient on the assignment indicator for those receiving treatment falls from 0.068 to 0.062 and remains significant at the five-percent level. Despite the inclusion of detailed information on labor supply in the 12 months prior to random assignment and other controls, the coefficient on the assignment to the treatment group is reduced only about nine percent. The observed variables examined here account for little of the selection.

6. Conclusion

In this paper, we have derived a simple test for selection on unobserved variables when using instrumental variables. The test is simple; any well-trained undergraduate can implement it. Using a Wald-like estimator, one can use these estimates to assess the magnitude of the selection

bias as well. This allows researchers a much better understanding of the precise nature of selection on unobserved variables.

Our two empirical applications nicely demonstrate what can be learned from these tests. First, we revisit the Angrist and Evans (1998) analysis of the impact of children on married women's labor supply using the sex composition of the first two children as an instrument. To our considerable surprise, we find little evidence of selection into having three children despite the relatively modest set of covariates available in the census. In contrast, those who complied with the instrument and had a third child seem extremely different from those who always stop at two children. The labor market earnings of never takers are about $2,000 more per year than the earnings of women who complied with the instruments. Surprisingly, we find that there are no (statically or substantively) significant differences between the always takers and the compliers.

Our second application also contained a surprise. We reanalyzed the probability of employment for adult women in the JTPA experiment 18 months after random assignment. While the impact of treatment on the treated showed a sizeable impact on the probability of employment (0.136), the selection of participants into treatment status was even larger than the treatment effect. Those who complied with the treatment protocol had a much higher employment rate when treated (0.145 higher) than those women who always took treatment, a 29 percent higher rate of employment. Even after conditioning on an extensive set of predetermined covariates, the selection effect remained a bit larger than the impact of treatment itself.

The ability to assess the magnitude of the potential selection bias is an important feature of our approach. Statisticians and economists have long recognized that there is a fundamental tradeoff in, say, nonparametric estimation between variance and bias. When deciding on whether to use IV estimators or estimators that rely on selection on observed variables, however,

there has not been a similar discussion. This omission arises undoubtedly because of the lack of a simple means to assess the magnitude of the selection biases. With our simple methods of assessing the bias, perhaps we may improve these discussions.

Since the publication of Vytlacil (2002), we have understood the equivalence between the assumptions necessary for IV and latent index model estimation. But IV estimation has always seemed to provide less information about the nature of the selection bias than control function estimation. In this paper, however, we showed that simple auxiliary regressions will produce rich insights into the nature and magnitude of the selection bias when using IV estimation.

Table 1: Causal Impact of Having More than Three Children on Mother's Labor Supply, Angrist and Evans 1998

	More kids coefficient, parametric OLS model	More kids coefficient, parametric IV model	More kids coefficient, semiparametric model	More kids coefficient, semiparametric IV model
Worked last year	-0.176***	-0.120***	-0.174***	-0.117***
	(0.00162)	(0.0249)	(0.00164)	(0.0250)
Weeks worked	-8.97***	-5.66***	-8.90***	-5.53***
	(0.0707)	(1.108)	(0.0727)	(1.109)
Hours worked	-6.66***	-4.59***	-6.59***	-4.45***
	(0.0611)	(0.9452)	(0.0620)	(0.9461)
Income	-3,768***	-1,960***	-3,739	-1,915***
	(33.45)	(541.5)	(35.47)	(542.0)
First Stage: Same sex coefficient	---	0.062***	----	0.062***
		(0.0015)		(0.0015)
N	394,835	394,835	394,835	394,835

*5 percent significance level, ** 1 percent significant level, *** 0.1 percent significance level

Notes: Covariates in the parametric model include age of the mother, the age of the mother at first birth, indicators for whether the mother is black or non-black and nonwhite, an indicator for whether the mother is Hispanic, an indicator for whether the first child was a boy, and an indicator for whether the second child was a boy. For the semiparametric model, we drop the indicator for the second child being a boy to avoid perfect colinearity with the instrument, an indicator that both of the first two children are the same sex. The semiparametric IV regression model uses a fully saturated model in the covariates and an additively separable term for having more children. The F-statistics on the instrument for the parametric model is 1,711. For the semiparametric model it is 1,702. For the semiparametric model, 72 cases have predicted values of one for the probability of having more children and 168 have a predicted probability of zero. Our parametric estimates exactly match those of Angrist and Evans, Table 7, columns (1) and (2).

Table 2: Test of CIA using Angrist and Evans (1998) Data

	OLS		Semiparametric	
Worked last year	**(CIA^{ATET}) test**	**(CIA^{ATEN}) test**	**(CIA^{ATET}) test**	**(CIA^{ATEN}) test**
Dependent variable	Y_0	Y_1	Y_0	Y_1
Coefficient on instrument	0.0046*	0.0015	0.0051**	0.0017
(standard error)	(0.00197)	(0.0025)	(0.0020)	(0.0025)
[bias]	[0.047]	[0.011]	[0.052]	[0.012]
N	236,092	158,743	236,092	158,743
Weeks worked	**(CIA^{ATET}) test**	**(CIA^{ATEN}) test**	**(CIA^{ATET}) test**	**(CIA^{ATEN}) test**
Dependent variable	Y_0	Y_1	Y_0	Y_1
Coefficient on instrument	0.297***	0.053	0.315***	0.063
(standard error)	(0.0902)	(0.1043)	(0.0903)	(0.1047)
[bias]	[3.01]	[0.37]	[3.19]	[0.44]
N	236,092	158,743	236,093	158,743
Hours worked	**(CIA^{ATET}) test**	**(CIA^{ATEN}) test**	**(CIA^{ATET}) test**	**(CIA^{ATEN}) test**
Dependent variable	Y_0	Y_1	Y_0	Y_1
Coefficient on instrument	0.205**	-0.0004	0.221**	0.016
(standard error)	(0.0753)	(0.0925)	(0.0753)	(0.0927)
[bias]	[2.08]	[0.00]	[2.24]	[0.11]
N	236,092	158,743	236,093	158,743
Income	**(CIA^{ATET}) test**	**(CIA^{ATEN}) test**	**(CIA^{ATET}) test**	**(CIA^{ATEN}) test**
Dependent variable	Y_0	Y_1	Y_0	Y_1
Coefficient on instrument	188***	-7.01	194***	-5.49
(standard error)	(45.43)	(48.28)	(45.40)	(48.41)
[bias]	[1,904]	[-49]	[1,965]	[-38]
N	236,092	158,743	236,092	158,743

*5 percent significance level, ** 1 percent significant level, *** 0.1 percent significance level

Table 3: Impact of Training for Adult Women Selected for Classroom training in the National JTPA Study

	(1)	(2)
Covariates	**No**	**Yes**
Mean of employment for control group	0.505	0.505
Mean training for control group	0.344	0.344
Intent to treat	0.041**	0.037*
(standard error)	(0.0203)	(0.0198)
(n=2,374)		
Bloom estimator		
First-stage treatment indicator	0.305***	0.305***
(standard error)	(0.0194)	(0.0191)
[F-statistic on instrument]	[246]	[246]
(n=2,374)		
Impact of classroom training on compliers	0.136**	0.122*
(standard error)	(0.0670)	(0.0650)
[n=2,374]		
Treatment group indicator for Y_0 regression	0.006	0.005
(standard error)	(0.0285)	(0.0273)
[bias]	[0.013]	[0.011]
(n=1,233)		
Treatment group indicator for Y_1 regression	0.068**	0.062**
(standard error)	(0.0318)	(0.0313)
[bias]	[0.145]	[0.132]
(n=1,501)		

Note: Dependent variable is an indicator variable for whether the participant is employed 18 months after random assignment. The treatment indicator is one when the participant is assigned to the treatment group. The classroom training variable is an indicator for whether the participant received classroom training in the first 9 months after random assignment. For the specification with covariates, the set of covariates include age and the square of age and a vector of indicator variables. The indicator variables are whether the participant has never been married, whether the participant is currently married, whether the participant is a non-Hispanic black, whether the participant is Hispanic, whether the participant is another race/ethnicity (white, non-Hispanic is the excluded category), whether the participant has less than a high school degree, whether the participant has a GED, whether the participant has more than a high school degree (high school degree is the excluded category), whether the participant was on AFDC at the time of random assignment, whether the participant was on food stamps at the time of random assignment, whether the participant was on AFDC for two years or more, whether the participant had children under five years of age in the household, whether the participant had children under 18 in the household, whether the participant reported problems with her English skills, whether the participant reported never working for pay, whether the participant reported never working full time, whether the participant worked in the previous 12 months prior to random assignment, a cubic in the fraction of the year that the participant worked prior to random assignment, and 15 indicators for the site of the experiment. To avoid dropping observations, if a variable was missing we set its value to zero and added an indicator variable equal to one when the variable was missing.

References

Angrist, Joshua D. and William N. Evans, 1998. "Children and Their Parent's Labor Supply: Evidence from Exogenous Variation in Family Size" *American Economic Review* 88(3) 450-77.

Angrist, Joshua D., 2004. "Treatment Effects Heterogeneity in Theory and Practice" *Economics Journal* 114(494) C52-C83.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin, 1996. "Identification of Causal Effects using Instrumental Variables" (with discussion) *Journal of the American Statistical Association* 91 444-72.

Bertanha, Marinho and Guido W. Imbens, 2014. "External Validity in Fuzzy Regression Discontinuity Designs", Working Paper, Stanford University, December.

Black, Dan A., Seth G. Sanders, Evan J. Taylor, and Lowell J. Taylor, 2015. "The Impact of the Great Migration on the Mortality of African-Americans: Evidence from Deep South" *American Economics Review* February 105(2) 477-503.

Black, Dan A. and Jeffrey A. Smith, 2004. "How Robust is the Evidence on the Effects of College Quality? Evidence from Matching" *Journal of Econometrics* July/August 121(1-2) 99-121.

Bloom, Howard S., 1984. "Accounting for No-Shows in Experimental Evaluation Designs" *Evaluation Review* 8(2) 225-46.

Bloom, Howard S., Larry L. Orr, Stephen H. Bell, George Cave, Fred Doolittle, Winston Lin, Johannes M. Bos, 1997. "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Study" *Journal of Human Resources* 32(3) 549-76.

Blundell, Richard, Lorrain Dearden, and Barbara Sianesi, 2005. "Evaluating the Effect of Education on Earnings: Models, Methods and Results from the National Child Development Survey" *Journal of the Royal Statistical Association* 167(3) 473-512.

Busso, Matias, John DiNardo, Justin McCrary, 2014. "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators" *Review of Economics and Statistics* 96(5) 885-897.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik, 2009. "Dealing with the Limited Overlap in Estimation of Average Treatment Effects" *Biometrika* 96(1) 187-99.

Diaz, Juan J. and Sudhanshu Handa, 2006. "An Assessment of Propensity Score Matching as a Nonexperimental Estimator" *Journal of Human Resources* 41(2) 319-45.

Eberwein, Curtis, John C. Ham, and Robert J. LaLonde, 1997. "The Impact of Being Offered and Receiving Classroom Training on the Employment Histories of Disadvantage Women: Evidence from Experimental Data" *Review of Economic Studies* 64(4) 655-82.

Heckman, James J., 1979. "Sample Selection Bias as a Specification Error" *Econometrica* 47(1) 206-248.

Heckman, James J. 1996. "Randomization as an Instrumental Variable" *Review of Economics and Statistics* 78(2) 336-340.

Heckman, James J., N. Hohmann, Jeffrey A. Smith, and M. Khoo, 2000. "Substitution and Drop Out Bias in Social Experiments: A Study of an Influential Social Experiment" *Quarterly Journal of Economics* 115(2) 651-94.

Heckman, James J., Hidehiko Ichimura, Jeffrey A. Smith, and Petra Todd, 1998. "Characterizing Selection Bias Using Experimental Data" *Econometrica* 66(5) 1017-98.

Heckman, James J., Hidehiko Ichimura and Petra E. Todd, 1997. "Matching As An Econometric Estimator: Evidence from Evaluating a Job Training Program" *Review of Economic Studies* 64(4) 487-36.

Heckman, James J., Hidehiko Ichimura and Petra E. Todd, 1998. "Matching As An Econometric Estimator *Review of Economic Studies* 65(2) 261-94.

Heckman, James J., Robert J. Lalonde and Jeffrey A. Smith, 1999. "The Economics and Econometrics of Active Labor Market Programs,"' in *Handbook of Labor Economics*, Volume 3, eds. Orley Ashenfelter and David Card. Amsterdam: North-Holland. 1865-2097.

Heckman, James J. and Navarro, 2004. "Using Matching, Instrumental Variables, and Control Function to Estimate Economics Choice Models" *Review of Economics* 86(1) 30-57.

Heckman, James J. and Jeffrey A. Smith. 1998. "Evaluating the Welfare State" in *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial*, ed. Steiner Strom. Cambridge University Press for Econometric Society Monograph Series. 241-318.

Heckman, James J., Sergio Urzua, and Edward J. Vytlacil, 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity" *Review of Economics and Statistics* 88(3) 389-432.

Heckman, James J. and Edward J. Vytlacil, 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation" *Econometrica* 73(3) 669-738.

_____, 2007a. "Econometric Evaluation of Social Programs, Part1: Causal Models, Structural Models and Econometric Policy Evaluation" *Handbook of Econometrics* vol 6B, eds. Jame J. Heckman and Edward E. Leamer, 4780-4873.

_____, 2007b. “Econometric Evaluation of Social Programs, Part2: Using Marginal Treatment Effect to Organize Alternative Estimators to Evaluate Social Programs and to Forecast their Effects in New Environments” *Handbook of Econometrics* vol 6B, eds. James J. Heckman and Edward E. Leamer, 4875-5143.

Heckman, James J., Daniel Schmierer, and Sergio Urzua, 2010. “Testing the Correlated Random Coefficients Model” *Journal of Econometrics* 158(2) 177-203.

Huber, Martin, 2013. “A Simple Test for Ignorability of Noncompliance in Experiments” *Economic Letters* 120(3) 389-91.

Huber, Martin, Michael Lechner, and Connie Wunsch, 2013. “The Performance of Estimators Based on the Propensity Score” *Journal of Econometrics* 175(1) 1-21

Imbens, Guido W. and Joshua D. Angrist, 1994. “Identification and the Local Average Treatment Effects,” *Econometrica* 62(2) March 1994: 467-476.

Imbens, Guido W., 2004. “Nonparametric Estimation of the Average Treatment Effects Under Exogeneity: A Review” *Review of Economics and Statistics* 86(1) 4-29.

Imbens, Guido W. and Thomas Lemieux, 2008. “Regression Discontinuity Designs: A Guide to Practice” *Journal of Econometrics* 142(2) 615-635.

Joo, Joonhwi, and Robert LaLonde. 2014. “Testing for Selection Bias” IZA Discussion Paper No. 8455.

Lee, David S. and Thomas Lemieux, 2010. “Regression Discontinuity Designs in Economics” *Journal of Economic Literature* 48(2) 281-355

Murphy, Kevin M. and Robert H. Topel, 2002. “Estimation and Inference in Two-step Models” *Journal of Business and Economic Statistics* 20(1) 88-97.

Romano, Joseph P., and Azeem Shaikh, 2006. “Stepup Procedures for Control of Generalization of Familywise Error Rates” *Annals of Statistics* 34(4) 1850-73.

Romano, Joseph P. and Michael Wolf, 2005. “Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing” *Journal of the American Statistical Association* 100(469) 93-108.

Roy A. D., 1951 “Some Thoughts on the Distribution of Earnings” *Oxford Economics Papers* 3(2) 135-146.

Smith, Jeffrey A. and Petra E. Todd, 2005. “Does Matching Overcome the LaLonde’s Critique of Nonexperimental Estimators?” *Journal of Econometrics* 125(1) 305-53.

Smith, Jeffrey and Alexander Whalley. 2015. "How Well Do We Measure Public Job Training?" Unpublished manuscript, University of Michigan.

Vytlacil, Edward J., 2000. *Three Essays on the Nonparametric Evaluation of Treatment Effects* Dissertation, University of Chicago.

_____, 2002. “Independence, Monotonicity, and Latent Index Models: An Equivalence Result.” *Econometrica*, 70(1) 331–41.