Simple Tests for Selection:

Learning More from Instrumental Variables

Dan A. Black
University of
Chicago,
IZA, and NORC

Joonhwi Joo
University of
Chicago

Robert LaLonde
University of
Chicago
and IZA

Jeffrey A. Smith
University of
Michigan,
NBER, IZA and
CESifo

Evan J. Taylor
University of
Chicago

First Draft: December 5, 2014

Current Draft: September 14, 2017

*We thank seminar participants at the University of Chicago, the University of Illinois at Chicago, the University of North Carolina at Chapel Hill, IZA, and the Hitotsubashi Institute for Advanced Study, along with Josh Angrist, Ben Feigenberg, Anthony Fisher, Anthony Fowler, Darren Lubotsky, and Gerard van den Berg for helpful comments. We thank Bill Evans for providing us with the data from Angrist and Evans (1998).

Simple Tests for Selection: Learning More from Instrumental Variables

Abstract

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 impact of training on adult women from the Job Training Partnership Act (JTPA) experiment.

Key words: instrumental variable, local average treatment effect, selection, test

1. Introduction

In the 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 the assumption that instruments have a "monotonic" impact on behavior. Put differently, the instruments must induce all agents to behave in a weakly uniform manner when subjected to a change in the value of the instrument. Informally, 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 the IV and Ordinary Least Squares (OLS) estimates is not valid in a LATE framework. More broadly, the relationship between the OLS and IV estimates becomes less informative about the existence of selection within the LATE framework. Thus, researchers face the paradox of using IV estimation to correct for selection on unobserved variables, but with no clear evidence to demonstrate that such selection exists.

To see why, consider the framework of Angrist et al. (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:

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

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

$$C = \{i : D_i(Z_i = 1) = 1; D_i(Z_i = 0) = 0\},$$

where $D_i = D(Z_i)$ denotes the treatment choice of agent "*i*" as a function of Z_i, with $D_i = 1$ for treatment and $D_i = 0$ for no treatment. In this framework, the Wald estimator corresponds to a LATE estimator

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

where Y_{1i} denotes the treated potential outcome of the i^{th} agent, Y_{0i} denotes the untreated potential outcome of the i^{th} agent, and $Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$ denotes the observed outcome.

Selection on unobserved variables means that one or more of these four conditions fails:

$$E(Y_{0i} \mid N) = E(Y_{0i} \mid C) = E(Y_{0i} \mid A);$$

$$E(Y_{1i} \mid N) = E(Y_{1i} \mid C) = E(Y_{1i} \mid A).$$

These conditions do not imply the equivalence of the OLS and IV estimands, which could differ for a number of reasons including omitted subgroup interactions combined with a different distribution of compliers across subgroups than among the always and/or never takers. Nor does equivalence of the estimands imply the conditions. To see this consider the following example: Suppose that $P(C) = P(A) = P(N) = 1/3$ and that $P(Z) = 1/2$. 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 expected value of the OLS estimate of the impact of treatment equals zero. The expected value of the IV estimate, however, also equals zero, but $E(Y_1 \mid A) > E(Y_1 \mid C)$ and $E(Y_0 \mid N) > E(Y_0 \mid C)$ so we clearly have 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. These tests consider two of the four conditions above; the data do not provide information regarding the other two. As such, we test necessary but not sufficient conditions for the absence of selection. For simplicity we focus on the case of conditional mean independence; the straightforward generalization to full conditional independence does not add any substantive insights. Relative to the traditional Durbin-Wu-Hausman test that compares the IV and OLS estimates, our tests reveal substantively relevant information regarding whether selection occurs on the treated outcome, the untreated outcome, or both.

Drawing on the work of Black et al. (2015), our tests come in two forms. First, conditional on covariates, we compare the outcomes of the set of agents who comply with the instrument 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 effect as well.

Our tests resemble those in Heckman's (1979) seminal paper on the bivariate normal selection model. In the two-step estimator for the normal selection model with a common treatment effect, the inverse Mills' ratio represents the control function, and the coefficient on the inverse Mills' 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 asks if the coefficient on the inverse Mills' 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 as in the estimators examined in Newey, Powell, and Walker (1990) and the related literature, but the nature of the test remains the same.

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

Our paper is closely related to Heckman et al. (2010), hereinafter HSU, who derive both parametric and nonparametric tests for the correlated random coefficient model. Formally, HSU develop a test for the independence of treatment status and the idiosyncratic effect of treatment conditional on covariates. Drawing on the work of Heckman and Vytlacil (2005, 2007a,b), who show that conditional independence of Y_0 and Y_1 implies constant marginal treatment effects, 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. But as HSU note, their 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.

Similarly, in the context of a Marginal Treatment Effects (MTE) model, Brinch, Mogstad, and Wiswall (2017) propose testing for a constant MTE by regressing

$Y_i = D_i Y_{0i} + (1 - D_i)(Y_{1i} - Y_{0i})$ against D_i, Z_i and their interactions. As Brinch, Mogstad, and Wiswall note, the extension of their test to a model with covariates is straightforward, but for a linear parametric model the test would involve the estimation of a very large number of interaction terms.

Bertanha and Imbens (2014) consider closely related 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 a Wald test for the exogeneity of noncompliance in experiments closely related 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 the 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. Guo et al. (2014) provide a substantially more complex test in the IV context. Battistin and Rettore (2008) and Costa Dias et al. (2013) consider related tests that exploit particular empirical contexts.

We find it peculiar that while the LATE revolution has led to a more sophisticated interpretation of IV estimates, researchers rarely make an empirical case via testing for the use of instrumental variables methods. Heckman et al. (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 lack of common support and from differences in the distributions of covariates, leaving selection on unobserved variables to account for only about seven percent of the difference. Blundell et al. (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 in the same way, Diaz and Handa (2006) report that their propensity score matching estimates matched the experimental evidence from the famous PROGRESA experiment with their set of conditioning variables. While by no means conclusive, matching on a rich set of covariates motivated by theory and the institutional context, and limiting the analysis to the region of common support between treated and untreated units, substantially reduce bias in many substantive contexts. Put differently, sometimes unconditional differences in mean outcomes between treated and untreated units arise mainly from selection on observed variables and lack of common support rather than from selection on unobserved variables. Indeed, given the necessary, and often empirically quite large, increase in the variance of estimates when using instrumental variables methods, a researcher might well prefer a precise but modestly biased OLS estimate to a consistent but imprecise IV estimate, much as in nonparametric estimation, a researcher trades off bias and variance via the choice of a bandwidth or other tuning parameter.

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, 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. Our presentation builds on Heckman et al. (1997), Heckman and Smith (1998), and Heckman et al. (1999). Using the notation introduced above, we define the causal impact of the treatment on agent “*i*” 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 the literature refers to as the “missing counterfactual.”

Matching estimators represent one intuitive class of estimators for generating the missing counterfactuals. These estimators rely on the assumption that researchers have sufficiently rich covariates that any differences in the treatment decisions of the agents are independent of the agents’ potential outcomes conditional on the covariates. Let *X* denote those covariates. Formally, matching estimators rely on two assumptions. First, and most vexing, matching estimators require the Conditional Independence Assumption (CIA) for a causal interpretation. Various “flavors” of the CIA correspond to different parameters of interest. The strongest flavor demands that:

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

where “ $\perp$ ” denotes statistical independence. This version of the CIA applies to the Average Treatment Effect (ATE), or

$$\Delta^{ATE} = E(Y_{1i} - Y_{0i}) \tag{2}$$

The CIA for Y_0 assumes

$$Y_{0i} \perp D_i \mid X_i \tag{CIA^0}$$

This version of the CIA allows the estimation of

$$\Delta^{ATET} = E(Y_{1i} - Y_{0i} \mid D_i = 1) \tag{3}$$

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

$$Y_{1i} \perp D_i \mid X_i \tag{CIA^1}$$

which allows the estimation of

$$\Delta^{ATEN} = E(Y_{1i} - Y_{0i} \mid D_i = 0) . \tag{4}$$

Of course, the (CIA) implies that both (CIA^1) and (CIA^0) hold.

Second, matching estimators also require the Common Support Assumption (CSA) or

$$0 < \Pr(D_i = 1 \mid X_i) < 1 \ . \tag{CSA}$$

In other words, the CSA simply requires an untreated comparison unit with approximately the same realization of the covariates as each treated unit and, when estimating Δ^{ATE} rather than Δ^{ATET}, vice versa. Of course, the CSA is a testable assumption. When the CSA fails in practice, as it sometimes does, researchers generally change the definition of the relevant population to that over which the CSA holds, reflecting the limited variation that the data provide; see the discussions in Black and Smith (2004) and Crump et al. (2009).

When applying semiparametric or nonparametric matching methods, researchers commonly specify the functions that determine the potential outcomes (Y_{0i}, Y_{1i}) as

$$Y_{1i} = g_1(X_i) + \varepsilon_{1i} \tag{5}$$

$$Y_{0i} = g_0(X_i) + \varepsilon_{0i} \tag{6}$$

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

A common alternative to matching methods uses OLS to estimate parametric linear models. In this approach, researchers specify the functional form of the conditional mean function as

$$g_1(X_i) = X_i'\beta_1 \tag{7}$$

$$g_0(X_i) = X_i'\beta_0 \ . \tag{8}$$

In these models, the researcher avoids invoking the CSA (but not the CIA) by instead making assumptions about the functional form.

The common criticism of estimates obtained by matching or by OLS estimation of a parametric linear model is that they rely on the CIA, which appears implausible in many substantive contexts given the available data on conditioning variables. To avoid making the CIA, applied researchers often turn to IV estimation. While traditional IV methods require a common treatment effect for all units, Imbens and Angrist (1994) demonstrate that under different assumptions IV estimation allows for heterogeneous treatment effects. Researchers now routinely invoke their LATE framework when applying IV methods.

It is difficult to overemphasize the importance of this advance. Models that omit selection into treatment based upon (possibly very partial) knowledge of heterogeneous treatment effects seem incapable of capturing the complexity of human behavior. Incorporating such treatment effect heterogeneity allows researchers to consider and estimate far more plausible and interesting models, including the justifiably famous Roy (1951) model. Indeed, Heckman et al. (2006) term such heterogeneous impacts “essential heterogeneity.”

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

We consider the possible decisions of agent "*i*" for any value of Z_i, which is the set $\{D_i(z) \mid z \in \mathbb{Z}\}$. We may now state the assumptions of the LATE estimator as the Existence of Instruments (EI) and Monotonicity (M). Formally,

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

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

The (M) assumption requires that all agents respond to the instrument in the same direction, not that the function $\Pr(D_i = 1 \mid X_i, Z_i)$ be monotone in Z; this led Heckman et al. (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 a mixture of treatment effects associated with agents who both enter and leave the treatment as the instrument varies. To keep the notation simple, we continue to assume $Z_i \in \{0,1\}$; our arguments, however, generalize to continuous instruments.

Imbens and Angrist note 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

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

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

where $1(\cdot)$ is an indicator function for the logical condition inside the parentheses 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

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

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

where $P(Z_i, X_i) = \Pr(D_i = 1 \mid X_i, Z_i)$ is the conditional probability of treatment or propensity score. Unlike the propensity score used by propensity score matching estimators under the CIA, this propensity score also includes at least one instrument; see Heckman and Navarro-Lazano (2004) for further discussion. We denote the control functions that embody the conditional means of ε_1 and ε_0 by $c_1(\cdot)$ and $c_0(\cdot)$; including them in the conditioning implies $E(e_{1i}) = E(e_{0i}) = 0$.

The control function approach allows an easier interpretation of the independence assumption $(Y_{0i}, Y_{1i}, \{D_i(z) \mid \mathrm{z} \in \mathbb{Z})\} \perp Z_i \mid X_i)$ 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 typically invoke them without looking for evidence of selection on unobserved variables. In contrast, the VIA assumptions allow researchers to consistently estimate LATEs for those individuals who comply with the instruments. In the case of a linear parametric model with a single instrument we would augment equations (7) and (8) to obtain

$$E(Y_{1i} \mid X_i, Z_i, D_i = 1) = X_i\beta_1 + \alpha_1 Z_i \tag{11}$$

$$E(Y_{0i} \mid X_i, Z_i, D_i = 0) = X_i\beta_0 + \alpha_0 Z_i . \tag{12}$$

With non-binary instruments researchers may wish to add higher order terms – replace Z_i with $f(Z_i)$ – though this raises subtle but important issues of model selection that lie outside the scope of this paper. With multiple instruments, researchers would probably want to replace Z_i with the estimated propensity score, $\hat{p}(Z_i, X_i)$ and adjust the standard errors for generated regressors as in Murphy and Topel (2002). Joo and LaLonde (2014) present a control function version of our test along these lines.

The model behind equations (11) and (12) represents an important departure from the canonical model used in IV applications, given by

$$Y_i = X_i\beta + \delta D_i + \varepsilon_i \tag{13}$$

In contrast, the model underlying equations (11) and (12) consists of

$$Y_{1i} = X_i\beta_1 + \varepsilon_{1i} \tag{14}$$

$$Y_{0i} = X_i\beta_0 + \varepsilon_{0i}\,. \tag{15}$$

Equations (14) and (15) allow researchers to estimate heterogeneous treatment effects, $\Delta(X_i)$, that differ with the realization of the covariates while still maintaining the CIA. There is, of course, generally no theoretical reason to prefer equation (13) to equations (14) and (15), but the demands on instrument strength usually dissuade researchers from using the model described by equations (14) and (15) when they resort to IV estimation because they fear selection on unobserved variables.

In the case of matching estimators, we would augment equations (5) and (6) and specify the conditional mean functions as

$$E(Y_{1i} \mid X_i, Z_i, D_i = 1) = g_1(X_i) + \alpha_1 Z_i \qquad (16)$$
$$E(Y_{0i} \mid X_i, Z_i, D_i = 1) = g_0(X_i) + \alpha_0 Z_i \qquad (17).$$

To clarify the relationship among the various forms of the CIA and our test, it is useful to outline the samples used and hypotheses involved when estimating these auxiliary regressions. Formally, we estimate (12) or (17) 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 (16) 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 $D_i(z_i = 1) \geq D_i(z_i = 0)$ and divide agents under the VIA into the three types defined in the introduction: "always takers," "never takers," and "compliers." The test given in either equation (11) or equation (16) simply compares $E(Y_1 | x, A)$ to $E(Y_1 | x, C)$. As Black et al. (2015) note, this is easily done because $E(Y_1 | x, z = 0) = E(Y_1 | x, A)$ and $E(Y_1 | x, z = 1) = E(Y_1 | x, A \cup C)$. Thus, at $X = x$ we have that

$$\alpha_1(x) \equiv \frac{\Pr(C | x)(E(Y_{1i} | x, C) - E(Y_{1i} | x, A))}{\Pr(C | x) + \Pr(A | x)} .$$

The regression coefficient in either equation (11) or equation (16) 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 non-constant control function in equation (9), which constitutes evidence that unobserved variables affect the outcomes. A parallel argument applies to α_0.

The finding that either $\alpha_0 \neq 0$ or $\alpha_1 \neq 0$ constitutes evidence of either selection or violation of the exclusion restrictions (i.e., the failure of EI) or both. Assuming the validity of the exclusion restriction, rejection of one or both of the null hypotheses provides simple and compelling evidence for violation of the CIA. 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 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 signs of the relevant selection effects and their magnitudes. This allows researchers to provide a much more nuanced discussion of the nature of the agents' choice behavior. Given the equivalence that Vytlacil (2000, 2002) demonstrates, it is perhaps not surprising that we may learn more about the selection problem using IV methods than we learn from current practices.

In the next two subsections, we show how to adapt these tests 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, treatment depends on a running variable S_i and has the feature that the probability of treatment jumps (i.e. has a discontinuity) at some particular value of S_i. We assume that the jump in the probability of treatment occurs at $S_i = 0$. To use both of our tests we require fuzziness on both sides of the discontinuity; put differently, we need both treated and untreated units on both sides of the cutoff. Formally, we require

$$1 > \lim_{S \downarrow 0} \Pr(D = 1 \mid X, S) > \lim_{S \uparrow 0} \Pr(D = 1 \mid X, S) > 0$$

or the same condition but with the two limits reversed. With both treated and untreated units on only one side of the discontinuity, a researcher can apply our test for one of (Y_0, Y_1) but not both. As emphasized by Imbens and Lemieux (2008) and Lee and Lemieux (2010), when faced with a fuzzy RD, researchers who use the discontinuity at $S = 0$ as an instrument for treatment estimate a LATE at $S = 0$.

Because of the discrete change in the treatment probability at $S = 0$, under selection on unobserved variables we would expect a jump in the control function at the same point. More formally, selection on unobserved variables implies a jump in the value of $E(Y_0 \mid S, D = 0)$ as S

crosses zero, while the (CIA^0) assumption implies a smooth $E(Y_0 \mid S, D=0)$ function around zero. This suggests a simple test based on a model of the form

$$E(Y_{0i} \mid D_i = 0) = g_0(X_i, S_i) + \alpha_0 1(S_i \geq 0) \tag{18}$$

with the null hypothesis being that $\alpha_0 = 0$ or the corresponding version for testing CIA^1

$$E(Y_{1i} \mid D_i = 1) = g_1(X_i, S_i) + \alpha_1 1(S_i \geq 0) \tag{19}$$

with the null hypothesis being that $\alpha_1 = 0$. Estimation of the sample analogue of (18) makes use only of untreated observations, which constitute a mixture of compliers and never takers. Similarly, estimation of the sample analogue of (19) uses only treated observations, which constitute a mixture of compliers and always takers.

As noted in the introduction, Bertanha and Imbens (2014) consider closely related 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, random assignment creates an instrument for treatment. Because many social experiments have imperfect compliance – Heckman et al. (2000) lists numerous examples – with both treatment group dropout and control group substitution into similar treatments provided elsewhere, one could easily implement our tests to check for selection on Y_1 or Y_0 in experiments. For instance, Table II of Heckman et al. (2000) reports that, among those recommended for classroom training prior to random assignment, somewhere between 49 and 59 percent of the treatment group in the Job Training Partnership Act (JTPA) experiment received services, depending on the demographic group, while between 27 and 40 percent of the control group 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 random assignment to the treatment group as an instrument for the receipt of treatment. As random assignment provides a binary instrument, the Wald estimator recovers the LATE for those who comply with the experimental protocol. Huber (2013) provides a Wald test of the exogeneity of non-compliance in experiments closely related to our own analysis.

D. Recovering Estimates of the Magnitude of the Selection Effect

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

$$E(Y_{0i} \mid Z_i = 0, D_i = 0) = \frac{\Pr(C)}{\Pr(C)+\Pr(N)} E(Y_{0i} \mid C) + \frac{\Pr(N)}{\Pr(C)+\Pr(N)} E(Y_{0i} \mid N) \tag{20}$$

while

$$E(Y_{0i} \mid Z_i = 1) = E(Y_{0i} \mid N) \tag{21}$$

so that

$$\alpha_0 \equiv E(Y_{0i} \mid Z_i = 1, D_i = 0) - E(Y_{0i} \mid Z_i = 0, D_i = 0) = \frac{\Pr(C)}{\Pr(C)+\Pr(N)} (E(Y_{0i} \mid N) - E(Y_{0i} \mid C)) \text{ .} \tag{22}$$

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

$$B_0 = E\left(Y_0 \mid N\right) - E\left(Y_0 \mid C\right) = \frac{\Pr(C)+\Pr(N)}{\Pr(C)} \alpha_0 . \tag{23}$$

Similarly, we have

$$E(Y_{1i} \mid Z_i = 1, D_i = 1) = \frac{\Pr(C)}{\Pr(C)+\Pr(A)} E(Y_{1i} \mid C) + \frac{\Pr(A)}{\Pr(C)+\Pr(A)} E(Y_{1i} \mid A) \tag{24}$$

while

$$E(Y_{1i} \mid Z_i = 0, D_i = 1) = E(Y_{1i} \mid A) \tag{25}$$

so that

$$\alpha_1 \equiv E(Y_{1i} \mid Z_i = 1, D_i = 1) - E(Y_{1i} \mid Z_i = 0, D_i = 1) = \frac{\Pr(C)}{\Pr(C)+\Pr(A)} (E(Y_{1i} \mid C) - E(Y_{1i} \mid A)) \ . \tag{26}$$

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

$$B_1 = E\left(Y_1 \mid C\right) - E\left(Y_1 \mid A\right) = \frac{\Pr(C)+\Pr(A)}{\Pr(C)} \alpha_1 \ . \tag{27}$$

To implement these measures empirically, we may use the OLS estimates of (α_0, α_1). We know that $\Pr(A) = \Pr(D_i = 1 \mid \mathrm{Z}_i = 0)$, $\Pr(N) = \Pr(D_i = 0 \mid Z_i = 1)$, and $\Pr(C) = 1 - \Pr(N) - \Pr(A)$ so we have sample analogues of all the terms on the right-hand sides of equations (23) and (27).

E. Complications with Instruments

Applied researchers know from painful experience the numerous complications that attend IV estimation. In this subsection, we discuss how three of these complications affect our test: weak instruments, failure of the monotonicity (M) assumption, and failure of the assumption of the existence of instruments (EI) (or exclusion restriction) assumption.

Weak instruments imply a relatively small number of compliers which in turn implies relatively low power for our test. Put differently, with a weak instrument, comparing the conditional means of, say, always takers and compliers will provide only noisy evidence regarding the null of no selection in the absence of a very large selection effect, a very large sample, or both.

The case for the monotonicity (M) assumption typically rests on some combination of institutional knowledge and economic theory specific to a particular empirical context. Our test provides no help in detecting failures of the (M) assumption. When it does fail, the untreated units include what Angrist et al. (1996) call defiers, agents who change treatment status when the value of the instrument changes but in an unexpected way, in addition to compliers and never takers. Similarly, the treated units now comprise always takers, compliers and defiers. The presence of the defiers undoes the LATE interpretation of the IV estimand.

Finally, because our test implicitly represents a joint test of the (EI) assumption and the null of no selection bias, failure of the (EI) assumption can lead to incorrect inferences regarding the presence or absence of selection. When failure of the EI assumption leads the test to reject the joint null, researchers may proceed to place heavy weight on the IV estimates, when in fact they provide an unknown mixture of the population treatment effect and the bias associated with the invalid instrument.

5. Empirical Applications

A. Angrist and Evans (1998) data

Our first application draws on Angrist and Evans (1998). This paper uses data from the 1980 and 1990 US Censuses to measure the causal impact of children on maternal labor supply. Because fertility is likely to be endogenous with respect to women's labor supply decisions, Angrist and Evans devise 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 six 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.

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 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 in 50.6% of families the first two children are of the same sex. Formally, the system that Angrist and Evans estimate is:

$$y_i = x_i'\beta + \delta morekids_i + \varepsilon_i \tag{28}$$

$$morekids_i = x_i'b + \gamma samesex_i + u_i \tag{29}$$

where *morekids* is an indicator for having more than two children. The covariates 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 two boys or two 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. We set all of these variables to zero for women who 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

$$y_i = g(x_i) + \delta morekids_i + \varepsilon_i \tag{30}$$

$$morekids_i = h(x_i) + \gamma samesex_i + u_i \tag{31}$$

where $g(\cdot)$ and $h(\cdot)$ are unknown functions. Because we have only discrete conditioning variables, we estimate $g(x_i)$ by a fully saturated regression. 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.

Of course, to interpret the IV estimand as a LATE 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 provides an exclusion restriction in the sense that having the first two children of the same sex does not directly affect 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 (strong) assumptions, we may now implement our parametric tests of the CIAs using:

$$y_{0i} = x_i'\beta_0 + \alpha_0 samesex_i + \varepsilon_{0i} \tag{32}$$
$$y_{1i} = x_i'\beta_0 + \alpha_1 samesex_i + \varepsilon_{1i} \tag{33}$$

and our semiparametric tests using

$$y_{0i} = g_0(x_i) + \alpha_0 samesex_i + v_{0i} \tag{34}$$
$$y_{1i} = g_1(x_i) + \alpha_1 samesex_i + v_{1i} \tag{35}$$

where for our semiparametric analysis we need to drop the indicator for having a boy as the second child in order to avoid making the *samesex* variable perfectly collinear with the x_i vector.

We estimate equations (32) and (34) on the sample of 236,092 women who have two children, and equations (33) and (35) using the sample of 158,743 who have three or more children.

Table 2 presents our results. 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. In each case, we estimate a positive coefficient α_0 on *Z*, where *Z* = 1 among the non-treated corresponds to the never-takers. Thus, we find that the never takers have higher earnings, hours worked, and weeks worked, and are more likely to work at all conditional on our covariates relative to the compliers who do not have a third child.

In contrast, we find little evidence against the (CIA^{ATEN}). Unlike the estimates of α_0 , our estimates of α_1 are statistically insignificant and economically very small. Thus, we find no evidence of selection 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.

To describe the magnitude of the selection effects we use the nonparametric estimates in column (3) of Table 2. Compared to the compliers, we find that the never takers are five percentage points more likely to have worked last year, worked about three weeks more, worked about two hours more per week, and earned $1,965 more per year. Comparing the compliers to the always takers, we find that compliers were one percentage point more likely to work last year, they worked about 0.4 extra weeks per year, they worked a tenth of an hour more per week, and earned $38 dollars less per year than the always 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.

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}) in the context of a canonical model such as equation (13), which rules out by assumption any heterogeneous treatment effects associated with the covariates, X_i. Even with the relatively parsimonious equation (13) specification, the Angrist tests fails to detect selection in our four outcomes at the five percent level, although the tests do detect selection at a 10 percent level in two of the four outcomes.

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

When facing 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 difference in a dependent variable 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 using assignment to the treatment group as an instrument for the receipt of treatment. Because assignment to the treatment group is independent of the potential outcomes (Y_0, Y_1), it 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 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), consists of women recommended for classroom training (the "CT-OS treatment stream" in the jargon of the experiment) prior to random assignment. We measure training as the onset of self-reported classroom training within nine months of randomization. We focus on classroom training and ignore other (usually much less intensive) services, such as job search assistance, received by some members of both the treatment and control groups for simplicity. We rely on the self-reported training data for both groups, rather than the self-reports for the controls and the JTPA administrative data for the treatment group, for comparability; Smith and Whalley (2017) offer a depressing exploration of the concordance of administrative and self-reported measures of service receipt in the JTPA study data. The administrative data from the experiment provides the indicator for random assignment to the treatment group rather than the control group. For our outcome variable, we use an indicator for self-reported employment in the eighteenth month after random assignment. There was much non-compliance in this sample. Only about 65 percent of the treatment group reports receiving classroom training in the first nine months after random assignment. There was much control group substitution as well: about 34 percent of the control group reports receiving 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

$$y_i = \beta + \delta R_i + \varepsilon_i , \tag{36}$$

where y_i is the outcome variable, 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 estimated intent-to-treat parameter, $\hat{\delta}$, equals 0.041 and statistically differs from zero 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 differential arises, of course, because random assignment only increases the rate of treatment by about 0.305, the coefficient on the indicator for random assignment to the treatment group from the first-stage of our TSLS Bloom estimator.

Nothing in this analysis, however, informs the researcher regarding the presence or absence of selection into treatment based on unobserved variables. Toward that end, we next estimate the following equations

$$y_{0i} = \beta_0 + \alpha_0 R_i + \varepsilon_{0i} \tag{37}$$

$$y_{1i} = \beta_1 + \alpha_1 R_i + \varepsilon_{1i} \tag{38}$$

where (y_{0i}, y_{1i}) are the outcomes of those not receiving training and receiving training. We estimate Equation (37) using the 1,233 adult women who do not receive training and estimate equation (40) using the 1,501 who do receive training. We find little evidence that the compliers have different Y_0 than those who never take training. The coefficient on the indicator for assignment to the treatment group in equation (37) is small, 0.006, and statistically insignificant at the five-percent level. In contrast, the coefficient on the indicator for assignment to the treatment group in equation (38) 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.

A finding of substantively large selection on unobserved variables in the absence of covariates will hardly surprise most readers. Thus, we augment our equations with a vector of variables measuring the educational and demographic characteristics of those randomly assigned, as well as their pre-random assignment labor market activity and transfer payment receipt; see the notes 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. More surprisingly, the results of our tests for selection on unobserved variables also change very little when we add covariates. 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 falls by only about nine percent. The observed variables examined here account for little of the selection. In contrast, our test consistently fails to detect unobserved differences between the compliers and the never takers.

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. It generalizes various existing tests in the literature. Using a Wald-like estimator, one can use the estimates generated by our test to assess the magnitude of the selection effect as well and thereby gain a much better understanding of the precise nature of any selection on unobserved variables.

Magnitudes matter (in addition to signs and stars). As such, calculation of the selection effect may prove extremely valuable in many substantive contexts. As Cameron and Trivedi (2005, p. 107) reminds us, with one instrument we may compare the variance of the parameter of interest from the canonical model (equation (13)) estimated using instrumental variables to the variance of the parameter estimated using OLS using the equation:

$$SE(\hat{\delta}^{IV}) = \frac{SE(\hat{\delta}^{OLS})}{\rho(\tilde{D},\tilde{Z})}$$

where $\rho(\tilde{D},\tilde{Z})$ denotes the partial correlation coefficient after removing the variation correlated with the other covariates *X*, what Black and Smith (2006) term the “Yulized residuals” in honor of Yule’s (1907) brilliant paper. This partial correlation coefficient may be quite modest, 0.10 or even 0.05 depending on the strength of the instrument, suggesting dramatic decreases in the precision of the estimates when using IV methods. For example, Black et al. (2015) estimate that $\rho(\tilde{D},\tilde{Z})$ equals only 0.059 in their application. In many situations, serious researchers will trade off increases in bias for reductions in variance. Our method provides researchers with a means of assessing this bias and so allows them to make a quantitatively informed decision regarding whether or not the “IV” cure for the “OLS bias” disease is worse than the disease itself.

Since the publication of Vytlacil (2002), we have understood the equivalence between the assumptions necessary for the LATE interpretation of IV estimates and models of selection into treatment based on latent indices. But IV estimation has always seemed to provide less information about the nature of the selection effect 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 effect when using IV estimation.

Our two empirical applications nicely demonstrate the knowledge our tests produce. 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 more than two children despite the relatively modest set of covariates available in the census. In contrast, those who complied with the instrument and had at least one more child seem extremely different from those who always stop at two children. The labor market earnings of never takers exceed those of women who complied with the instruments by about $2,000 per year. Surprisingly, we find no (statistically or substantively) significant differences between the always takers and the compliers.

Our second application also yielded a surprise. We reanalyzed the probability of employment 18 months after random assignment for adult women in (one part of) the JTPA experiment. While we found a sizeable impact of training on the compliers (0.136 without covariates), we estimated an even larger selection effect, as the employment rate of compliers when trained exceeds that of always takers when trained by 0.145 (again, without covariates). Even after conditioning on an extensive set of predetermined covariates, the selection effect remained larger than the treatment effect on the compliers.

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*** (0.00162)	-0.120*** (0.0249)	-0.174*** (0.00164)	-0.117*** (0.0250)
Weeks worked	-8.97*** (0.0707)	-5.66*** (1.108)	-8.90*** (0.0727)	-5.53*** (1.109)
Hours worked	-6.66*** (0.0611)	-4.59*** (0.9452)	-6.59*** (0.0620)	-4.45*** (0.9461)
Income	-3,768*** (33.45)	-1,960*** (541.5)	-3,739 (35.47)	-1,915*** (542.0)
First Stage: Same sex coefficient	---	0.062*** (0.0015)	----	0.062*** (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 the age of the mother, the age of the mother at first birth, indicators for whether the mother is black or non-black and non-white, 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-statistic on the instrument for the parametric model equals 1,711. For the semiparametric model it equals 1,702. For the semiparametric model, 72 cases have predicted values of one for the probability of having more children and 168 have predicted probabilities 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)
[selection effect]	[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)
[selection effect]	[3.01]	[0.37]	[3.19]	[0.44]
N	236,092	158,743	236,092	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)
[selection effect]	[2.08]	[0.00]	[2.24]	[0.11]
N	236,092	158,743	236,092	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)
[selection effect]	[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 Recommended 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)
[selection effect]	[0.013]	[0.011]
(n=1,233)		
Treatment group indicator for Y_1 regression	0.068^{**}	0.062^{**}
(standard error)	(0.0318)	(0.0313)
[selection effect]	[0.145]	[0.132]
(n=1,501)		

Note: The dependent variable is an indicator variable for whether the participant is employed in the 18th month after random assignment. The treatment indicator equals 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 nine 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 indicate 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 General Education Development degree, whether the participant has more than a high school degree (high school degree is the excluded category), whether the participant was on Aid to Families with Dependent Children (AFDC) at the time of random assignment, whether the participant was on AFDC for two years or more, whether the participant was on food stamps at the time of random assignment, 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 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 experimental sites. 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” *Economic 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.

Battistin, Eric and Enrico Rettore. 2008. “Ineligible and Eligible Non-Participants as a Double Comparison Group in Regression Discontinuity Designs.” *Journal of Econometrics* 142 715-730.

Bertanha, Marinho and Guido W. Imbens, 2014. “External Validity in Fuzzy Regression Discontinuity Designs” NBER Working Paper No. 20773.

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 Economic Review* 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* 121(1-2) 99-121.

Black, Dan A. and Jeffrey A. Smith. 2006. “Estimating the Returns to College Quality with Multiple Proxies for Quality” *Journal of Labor Economics* 24(3) 701-28.

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 Society, Series A* 167(3) 473-512.

Brinch, Christian, Magne Mogstad, and Matthew Wiswall, 2017. “Beyond LATE with a Discrete Instrument.” *Journal of Political Economy* 125(4) 985-1039.

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.

Cameron, A. Colin and Pravin K.Trivedi, 2005. *Microeconometrics: Methods and Applications* Cambridge, UK: Cambridge University Press.

Costa Dias, Monica, Hidehiko Ichimura, and Gerard van den Berg. 2013. "Treatment Evaluation with Selective Participation and Ineligibles." *Journal of the American Statistical Association* 142 715-730.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik, 2009. "Dealing with 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 Disadvantaged Women: Evidence from Experimental Data" *Review of Economic Studies* 64(4) 655-82.

Guo, Zijian, Jing Cheng, Scott Lorch, and Dylan Small, 2014. "Using an Instrumental Variable to Test for Unmeasured Confounding" *Statistics in Medicine* 33(20) 3528-3546.

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., Neil Hohmann, Jeffrey A. Smith, and Michael. 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 Ichimra, and Petra Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme" *Review of Economic Studies* 64(4) 605-654.

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 Salvador Navarro-Lazano, 2004. "Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models" *Review of Economics and Statistics* 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.

Heckman, James J. and Edward J. Vytlacil, 2007a. “Econometric Evaluation of Social Programs, Part1: Causal Models, Structural Models and Econometric Policy Evaluation” *Handbook of Econometrics, Volume 6B*, eds. Jame J. Heckman and Edward E. Leamer, 4780-4873.

Heckman, James J. and Edward J. Vytlacil, 2007b. “Econometric Evaluation of Social Programs, Part 2: Using Marginal Treatment Effect to Organize Alternative Estimators to Evaluate Social Programs and to Forecast their Effects in New Environments” *Handbook of Econometrics, Volume 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 Estimation of Local Average Treatment Effects,” *Econometrica* 62(2) 467-475.

Imbens, Guido W., 2004. “Nonparametric Estimation of 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.

Newey, Whitney, James Powell, and James Walker. 1990. "Semiparametric Estimation of Selection Models: Some Empirical Results." *American Economic Review* 80(2): 324-328.

Romano, Joseph P., and Azeem Shaikh, 2006. "Stepup Procedures for Control of Generalizations of the Familywise Error Rate" *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) 94-108.

Roy Andrew 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 LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125(1) 305-53.

Smith, Jeffrey and Alexander Whalley. 2017. "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.

Vytlacil, Edward J., 2002. "Independence, Monotonicity, and Latent Index Models: An Equivalence Result" *Econometrica* 70(1) 331–41.

Yule, Udry, 1907. "On the Theory of Correlation for Any Number of Variables, Treated by a New System of Notation" *Proceedings of the Royal Society* 79(529) 181-93.