NBER WORKING PAPER SERIES

SIMPLE TESTS FOR SELECTION: LEARNING MORE FROM INSTRUMENTAL VARIABLES

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

Working Paper 30291
http://www.nber.org/papers/w30291

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2022

We thank seminar participants at the CESifo Education Group Meetings (especially Georg Graetz and Edwin Leuven), Chicago, UIC, Indiana, IZA, Kentucky, UNC, UC-Riverside, and the Hitotsubashi Institute for Advanced Study, along with Josh Angrist, Chris Berry, Ben Feigenberg, Anthony Fisher, Anthony Fowler, Martin Huber, David Jaeger, Darren Lubotsky, Lois Miller, Nikolas Mittag, Douglas Staiger, and Gerard van den Berg for helpful comments. We thank Bill Evans for providing us with the data from Angrist and Evans (1998). Sadly, Bob LaLonde passed away while we were working on a revision of this paper. He is greatly missed. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Simple Tests for Selection: Learning More from Instrumental Variables
Dan A. Black, Joonhwi Joo, Robert LaLonde, Jeffrey A. Smith, and Evan J. Taylor
NBER Working Paper No. 30291
July 2022
JEL No. C26,C52,C93

ABSTRACT

We provide simple tests for selection on unobserved variables in the Vytlacil-Imbens-Angrist framework for Local Average Treatment Effects (LATEs). Our setup allows 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. We show that it applies to the standard binary instrument case, as well as to experiments with imperfect compliance and fuzzy regression discontinuity designs, and we link it to broader discussions regarding instrumental variables. We illustrate the substantive value added by our framework with three empirical applications drawn from the literature.

Dan A. Black
Harris School
University of Chicago
1307 E. 60th St.
Chicago, IL 60637
danblack@uchicago.edu

Joonhwi Joo
University of Texas at Dallas
800 W. Campbell Rd.
JSOM 13-326
Richardson, TX 75080
joonhwi.joo@utdallas.edu

Robert LaLonde
University of Chicago

Jeffrey A. Smith
Department of Economics
University of Wisconsin-Madison
1180 Observatory Drive
Madison, WI 53706-1393
and IZA
and also NBER
econjeff@ssc.wisc.edu

Evan J. Taylor
Department of Economics
University of Arizona
Tucson, AZ 85718
evantaylor@arizona.edu

1 Introduction

In the years since the publication of Imbens and Angrist (1994), applied researchers have embraced the interpretation of Instrumental Variables (IV) estimators as measuring the impact of treatment on the subset of respondents who comply with the instrument. For binary instruments, Imbens and Angrist call this the Local Average Treatment Effect, or LATE. The LATE framework allows researchers to consistently estimate interpretable causal parameters in the presence of treatment effect heterogeneity.

The LATE framework, however, 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, relative to approaches based on conditional independence, LATE identifies a treatment effect only for a very specific sub-population of the treated. Third, treatment effect heterogeneity invalidates the traditional Durbin-Wu-Hausman test for selection: The relationship between the Ordinary Least Squares (OLS) and IV estimands becomes completely uninformative about the existence of selection within the LATE framework.[1] Thus, researchers face the paradox of using IV to address potential selection on unobserved variables, but with no clear evidence that such selection exists.

Consider the binary treatment framework of Angrist, Imbens, and Rubin (AIR, subsequently) (1996), with a binary instrument $Z_i \in \{0, 1\}$ and with $D_i(z_i)$ indicating the treatment choice of agent i when $Z_i = z_i$. 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 (A), defined as the sub-population with $D_i(1) = D_i(0) = 1$, the never-takers (N), defined as the sub-population with $D_i(1) = D_i(0) = 0$, and the compliers

[1]The "OLS estimand" equals the coefficient on D in a parametric linear model with Y as the dependent variable and X and D as independent variables. The "IV estimand" equals the coefficient on D in the same linear model estimated by two-stage least squares with Z as the instrument.

(C), defined as the sub-population with $D_i(1) = 1$ and $D_i(0) = 0$, where $D_i = D_i(Z_i)$ denotes the treatment choice of agent i as a function of the binary instrument Z_i, with $D_i = 1$ for treatment and $D_i = 0$ for no treatment. In this framework, the Wald estimand corresponds to a LATE or

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

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

Selection on unobserved variables in AIR's (1996) binary instrument and binary treatment framework means that at least one of the following four conditions fails:

$$E(Y_{0i}|N, X_i) = E(Y_{0i}|C, X_i) \text{ and } E(Y_{0i}|C, X_i) = E(Y_{0i}|A, X_i) \tag{2}$$

$$E(Y_{1i}|N, X_i) = E(Y_{1i}|C, X_i) \text{ and } E(Y_{1i}|C, X_i) = E(Y_{1i}|A, X_i) \tag{3}$$

where X_i denotes a vector of conditioning variables not affected by the treatment. These conditions do not imply the equivalence of the OLS and IV estimands: By construction, the set of compliers differs from the set of people receiving treatment.[2] Nor does equivalence of the estimands imply the conditions.[3] How then do we test for such selection?

In this paper, we provide a set of simple tests for the presence of selection on the treated and untreated outcomes and derive measures of the magnitude of the selection in each case. We care about the incidence and magnitude of selection because (i) they inform our understanding of the underlying behavior, (ii) they affect the reasonableness of generalizing impact estimates obtained via instrumental variables beyond the compliers, and (iii) they

[2]Technically, we would see equality of the OLS and IV estimands in the simple linear model only under empirically implausible conditions, such as homoskedastic errors plus equivalent distributions of X in the two relevant groups (e.g., compliers and always-takers) or homoskedastic errors plus no variation in the treatment effect with X.

[3]To see this, consider the following example: Suppose that $Pr(C) = Pr(A) = Pr(N) = 1/3$ and that $Pr(Z_i = 1) = 1/2$. Further, let $E(Y_{1i}|A) = 1$, $E(Y_{1i}|C) = 0$, $E(Y_{0i}|C) = 0$, and $E(Y_{0i}|N) = 1$. In this case, the OLS estimand equals zero. The IV estimand also equals zero but $E(Y_{1i}|A) > E(Y_{1i}|C)$ and $E(Y_{0i}|N) > E(Y_{0i}|C)$ so we clearly have selection on Y_{1i} and Y_{0i}.

affect the relative attractiveness of estimates based on IV and on conditional independence in particular substantive contexts. We illustrate (i) in our empirical applications and expand on (ii) and (iii) in the penultimate section of the paper.

Our tests consider two of the four conditions in equations (2) and (3) (or their distributional analogues) as the data do not provide information about $E(Y_{0i}|A, X_i)$ and $E(Y_{1i}|N, X_i)$. We provide a quick and simple diagnostic that would indicate the presence of selection on unobservables affecting the outcome; but failure to detect the presence of selection effects from our test does not prevent the existence of selection on unobservables per se. Put differently, we test necessary (but not sufficient) conditions for the absence of selection. We develop and apply tests of both conditional mean independence and conditional independence of distributions. 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, Sanders, Taylor, and Taylor (2015), our tests of conditional mean independence come in two forms. First, we compare the conditional mean outcomes of agents who comply with the instrument when not treated to those of agents who never take treatment. Second, we compare the conditional mean outcomes of agents who comply with the instrument when treated to those of agents who always take treatment. Mechanically, we implement these tests 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 adjustment, our tests allow researchers to assess the economic magnitude of any selection effects as well. Our distributional tests proceed along similar lines, but replace the dependent variable with other features of the outcome distribution.

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 equals zero. Allowing selection on both the treated and untreated outcomes in a trivariate normal framework, as in Björklund and Moffitt (1987), makes the parallel to our tests an exact one.[4]

Our paper is closely related to Heckman, Schmierer, and Urzua (2010), hereinafter HSU, who derive both parametric and non-parametric 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. Note that selection on the idiosyncratic component of the treatment effect implies selection on one or both of Y_{1i} and Y_{0i}, while selection on outcomes need not imply selection on the idiosyncratic component of the treatment effect.

Building on the work of Heckman and Vytlacil (2005, 2007a,b), who show that conditional independence of Y_{0i} and Y_{1i} 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 non-parametric tests suffer from low power in sample sizes common in empirical studies.[5] 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_{1i} and Y_{0i}.

Similarly, in the context of a Marginal Treatment Effects (MTE) model, Brinch, Mogstad, and Wiswall (2017) propose testing for a constant MTE by regressing $Y_i = Y_{0i} + D_i(Y_{1i} - Y_{0i})$ against D_i, Z_i and their interactions. As they note, the extension of their test to a model with covariates is straightforward, but for a linear parametric model the test could involve

[4]See Blundell, Dearden, and Sianesi, (2005) for implementation. In more general selection models that retain the common effect assumption but dispense with normality, 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). Yet the nature of the test remains the same.

[5]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) as well as the step-down method of multiple hypothesis testing in Romano and Shaikh (2006).

the estimation of a large number of interaction terms. Kowalski (2018a) describes alternative tests that relax the linearity assumptions in Brinch et al. (2017) and provides links to the literature in health economics that seeks to empirically distinguish moral hazard and adverse selection; Kowalski (2018b) applies her tests in the context of the Oregon Health Insurance experiment.

Three other papers build tests on the same basic intuition as our own: Bertanha and Imbens (2020) in the context of fuzzy regression discontinuity designs, de Luna and Johansson (2014) in the context of conventional instrumental variables analysis, and Huber (2013) in the context of experiments with imperfect compliance. Casting a broader net, Angrist (2004) proposes a test that compares the estimated treatment effect for compliers to an estimate obtained using the always-takers and the never-takers. His test does not distinguish among selection on one or both of Y_{1i} and Y_{0i} and assumes the magnitude of the treatment effect does not vary with covariates. Guo, Cheng, Lorch, and Smallet (2014) provide a substantially more complex test than ours in the IV context. Battistin and Rettore (2008) and Costa Dias, Ichimura, and van den Berg (2013) consider related tests that exploit particular empirical contexts. Donald, Hsu, and Lieli (2014) derive a test that uses differences between estimates of the Average Treatment Effect on the Treated (ATET) and LATE parameters in the context of one-sided non-compliance.[6]

We add to this literature in seven ways. First, we place the tests in Huber (2013), de Luna and Johansson (2014), Black, et al. (2015), and Bertanha and Imbens (2020) within a common framework that encompasses standard binary instruments, experiments with imperfect compliance, and fuzzy regression discontinuity designs. Second, we emphasize separate consideration of selection on Y_{1i} and Y_{0i}, which illuminates the underlying economics. Third, we develop and apply tests of conditional independence of distributions as well as tests of conditional mean independence. Fourth, we provide quantitative measures of selection that

[6]More distant relations include Kitagawa (2015) and Huber and Mellace (2015), both of which propose tests of instrument validity in the LATE context, as well as strategies that avoid the need for an instrument entirely, such as the bounds (a.k.a. set identification) in Manski (1990) or the sensitivity analyses proposed in Altonji, Elder and Taber (2005) and refined in Oster (2019).

complement the qualitative inferences provided by our tests. Fifth, we link our tests to the classical literature on selection models. Sixth, we relate our tests to ongoing discussions about instrumental variables methods and the trade-offs involved between approaches that build on instruments and those that build on conditional independence. Seventh, we undertake three meaningful empirical applications, one each for standard instrumental variables, fuzzy regression discontinuity, and experiments with imperfect compliance.

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, Ichimura, Smith, and Todd (1998) find that most of the difference between non-experimental 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 little evidence that their matching estimates suffer from selection bias when estimating their single treatment model using the very rich National Child Development Survey data. Similarly, when using comparable outcome measures for the treated and untreated units, Diaz and Handa (2006)'s propensity score matching estimates are quite similar to 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 reduces bias in many empirical contexts. 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. We see a parallel to non-parametric estimation, where researchers trade off bias and variance via the choice of a bandwidth. We develop a simple estimator of the magnitude of the bias implicit

in the OLS estimate which allows a quantitative comparison of the costs and benefits of IV relative to OLS.

In the next section, we outline the restrictions necessary for matching and OLS. In section 3, we outline the necessary assumptions for Imbens and Angrist's (1994) IV estimation and the latent index approach of Vytlacil (2002). In section 4, we outline our simple tests of conditional independence and provide formulae for the magnitude of the selection. Section 5 reports on our empirical applications. In section 6 we relate out tests to the broader literatures on instrumental variables and on the choice between IV and OLS estimation of the linear model. The final section concludes.

2 Matching, OLS, and Selection on Observed Variables

Using the notation introduced above, we define the causal impact of the treatment on agent i as $\delta_i = Y_{1i} - Y_{0i}$. The fundamental problem of evaluation is that we observe only one of the two potential outcomes; we must estimate the other, which the literature calls 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_i denote those covariates. Formally, matching estimators rely on two assumptions:

First, and most vexing, matching estimators require a Conditional Independence Assumption (CIA). Various "flavors" of CIA correspond to different parameters of interest. The strongest flavor demands that:

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

where $\perp\!\!\!\perp$ denotes statistical independence. This version of the CIA applies to the Average

Treatment Effect (ATE), or $\Delta^{ATE} = E(Y_{1i} - Y_{0i})$. The CIA for Y_{0i} assumes

$$Y_{0i} \perp\!\!\!\perp D_i | X_i. \tag{CIA^0}$$

Because researchers observe Y_{1i} for those who are treated, estimation of the $\Delta^{ATET} = E(Y_{1i} - Y_{0i}|D_i = 1)$ only requires the weaker (CIA^0) rather than the (CIA). Similarly, when estimating the Average Treatment Effect on the Non-treated (ATEN), researchers need only assume

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

which allows the estimation of $\Delta^{ATEN} = E(Y_{1i} - Y_{0i}|D_i = 0)$. Of course, the (CIA) implies that both (CIA^1) and (CIA^0) hold.

Technically, the ATE, ATET or ATEN do not require full conditional independence for identification. Rather, they "only" require conditional mean independence, which comes in three parallel flavors:

$$E(Y_{ji}|X_i, D_i = 1) = E(Y_{ji}|X_i, D_i = 0) \; j \in \{0, 1\} \tag{CMIA}$$

$$E(Y_{0i}|X_i, D_i = 1) = E(Y_{0i}|X_i, D_i = 0) \tag{CMIA^0}$$

$$E(Y_{1i}|X_i, D_i = 1) = E(Y_{1i}|X_i, D_i = 0). \tag{CMIA^1}$$

We mainly consider tests of (CMIA^0) and (CMIA^1), but also develop and implement tests of (CIA^0) and (CIA^1).

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

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

The (CSA) requires an untreated comparison unit with approximately the same realization of the covariates as each treated unit. Unlike the (CIA), researchers may test the (CSA)

in observational data. 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.[7]

When applying semi-parametric or non-parametric matching methods, researchers commonly specify the potential outcome functions as

$$Y_{1i} = g_1(X_i) + \epsilon_{1i} \tag{4}$$

$$Y_{0i} = g_0(X_i) + \epsilon_{0i} \tag{5}$$

where $\big(g_0(\cdot), g_1(\cdot)\big)$ denote the unknown conditional mean functions and $(\epsilon_{0i}, \epsilon_{1i})$ summarize the residual uncertainty associated with unobserved variables. With the (CSA) and the appropriate version of the (CIA) or (CMIA), researchers may use a variety of methods to estimate the unknown conditional mean functions. A common alternative to matching uses OLS to estimate parametric linear models. In this approach, researchers specify the functional form of the conditional mean function by replacing $g_1(X_i)$ in (4) with $X_i'\beta_1$ and replacing $g_0(X_i)$ in (5) with $X_i'\beta_0$, or by pooling the data and estimating a common β. With these substitutions, the researcher avoids invoking the (CSA) but not the (CMIA).[8]

Estimates identified via conditional independence often get criticized because such assumptions appear implausible in many substantive contexts given the available conditioning variables. To avoid a CIA, applied researchers often turn to IV estimation. While traditional IV assumptions presume a common treatment effect, 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 with binary instruments. It is difficult to overemphasize the importance of this advance. Models

[7] See the discussions in Black and Smith (2004) and Crump et al. (2009).

[8] Other estimators that build on conditional independence include Inverse Propensity Weighting (IPW), so-called "double robust" estimators that combine IPW with a linear model, and various estimators that exploit machine learning tools. See e.g. Heckman et al. (1999), Imbens (2004), Smith and Todd (2005), Huber et al. (2013), and Busso et al. (2014).

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, Urzua, and Vytlacil (2006) term such heterogeneous impacts "essential heterogeneity."

3 The IV and Control Function Approaches

We may state the assumptions of the LATE estimator as the Existence of Instruments (EI) and Monotonicity (M). To keep the notation simple, we continue to assume $Z_i \in \{0, 1\}$; our arguments, however, generalize to continuous instruments. Formally,

$$\begin{aligned} &\text{(i) } (Y_{0i}, Y_{1i}, D_i(0), D_i(1)) \perp\!\!\!\perp Z_i | X_i \\ &\text{(ii) } Pr(D_i = 1 | X_i, Z_i) \text{ is a non-trivial function of } Z_i \end{aligned} \tag{EI}$$

$$\text{Either } D_i(0) \geq D_i(1) \forall i \text{ or } D_i(0) \leq D_i(1) \forall i. \tag{M}$$

The (EI) assumption requires that, conditional on X_i, the instrument only affects the outcome through its effect on treatment $D_i(Z_i)$. The (M) assumption requires that all agents respond to the instrument in the same direction, not that the function $Pr(D_i = 1|X_i, Z_i)$ be monotone in Z_i; this led Heckman et al. (2006) to rename the condition uniformity, although the somewhat confusing term monotonicity was too well-established to be displaced. The (M) assumption is restrictive; it need not hold in all substantive contexts. 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.

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. In our notation, one may define the expectations of the errors in equations (4) and (5) as zero, or $E(\epsilon_{1i}|X_i) = E(\epsilon_{0i}|X_i) = 0$. This is, of course, a convenient normalization with any nonzero mean absorbed into the conditional mean functions. When we observe only one of the two potential outcomes, we do not know that the conditional expectations $E(\epsilon_{1i}|D_i = 1)$ and $E(\epsilon_{0i}|D_i = 0)$ equal zero because of the missing data problem. To formalize this argument a bit, we follow Vytlacil (2002) and let

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

$$Z_i \perp\!\!\!\perp (Y_{1i}, Y_{0i}, U_i)|X_i \tag{V2}$$

where 1(.) is the logical indicator function, U_i is an unobserved random variable that affects treatment choice, and $h(Z_i, X_i)$ is the index function. Selection on unobservables means that (Y_{1i}, Y_{0i}) are not independent of U_i conditional on X_i.

With assumptions (EI) and (M) (or the equivalent assumptions (V1) and (V2) for latent index models), we may write

$$E(\epsilon_{1i}|X_i, Z_i, D_i = 1) = c_1(X_i, Pr(X_i, Z_i)) \tag{6}$$

$$E(\epsilon_{0i}|X_i, Z_i, D_i = 0) = c_0(X_i, Pr(X_i, Z_i)) \tag{7}$$

where $Pr(Z_i, X_i) = Pr(D_i = 1|X_i, Z_i)$ is the conditional probability of treatment or propensity score[9] and where we denote the control functions that embody the conditional means of ϵ_{1i} and ϵ_{0i} by $c_1(\cdot)$ and $c_0(\cdot)$. We can obtain $E(\epsilon_{1i}|D_i = 1)$ and $E(\epsilon_{0i}|D_i = 0)$ by integrating out over X_i and Z_i in (6) and (7). The key to (6) and (7) is that Z_i enters only through the probability of treatment. In the absence of selection the control function equals zero everywhere; in the presence of selection, including the control function solves the selection

[9]Unlike the propensity score used in matching estimators under the (CIA), this propensity score also includes at least one instrument; see Heckman and Navarro-Lozano (2004) for further discussion.

problem.

The control function approach allows an easier interpretation of the independence assumption embedded in (EI); it 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. We turn now to our tests, which replace the control function with a simple function of the instrument.

4 Testing Conditional Independence

4.1 Instrumental Variables

As noted above, the various versions of the (CIA) and (CMIA) allow researchers to ignore the possibility of selection on unobserved variables, although they typically invoke them without looking for evidence of such selection. In contrast, the VIA assumptions allow researchers to consistently estimate LATEs for compliers. In the case of a single instrument we augment the linear parametric versions of equations (4) and (5) to obtain

$$E(Y_{1i}|X_i, Z_i) = X_i'\beta_1 + \alpha_1 Z_i \tag{8}$$

$$E(Y_{0i}|X_i, Z_i) = X_i'\beta_0 + \alpha_0 Z_i \tag{9}$$

In equations (8) and (9), $\alpha_1 Z_i$ and $\alpha_0 Z_i$ serve as simple linear approximations to the control functions in equations (6) and (7). Furthermore, $\alpha_0 = 0$ and $\alpha_1 = 0$ would be satisfied under conditional independence.[10] As such, we test $H_0 : \alpha_j = 0$ vs. $H_1 : \alpha_j \neq 0$.[11] Under

[10] Obtaining the exact functional form of the control function requires additional assumptions regarding the joint distribution of the unobserved components, such as the bivariate normality that underlies the classical two-step estimator of Heckman (1979). See Aradillas-Lopez, Honoré and Powell (2007) and Newey (2009) for approaches that avoid the specification of the functional forms of the control functions.

[11] In principle, one could add interactions between X and Z to (8) and (9) but we expect that doing so would provide few insights in most applications due to loss of statistical power. An alternative approach would estimate the equations completely separately for a small number of subgroups defined by X.

the VIA assumptions, rejecting this null can be taken as evidence for selection and against (CMIA^j).[12]

With non-binary instruments researchers may wish to add higher order terms (i.e., 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 could replace Z_i with the estimated propensity score, $\hat{Pr}(Z_i, X_i)$ and adjust the standard errors for generated regressors as in Murphy and Topel (2002).[13]

The model behind equations (8) and (9) represents an important departure from the canonical model used in IV applications, given by

$$Y_i = X_i'\beta + \phi D_i + \epsilon_i \tag{10}$$

In contrast, the model underlying equations (8) and (9) consists of

$$Y_{1i} = X_i'\beta_1 + \epsilon_{1i} \tag{11}$$

$$Y_{0i} = X_i'\beta_0 + \epsilon_{0i}. \tag{12}$$

Equations (11) and (12) allow estimation of heterogeneous treatment effects, $\Delta(X_i) = (\beta_1 - \beta_0)X_i$, that differ with the realization of the covariates while still maintaining the (CMIA). There is generally no theoretical reason to prefer equation (10) to equations (11) and (12), but the demands on instrument strength usually dissuade researchers from using the model described by equations (11) and (12)—or models embodying even more flexible forms of treatment effect heterogeneity—when they resort to IV estimation. Avoiding any interactions involving the treatment indicator represents a strong, substantive restriction on the model.

In the case of matching estimators, we augment equations (4) and (5) and specify the

[12]There are knife-edge cases where there is selection into treatment, i.e. equations (2) and (3) do not hold and $\alpha_j \neq 0$, but the CMIA is not violated. These cases depend on a specific distribution of the instrument which allows different forms of selection to "cancel out". Appendix A considers these pathological cases.

[13]Joo and LaLonde (2014) present a control function version of our test along these lines.

conditional mean functions as

$$E(Y_{1i}|X_i, Z_i, D_i = 1) = g_1(X_i) + \alpha_1 Z_i \tag{13}$$

$$E(Y_{0i}|X_i, Z_i, D_i = 0) = g_0(X_i) + \alpha_0 Z_i. \tag{14}$$

To clarify the relationship among the various forms of the (CMIA) and our test, it is useful to outline the samples used and hypotheses involved when estimating these auxiliary regressions. Formally, we estimate (9) or (14) using the sample of untreated observations to test the null that the (CMIA^0) holds, corresponding to $\alpha_0 = 0$, against the alternative that it does not hold, or $\alpha_0 \neq 0$. Similarly, we estimate equation (8) or (13) using the sample of treated observations to test the null that (CMIA^1) holds, corresponding to $\alpha_1 = 0$, against the alternative that it does not hold, or $\alpha_1 \neq 0$.

To develop some intuition for the tests, assume that $D_i(1) \geq D_i(0)$ and divide agents under VIA into the three types defined in the introduction: always-takers, never-takers, and compliers. The test given in either equation (8) or equation (13) simply compares $E(Y_{1i}|X, A)$ to $E(Y_{1i}|X, C)$. As Black et al. (2015) note, this is easily done because $E(Y_{1i}|X_i, Z_i = 0, D_i = 1) = E(Y_{1i}|X_i, A)$ and $E(Y_{1i}|X_i, Z_i = 1, D_i = 1) = E(Y_{1i}|X_i, A \cup C)$. Thus, at $X_i = x$ we have that

$$\alpha_1(x) \equiv \frac{Pr(C|x)}{Pr(C|x) + Pr(A|x)}[E(Y_{1i}|x, C) - E(Y_{1i}|x, A)]. \tag{15}$$

The regression coefficient in either equation (8) or equation (13) then simply integrates over the realizations of X_i, or $\alpha_1 = \int \alpha_1(x) w_1(x) dF(x)$, where $F(x)$ is the distribution of the covariates X_i and $w_1(x)$ is a weighting function that depends on the joint distribution of X and Z. Put differently, the tests look for evidence of a non-constant control function in equation (6), which constitutes evidence that unobserved variables affect the outcomes. A parallel argument applies to α_0. See Sloczyński (2020) for more on these weighting functions.[14]

[14]One can imagine attempting to strategically select the weights so as to maximize the power of the test.

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 (CMIA), and thus of the (CIA) as well.[15]

The tests allow researchers to assess whether any selection arises on Y_{0i}, which represents a violation of (CMIA^0), or on Y_{1i}, which represents a violation of (CMIA^1), or both. In addition, as with the tests for selection in the bivariate and trivariate normal models, our tests allow researchers to determine the signs of the relevant selection effects and their magnitudes, which we discuss in section 4.4. This allows researchers to provide a much more nuanced discussion of the nature of agents' underlying 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.

4.2 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 occurs at $S_i = 0$. To use both of our tests we require fuzziness on both sides of the discontinuity; formally, we require

$$1 > \lim_{S_i \downarrow 0} Pr(D_i = 1|X_i, S_i) > \lim_{S_i \uparrow 0} Pr(D_i = 1|X_i, S_i) > 0$$

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_{1i}, Y_{0i}) but not both. As emphasized by Imbens and Lemieux (2008) and Lee and Lemieux (2010),

We leave this bit of technocracy to future work.

[15] We find the separate tests of greater substantive interest than a joint test. Readers who disagree can follow the advice of de Luna and Johansson (2014) and combine them, or adopt the test in Angrist (2004).

when faced with a fuzzy RD, researchers who use the discontinuity at $S_i = 0$ as an instrument for treatment estimate a LATE at $S_i = 0$.

Because of the discrete change in the treatment probability at $S_i = 0$, 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 $E(Y_{0i}|X_i, S_i, D_i = 0)$ as S_i crosses zero, while the (CMIA^0) assumption implies a smooth $E(Y_{0i}|X_i, S_i, D_i = 0)$ function around zero. This suggests a simple test based on a model of the form

$$E(Y_{0i}|X_i, S_i, D_i = 0) = g_0(X_i, S_i) + \alpha_0 1(S_i \geq 0) \tag{16}$$

with $H_0 : \alpha_0 = 0$, or the corresponding version for testing (CMIA^1)

$$E(Y_{1i}|X_i, S_i, D_i = 1) = g_1(X_i, S_i) + \alpha_1 1(S_i \geq 0) \tag{17}$$

with $H_0 : \alpha_1 = 0$. Estimation of the sample analogue of (16) makes use only of untreated observations, which constitute a mixture of compliers and never-takers. Similarly, estimation of the sample analogue of (17) uses only treated observations, which constitute a mixture of compliers and always-takers.

As noted in the introduction, Bertanha and Imbens (2020) 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.

4.3 Experiments with Imperfect Compliance

As Heckman (1996) emphasizes, random assignment creates an instrument for treatment. As documented in Heckman, Hohmann, Smith, and Khoo (2000), many social experiments have imperfect compliance, with both treatment group dropout and control group substitution

into similar treatments provided elsewhere. In the presence of 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. One could easily implement our tests to check for selection on Y_{1i} or Y_{0i} in experiments such as these. As noted above, Huber (2013) provides a test closely related to ours.

4.4 Recovering Estimates of the Magnitude of the Selection Effect

To recover estimates of the magnitude of the selection effect, continue to assume that $Z_i = 1$ encourages treatment, and ignore covariates for notational simplicity. Rearranging equation (15) yields a measure of the selection effect for Y_{1i}, which we denote B_{1i}, given by

$$B_{1i} = E(Y_{1i}|C) - E(Y_{1i}|A) = \frac{Pr(C) + Pr(A)}{Pr(C)}\alpha_1. \tag{18}$$

An eerily similar derivation yields the corresponding measure for the untreated outcome

$$B_{0i} = E(Y_{0i}|N) - E(Y_{0i}|C) = \frac{Pr(C) + Pr(N)}{Pr(C)}\alpha_0. \tag{19}$$

To implement these measures empirically, we may use the OLS estimates of (α_0, α_1). We know that $Pr(A) = Pr(D_i = 1|Z_i = 0)$, $Pr(N) = Pr(D_i = 0|Z_i = 1)$, and $Pr(C) = 1 - Pr(N) - Pr(A)$ under assumption (M), so we have sample analogues of all the terms on the right-hand sides of equations (18) and (19).

4.5 Full Independence

Up to this point we have concentrated on testing (CMIA^0) and (CMIA^1). As noted above, the most common causal estimands require only one or both of these conditional mean inde-

pendence assumptions. In this subsection we develop additional tests of other implications of full conditional independence, i.e., tests of other implications of (CIA^0) and (CIA^1).[16] We have two motivations: first, some researchers lust after causal estimands that require full conditional independence; second, even researchers interested in the ATET, ATEN or ATE may value the additional information provided by these tests, especially in cases where our tests of conditional mean independence only marginally fail to reject the null. We continue to emphasize ease of implementation in the tests we propose.

One category of tests compares higher moments rather than means. For example, in the IV case with a binary instrument considered in Section 3.1, we can test the nulls

$$F(Y_{0i}|N, X_i) = F(Y_{0i}|C, X_i) \tag{20}$$

$$F(Y_{1i}|C, X_i) = F(Y_{1i}|A, X_i) \tag{21}$$

where $F(.)$ denotes a cumulative distribution function, by estimating versions of (8) and (9) or of (13) and(14) with higher moments of Y_{1i} and Y_{0i} as the dependent variables.

A second category of test builds on non-linear transformations of continuous outcomes. For example, a very simple version of this strategy replaces the levels of Y_{0i} and Y_{1i} with their natural logarithms in the tests already described. The variant we pursue builds on distribution regressions of the form

$$E(1(Y_{1i} \leq y_q|X_i, Z_i, D_i = 1)) = g_1^q(X_i) + \alpha_1^q Z_i \tag{22}$$

$$E(1(Y_{0i} \leq y_q|X_i, Z_i, D_i = 0)) = g_0^q(X_i) + \alpha_0^q Z_i \tag{23}$$

where $q = F_y(y_q)$ denotes some quantile of the unconditional distribution of Y_i. Under (CIA^1) the null implies $\alpha_1^q = 0$ while under (CIA^0) it implies $\alpha_0^q = 0$, in both cases for all $q \in (0, 1)$.

We implement both the higher moment tests and the distribution regression tests in our

[16] Of course, for binary outcomes, conditional mean independence implies full independence.

applications in Sections 5.1 and 5.2 where we have continuous outcomes. For the higher moment tests, we consider second, third and fourth moments about the mean.[17] For the distribution regressions, we look at either deciles or ventiles, taking care with the mass point at zero for the earnings outcome in Section 5.1.[18,19]

5 Empirical Applications

5.1 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 endogenous with respect to women's labor supply decisions, Angrist and Evans devise an ingenious 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.

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

[17] For numerical reasons, we divide the outcome by its standard deviation when calculating the third and fourth moments. Stopping at the fourth moment is arbitrary. We leave the derivation of a data-driven method for ex ante choice of the optimal number of moments to future work.

[18] Using deciles or ventiles rather than, say, percentiles, lessens concerns about non-independence of adjacent tests at some cost in statistical power. Both approaches to testing full independence raise multiple testing issues (as does our consideration of multiple outcomes in our tests of conditional mean independence). We view these issues as beyond our scope; Schochet (2008) and List, Shaikh, and Xu (2019) provide valuable overviews of the multiple testing literature aimed at applied readers.

[19] An alternative path would either try to combine our moment tests or our distribution regression tests to produce a single test statistic, or would replace them with a different joint test, such as one of the tests based on characteristic functions in Wang and Hong (2018).

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

$$Y_i = X_i'\beta + \phi M_i + \epsilon_i \tag{24}$$

$$M_i = X_i'b + \gamma Z_i + \eta_i \tag{25}$$

where M_i is an indicator for having more than two children. 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 non-black and non-white (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, Z_i, is an indicator for whether the first two children were either two boys or two girls. For dependent variables, Y_i, 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

$$Y_i = g(X_i) + \phi M_i + v_i \tag{26}$$

$$M_i = h(X_i) + \gamma Z_i + u_i \tag{27}$$

where, in general, $g(.)$ and $h(.)$ comprise fully saturated functions of our discrete conditioning variables. Our parametric results, both the OLS and two-stage least squares estimates, exactly match Angrist and Evans' findings.[20] Moreover, the semi-parametric estimates turn out virtually identical to the parametric estimates of Angrist and Evans, which is not too

[20]Keeping in mind that we rescaled the income variable to be in $1000s.

surprising given that nature makes the sex of women's offspring 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 document 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.[21] With these (strong) assumptions, we may now implement our parametric tests of the CIAs using

$$Y_{0i} = X_i'\beta_0 + \alpha_0 Z_i + \epsilon_{0i} \tag{28}$$

$$Y_{1i} = X_i'\beta_1 + \alpha_1 Z_i + \epsilon_{1i} \tag{29}$$

and our semi-parametric tests by repeating the estimation replacing $X_i'\beta_0$ with $g_0(X_i)$ in (28) and $X_i'\beta_1$ with $g_1(X_i)$ in (29). For our semiparametric analysis we drop the indicator for having a boy as the second child in order to avoid making the Z_i variable perfectly collinear. We estimate equation (28) on the sample of 236,092 women who have exactly two children, and equation (29) using the sample of 158,743 who have three or more children.

Table 2 presents the results of our tests of conditional mean independence. The data strongly reject the null that $\alpha_0 = 0$, thereby providing strong evidence against the (CMIA^0). For each of the four outcomes, we reject the null hypothesis of $\alpha_j = 0$ at a five-percent confidence level. In each case, we estimate a positive coefficient α_0 on Z_i, where $Z_i = 1$ 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 (CMIA^1). Unlike the estimates of α_0,

[21] Some subsequent authors have questioned the monotonicity assumption in this context; see e.g. de Chaisemartin (2017).

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_{1i}. Frankly, we find this result stunning. There appear to be no meaningful differences in the labor supply behavior of the always-takers and the compliers when we condition on the limited observed characteristics available in the US Census data. Before undertaking this analysis, we fully expected to find a two-sided selection problem. The data disagreed.

Table 3 adds the results from our tests of independence of distributions for the three continuous outcomes. The upper panel shows the results from estimating equations (22) and (23) for the mass point at zero and for each decile above the mass point. Both sets of distributional tests strongly corroborate the lessons from our tests of conditional mean independence: substantively large and statistically significant rejections of the null for Y_{0i} and substantively small and rather imprecise failures to reject the null for Y_{1i}. The lower panel shows the results from tests based on (20) and (21) with the 2^{nd}, 3^{rd}, and 4^{th} (central) moments replacing the mean. For completeness, we repeat the results for the 1^{st} moment (i.e., the mean) from Table 2. The tests based on higher moments shed much less light, as only two of 18 tests reject their null, one for Y_{0i} with the hours worked outcome and one for Y_{1i} with the income outcome.

To describe the magnitude of the selection effects we use the semi-parametric estimates from 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 while differing only very little from the always-takers.

5.2 Angrist and Lavy (1999) Data

Angist and Lavy (1999) address the perennial economics of education topic of the causal effect of class size on student outcomes using data from Israel. They exploit the use of Maimonides' rule that limits class sizes to 40 in Israeli public schools to construct regression discontinuity estimates of class size effects. Under Maimonides' rule enrollment in a particular combination of school, grade and academic year constitutes the running variable. When the running variable crosses from 40 to 41, average class size drops from 40 to 20.5. Similar things happen when the running variable moves from 80 to 81 or from 120 to 121. In practice, for a variety of institutional reasons (such as extra funds provided to schools serving relatively disadvantaged students and the fact that enrollment is measured in the fall of the academic year while class size is measured in the spring) the data present a fuzzy regression discontinuity rather than the strict one that would arise if Maimonide's rule alone determined class size.

For their primary estimates Angrist and Lavy (1999) operationalize class size as a continuous treatment, and do two-stage least squares with the first stage instrument consisting of the predicted enrollment based solely on Maimonide's rule. In keeping with our focus on the discontinuity framing of their analysis, we replicate the estimates in columns (5) and (11) of their Table V. These estimates refer to fourth graders in 1991 and use only observations on school-grade combinations with enrollments within five students (above or below) of a multiple of 40.[22] The outcomes consist of reading and math achievement tests scores in "natural" units rather than standard deviation units. Our replicate estimates appear in the top row of Table 4. Except for the third digit in one of the standard errors, we match their numbers exactly. In a common effect world, we can interpret the -0.098 estimate for reading comprehension as indicating that a one student increase in class size decreases the number of correct answers by about 0.1 in a context where, following their Table I, the mean math score equals 28.1 with a standard deviation of 6.5.[23] In a heterogeneous treatment effects

[22]Their paper provides a startling reminder that the now-ubiquitous applied econometric toolkit for regression discontinuity designs, including sophisticated bandwidth selection tools, did not yet exist in 1999.

[23]The value of 6.5 refers to the standard deviation of class-level mean math scores. The individual-level

world, the two-stage least squares estimator that Angrist and Lavy (1999) employ implicitly yields a weighted average of the LATEs at the multiple discontinuities.

To make their setup comparable to our other examples, we discretize both their continuous treatment and their continuous instrument. In particular, our binary treatment consists of having a class size of less than 35, while our binary instrument consists of having a predicted class size of less than 35. The bottom row of Table 4 shows the estimates we obtain using the same sample as in the top row, but with the binary treatment and binary instrument.[24] To interpret these estimates, first note that the mean change in class size for the compliers equals 3.18 students. Thus, under a LATE interpretation, a change in class size of that magnitude decreases expected math test scores by -1.34 and increases expected reading comprehension test scores by 1.37 within that group, although neither of these estimates approach statistical significance.

Table 5 presents the results from our tests of (CMIA^0) and (CMIA^1) using the Angrist and Lavy (1999) data. We find little evidence of selection in either case but note that, consistent with the standard errors on our estimates in Table 4, we do not have nearly as much power as we would like. The tests of (CIA^0) and (CIA^1) in Table 6 tell a modestly different story. In particular, our distributional regressions provide meaningful but not overwhelming evidence against (CIA^0) for the reading comprehension test and against (CIA^1) for the math test in the form of multiple (two out of four ventiles) rejections of the null in each case. As we would expect a large common component underlying test performance in the two subjects, these findings make us hesitant to extrapolate from the LATE interpretation to an ATET or ATEN interpretation in either subject; put differently, background knowledge leads us to think that the rejections should "spill over" across subjects.

variation in math scores is surely much larger. Thus, the estimated causal effect, expressed in terms of the common metric of standard deviations of individual scores, is rather small. See their Section V.B for more about magnitudes.

[24] Angrist and Lavy (1999) undertakes a similar exercise, see their Table VI. Ours differs in that we do not obtain separate estimates for each of the three discontinuities, preferring instead a more precisely estimated weighted average, at some cost in ease of interpretation.

5.3 Eberwein, Ham, and LaLonde (1997) Data

When facing control group substitution and treatment group dropout (i.e., two-sided non-compliance) in an experiment, researchers will often estimate two treatment parameters: the intent-to-treat parameter, estimated as the mean 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.

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.[25] 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.[26] The administrative data from the experiment provide our treatment indicator. For our outcome variable, we use an indicator for self-reported employment in the eighteenth month after random assignment.[27]

The data reveal much non-compliance. On the dropout side, only about 65 percent of the treatment group reports receiving classroom training in the first nine months after random assignment. On the substitution side, about 34 percent of the control group reports receiving classroom training over the same period.

[25]The reader should not compare our results directly to those in Donald et al. (2014) as their data on women also includes those recommended for services other than classroom training.

[26]Smith and Whalley (2021) offer a depressing exploration of the concordance of administrative and self-reported measures of service receipt in the JTPA study data.

[27]As noted in Section 4.5, with a binary outcome the conditional mean fully characterizes the distribution, leaving no scope for additional tests of conditional independence of distributions.

In Table 7, we provide two sets of estimates of the intent-to-treat parameter. In column (1) we provide the simple mean difference estimates given by

$$Y_i = \beta + \phi R_i + \epsilon_i, \tag{30}$$

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, ϕ) are parameters to be estimated. The estimated intent-to-treat parameter, $\hat{\phi}$, 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 achieves statistical significance 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.

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

$$Y_{0i} = \beta_0 + \alpha_0 R_i + e_{0i} \tag{31}$$

$$Y_{1i} = \beta_1 + \alpha_1 R_i + e_{1i} \tag{32}$$

where (Y_{0i}, Y_{1i}) denote the outcomes of those not receiving training and receiving training. We estimate Equation (31) using the 1,233 adult women who do not receive training and estimate equation (32) using the 1,501 who do receive training. We find little evidence that the compliers have different Y_{0i} than those who never take training. The coefficient on the indicator for assignment to the treatment group in equation (31) 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 (32) 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 in the treated state.

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 7 for a complete list. 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_{1i} 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 treatment group indicator 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 Discussion

6.1 Finite Sample Performance

Appendix B presents a modest but informative Monte Carlo analysis focused on the size and power of our test. Our simulated data come from a relatively generic generalized Roy model, varying the sample size, the strength of the first stage, and the extent of selection on the unobserved component of the outcome. We obtain the expected patterns. For example,

because weak instruments imply a relatively small number of compliers, they imply low power for our test, as indeed we find. More generally, a reasonable sample size, a strong first stage, and a lot of selection on the unobserved component of the outcome imply power close to one. The close-to-zero cost of our test makes it worth doing even in contexts that lack one or two of these features; in such contexts, the researcher will wisely pay more attention to rejections of the null than to failures to reject.

6.2 IV Assumptions and Our Test

What are the implications of faiure of one or both of the (EI) and (M) assumptions for our test? 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 AIR (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.[28] Similarly, the treated units now comprise always-takers, compliers and defiers. The presence of defiers undoes the LATE interpretation of the IV estimand. Klein (2010) provides a gateway into the literature that relaxes the monotonicity assumption.

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 converge to an unknown mixture of the population LATE and the bias associated with the invalid instrument.

[28]Dahl, Huber, and Mellace (2017) provides a convenient port of entry to the literature on defiers.

6.3 Bias, Variance, and Mean-squared Error

Statistics reminds us of the trade-off between bias and variance. Consider the canonical model in (10) and imagine for the moment either a common effect world or a world where the (CMIA) holds. In that world, as Cameron and Trivedi (2005, p. 107) note, with one instrument we may compare the variance of the parameter of interest in (10) estimated using instrumental variables to the variance of the same parameter estimated using OLS using the equation:

$$SE(\hat{\delta}^{IV}) = \frac{SE(\hat{\delta}^{OLS})}{\hat{\rho}(\tilde{D}, \tilde{Z})} \tag{33}$$

where $\hat{\rho}(\tilde{D}, \tilde{Z})$ denotes the partial correlation coefficient after removing the variation associated with the other covariates X, what Black and Smith (2006) term the "Yulized residuals" in honor of Yule's (1907) brilliant paper.

Many empirical contexts yield quite modest partial correlations, 0.10 or even 0.05 depending on the strength of the instrument, implying dramatic decreases in the precision of the estimates when using IV methods. For example, Black et al. (2015) estimate $\hat{\rho}(\tilde{D}, \tilde{Z}) = 0.059$. Others contexts yield larger values: for instance, the ratio of the standard errors for the OLS and IV estimates in Table 2 of Bertanha and Imbens (2020) implies a large partial correlation. In many situations, serious researchers will trade off increases in bias for reductions in variance. Our method provides researchers with a simple 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.

To see how nefarious the IV cure may be, recall the canonical common effects model given in (10). Let $\hat{\delta}^{OLS} = \delta + e^{OLS}$ and $\hat{\delta}^{IV} = \delta + e^{IV}$, where e^{OLS} and e^{IV} denote the errors in the two estimates. Researchers concerned only with the distance of their estimate from the true parameter value (i.e., with the absolute values of e^{OLS} and e^{IV}) should prefer the IV estimator only when

$$(|\hat{\rho}(\tilde{D}, \tilde{\epsilon})||\hat{\rho}(\tilde{D}, \tilde{Z})| - |\hat{\rho}(\tilde{Z}, \tilde{\epsilon})|) > 0, \tag{34}$$

where $\hat{\rho}(.,.)$ again denotes the sample correlation between its arguments.[29] Any correlation between the treatment indicator, D, and the error term ϵ due to omitted variables and sampling variation gets down-weighted by the relative strength of the instrument. As the estimated correlation $\hat{\rho}(\tilde{D}, \tilde{Z})$ equals 0.064 in our Angrist and Evans' application, the correlation $\hat{\rho}(\tilde{D}, \tilde{\epsilon})$ must be an order of magnitude larger than $\hat{\rho}(\tilde{Z}, \tilde{\epsilon})$ before we prefer the IV estimates.

We might turn to formal decision theory to determine which estimator to use. An oft-used metric in decision theory is mean squared error. The expression for

$$\hat{\delta}_{IV} = \frac{cov(\tilde{Y}, \tilde{Z})}{cov(\tilde{D}, \tilde{Z})} \tag{35}$$

illustrates a peculiar feature about IV estimation: Because of the $cov(\tilde{D}, \tilde{Z})$ term in the denominator of the estimator, the expectation of $\hat{\delta}_{IV}$ does not exist in the case with only one instrument.[30] In just-identified models, both the bias and the variance of the IV estimator are not defined.[31] In contrast, while the OLS estimator is biased and inconsistent, the bias and variance of the estimator are finite. In the terminology of decision theory, the IV estimator is inadmissible: we should always pick OLS estimation. In our view, this conclusion may say more about the use of mean squared error as a criterion than the choice of estimators.

6.4 Different Estimators, Different Estimands

In a heterogeneous treatment effects (i.e., realistic) context, the IV and OLS estimands correspond to different causal parameters. Among others, Deaton (2010), Heckman and Urzua (2010), and Imbens (2010) debate the value of these estimands at a relatively high level of generality. In our view, the relevance of the LATE parameter depends on the

[29] This is just the finite sample analogue of equation (7) in Bound, Jaeger and Baker (1995).

[30] This result dates back to the 1960s; see Nelson and Startz (1990) for an excellent discussion and some additional results.

[31] As noted by Nelson and Startz (1990), the failure of the existence of the first moment $\hat{\delta}_{IV}$ is the result of the estimator performing poorly when $\widehat{cov}(\tilde{D}, \tilde{Z}) \approx 0$, as the estimates fluctuate widely depending on the magnitude and sign of $\widehat{cov}(\tilde{Z}, \tilde{\epsilon})$.

substantive context and on the composition of the complier population. Our tests and the related measures of bias developed above provide a framework within which to think about generalizing from the LATE to the ATET without additional assumptions.[32]

Imbens and Rubin (1997) and Abadie (2003) demonstrate that one can recover estimates of the marginal distributions of Y_{1i} and Y_{0i} for the compliers and thus $E(Y_{1i}|C)$ and $E(Y_{0i}|C)$. In Table 8, we produce estimates of $E(Y_{0i}|N)$, $E(Y_{0i}|C)$, $E(Y_{0i}|N \cup C)$, $E(Y_{1i}|A)$, $E(Y_{1i}|C)$, and $E(Y_{1i}|A \cup C)$ for each of the four outcomes in the Angrist and Evans data. To facilitate comparisons, we use IPW to reweight the data so that each set (i.e., A, N, or C) has the same distribution of the covariates as the aggregated data.[33] Two features stand out: First, we find little substantive difference between the always-takers and compliers in the treated state, but, second, we find that the compliers and the never-takers have substantially larger differences in the untreated state, consistent with Table 2.

In the absence of evidence for selection on unobserved variables for ϵ_{1i}, OLS estimation of the regression

$$Y_{1i} = X_i'\beta_1 + \epsilon_{1i} \tag{39}$$

using the treated units produces unbiased and consistent estimates of β_1. We can then use the estimated $\hat{\beta}_1$ to predict a treated outcome for each untreated unit and then estimate $\hat{\Delta}^{ATEN}$ using the observed untreated outcomes and the predictions. But given the estimates in Table 8, this appears unwise. While we have no direct evidence against the hypothesis that $E(Y_{1i}|N, X_i) = E(Y_{1i}|A \cup C, X_i)$, we do have evidence that the never-takers and compliers differ in unobserved ways in the untreated state.

Given that we have evidence of the similarity of the always-takers and the compliers in the treated state, we may wish to use $E(Y_{0i}|X_i, C)$ to form the missing counterfactual for $E(Y_{0i}|X_i, A)$ and estimate Δ^{ATET} for the always-takers and compliers. This requires us to

[32] Brinch et al. (2017) and Kowalski (2018a,b) consider the value of additional assumptions in addressing this same issue.

[33] Doing so requires us to drop 3,104 observations (less than one percent of the data) to impose the common support condition.

assume that $E(Y_{0i}|X_i, C) = E(Y_{0i}|X_i, A)$ as well as maintaining the VIA assumptions (so we may use the test for selection bias). In contrast, the IV estimator simply maintains the VIA assumptions. But the estimand of the IV estimator in this case lacks intrinsic interest: It is the impact of having a third child for the set of women who would have stopped with two children had they had a boy and a girl for their first two children. In contrast, the Δ^{ATET}, even with the potential bias, retains its substantive interest.

This argument applies specifically to the Angrist-Evans application. In the JTPA experiment, the compliers behaved in accordance with the treatment protocol in the experiment. We care about their LATE because it tells us about the effect of classroom training on those induced by the program to sign up for it. This constitutes an economically interesting parameter, albeit one estimated with large standard errors in our data. Of course, interest in the LATE in a particular context does not imply lack of interest in Δ^{ATET} or Δ^{ATEN}.

6.5 Pre-testing and an Alternative

A researcher who follows the methodology outlined in the paper for conditional mean independence will end up with an IV estimate, an OLS estimate, two test statistics (with p-values), and two bias estimates, one each for selection on Y_{0i} and Y_{1i}. What should the researcher do with these statistics? One strategy, inspired by the way researchers used the Durbin-Wu-Hausman (DWH) test back in the 1970s and 1980s, would view each test as a pre-test and report only the estimates indicated by the test, interpreted in the appropriate way. Thus, the researcher would report the IV estimate, interpreted as a LATE, given evidence of selection on Y_{0i} or Y_{1i}, and would report an OLS (or other CIA-based) estimate, interpreted as an ATET (or even an ATE), in the absence of such evidence. This strategy has the virtue of simplicity, and performs better than one might expect in the Monte Carlo analyses in Donald et al. (2014).

We see two main reasons not to adopt this strategy. First, the estimated variances do

not incorporate the additional variance component induced by the pre-test.[34] Guggenberger (2010) shows that incorporating this component in the traditional common effect DWH world really matters. More importantly in our view, the traditional approach lacks nuance and hides valuable information from the reader. In many contexts, the test will reject or fail to reject only marginally; in many contexts, the data will imply only a modest bias, whether statistically significant or not. In such contexts, we think most readers will want to assign non-zero posterior weights to both estimates.

In addition, to aid in this process, the researcher might even formally combine the IV and OLS estimates in some reasonable way, along the lines suggested in Donald et al. (2014), who combine to minimize variance, or Hansen (2017), which combines based on the strength of the DWH test result. We think such combinations represent a promising avenue for future research.

7 Conclusion

In this paper, we derived simple tests for selection on unobserved variables when using instrumental variables. Our tests of conditional mean independence generalize various existing tests in the literature while our tests of independence of distributions go beyond what the existing literature provides. 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.

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 the context of the bivariate or trivariate normal model. In this paper, we showed that simple

[34] Another option conducts the pre-test using a random sub-sample of the data and the estimation using the remainder, as in Athey and Imbens (2016). This procedure removes any correlation between the pre-test and the estimates, at the cost of increasing the standard errors of the estimates due to a smaller sample size.

auxiliary regressions will produce rich insights into the nature and magnitude of the selection effect when using IV estimation. Our three empirical applications nicely demonstrate the knowledge our framework produces, as does its application in Azzam, Bates, and Fairris (2018).

8 Bibliography

Abadie, Alberto, 2003. "Semiparametric Instrumental Variable Estimation of Treatment Response Models" *Journal of Econometrics* 113 231-263.
https://doi.org/10.1016/S0304-4076(02)00201-4

Altonji, Joseph, Todd Elder, and Christopher Taber, 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools " *Journal of Political Economy* 113(1) 151-184.
https://doi.org/10.1086/426036

Angrist, Joshua, 2004. "Treatment Effects Heterogeneity in Theory and Practice" *Economic Journal* 114(494) C52-C83.
https://doi.org/10.1111/j.0013-0133.2003.00195.x

Angrist, Joshua and William Evans, 1998. "Children and Their Parent's Labor Supply: Evidence from Exogenous Variation in Family Size" *American Economic Review* 88(3) 450-77.
https://www.jstor.org/stable/116844

Angrist, Joshua, Guido Imbens, and Donald Rubin, 1996. "Identification of Causal Effects using Instrumental Variables" (with discussion) *Journal of the American Statistical Association* 91 444-72.
https://doi.org/10.2307/2291629

Angrist, Joshua, and Victor Lavy, 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement" *Quarterly Journal of Economics* 114(3) 533-575.
https://www.jstor.org/stable/2587016

Aradillas-Lopez, Bo E. Honor and James L. Powell, 2007."Pairwise Difference Estimation with Nonparametric Control Variables" *International Economic Review* 48(4) 1119-58.
https://doi.org/10.1111/j.1468-2354.2007.00457.x

Athey, Susan and Guido Imbens, 2016. "Recursive Partitioning for Heterogeneous Causal Effects" *Proceedings of the National Academy of Science* 113(27) 7353-7360.
https://doi.org/10.1073/pnas.1510489113

Azzam, Tarek, Michael Bates, and David Fairris, 2018. "Do Learning Communities Increase First Year College Retention? Testing the External Validity of Randomized Control Trials" Unpublished manuscript, University of California at Riverside.

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.
https://doi.org/10.1016/j.jeconom.2007.05.006

Bertanha, Marinho and Guido Imbens, 2020. "External Validity in Fuzzy Regression Discontinuity Designs" *Journal of Business and Economic Statistics* 38(3) 593-612.
https://doi.org/10.1080/07350015.2018.1546590

Björklund, Anders and Robert Moffitt, 1987. "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models" *Review of Economics and Statistics* 69(1) 42-49.
https://doi.org/10.2307/1937899

Black, Dan, Seth Sanders, Evan Taylor, and Lowell 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.
https://doi.org/10.1257/aer.20120642

Black, Dan and Jeffrey Smith, 2004. "How Robust is the Evidence on the Effects of College Quality? Evidence from Matching" *Journal of Econometrics* 121(1-2) 99-121.
https://doi.org/10.1016/j.jeconom.2003.10.006

Black, Dan and Jeffrey Smith. 2006. "Estimating the Returns to College Quality with Multiple Proxies for Quality" *Journal of Labor Economics* 24(3) 701-28.
https://doi.org/10.1086/505067

Bloom, Howard, 1984. "Accounting for No-Shows in Experimental Evaluation Designs" *Evaluation Review* 8(2) 225-46.
https://doi.org/10.1177%2F0193841X8400800205

Bloom, Howard, Larry Orr, Stephen Bell, George Cave, Fred Doolittle, Winston Lin, Johannes 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.
https://doi.org/10.2307/146183

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.
https://doi.org/10.1111/j.1467-985X.2004.00360.x

Bound, John, David Jaeger, and Regina Baker, 1995. "Problems with Instrumental Variables Estimation when the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak" *Journal of the American Statistical Association* 90(430) 443-450.
https://doi.org/10.2307/2291055

Brinch, Christian, Magne Mogstad, and Matthew Wiswall, 2017. "Beyond LATE with a Discrete Instrument." *Journal of Political Economy* 125(4) 985-1039.
https://doi.org/10.1086/692712

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.
https://doi.org/10.1162/REST_a_00431

Cameron, Colin and Pravin Trivedi, 2005. *Microeconometrics: Methods and Applications* Cambridge, UK: Cambridge University Press.
https://doi.org/10.1017/CBO9780511811241

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.
https://doi.org/10.1080/01621459.2013.795447

Crump, Richard, V. Joseph Hotz, Guido Imbens, and Oscar Mitnik, 2009. "Dealing with Limited Overlap in Estimation of Average Treatment Effects" *Biometrika* 96(1) 187-99.
https://doi.org/10.1093/biomet/asn055

Dahl, Christian, Martin Huber, and Giovanni Mellace, 2017. "It's Never Too LATE: A New Look at Local Average Treatment Effects With or Without Defiers." University of Southern Denmark Discussion Papers on Business and Economics No. 2/2017.

de Chaisemartin, Clment, 2017. "Tolerating Defiance? Local Average Treatment Effects without Monotonicity" *Quantitative Economics* 8(2) 367-396.
https://doi.org/10.3982/QE601

de Luna, Xavier and Per Johansson. 2014. "Testing for the Unconfoundedness Assumption Using an Instrumental Assumption." *Journal of Causal Inference* 2(2): 187-199.
https://doi.org/10.1515/jci-2013-0011

Deaton, Angus, 2010. "Instruments, Randomization, and Learning about Development" *Journal of Economic Literature* 48(2) 424-455.
https://doi.org/10.1257/jel.48.2.424

Diaz, Juan Jose and Sudhanshu Handa, 2006. "An Assessment of Propensity Score Matching

as a Nonexperimental Estimator" *Journal of Human Resources* 41(2) 319-45.
https://doi.org/10.3368/jhr.XLI.2.319

Donald, Stephen, Yu-Chin Hsu, and Robert Lieli, 2014. "Testing the Unconfoundedness Assumption via Inverse Probability Weighted Estimators of (L)ATT" *Journal of Business and Economic Statistics* 23(3) 395-415.
https://doi.org/10.1080/07350015.2014.888290

Eberwein, Curtis, John Ham, and Robert 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.
https://doi.org/10.2307/2971734

Guggenberger, Patrick, 2010. "The Impact of a Hausman Pretest on the Asymptotic Size of a Hypothesis Test" *Econometric Theory* 26(2) 369-382.
https://doi.org/10.1017/S0266466609100026

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.
https://doi.org/10.1002/sim.6227

Hansen, Bruce, 2017. "Stein-like 2SLS Estimator" *Econometric Reviews* 36(6-9) 840-852.
https://doi.org/10.1080/07474938.2017.1307579

Heckman, James, 1979. "Sample Selection Bias as a Specification Error" *Econometrica* 47(1) 206-248.
https://doi.org/10.2307/1912352

Heckman, James, 1996. "Randomization as an Instrumental Variable" *Review of Economics and Statistics* 78(2) 336-340.
https://doi.org/10.2307/2109936

Heckman, James, Neil Hohmann, Jeffrey 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.
https://doi.org/10.1162/003355300554764

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd, 1998. "Characterizing Selection Bias Using Experimental Data" *Econometrica* 66(5) 1017-98.
https://doi.org/10.2307/2999630

Heckman, James, Robert Lalonde and Jeffrey 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.
https://doi.org/10.1016/S1573-4463(99)03012-6

Heckman, James and Salvador Navarro-Lozano, 2004. "Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models" *Review of Economics and Statistics* 86(1) 30-57.
https://doi.org/10.1162/003465304323023660

Heckman, James, Daniel Schmierer, and Sergio Urzua, 2010. "Testing the Correlated Random Coefficients Model" *Journal of Econometrics* 158(2) 177-203.
https://doi.org/10.1016/j.jeconom.2010.01.005

Heckman, James and Sergio Urzua, 2010. "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify" *Journal of Econometrics* 156(1) 27-37.
https://doi.org/10.1016/j.jeconom.2009.09.006

Heckman, James, Sergio Urzua, and Edward Vytlacil, 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity" *Review of Economics and Statistics* 88(3) 389-432.
https://doi.org/10.1162/rest.88.3.389

Heckman, James and Edward Vytlacil, 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation" *Econometrica* 73(3) 669-738.
https://doi.org/10.1111/j.1468-0262.2005.00594.x

Heckman, James and Edward Vytlacil, 2007a. "Econometric Evaluation of Social Programs, Part1: Causal Models, Structural Models and Econometric Policy Evaluation" in *Handbook of Econometrics, Volume 6B*, eds. James Heckman and Edward Leamer, 4780-4873.
https://doi.org/10.1016/S1573-4412(07)06070-9

Heckman, James and Edward 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" in *Handbook of Econometrics, Volume 6B*, eds. James Heckman and Edward Leamer, 4875-5143.
https://doi.org/10.1016/S1573-4412(07)06071-0

Huber, Martin, 2013. "A Simple Test for the Ignorability of Non-compliance in Experiments" *Economic Letters* 120(3) 389-91.
https://doi.org/10.1016/j.econlet.2013.05.018

Huber, Martin, Michael Lechner, and Connie Wunsch, 2013. "The Performance of Estimators Based on the Propensity Score" *Journal of Econometrics* 175(1) 1-21.
https://doi.org/10.1016/j.jeconom.2012.11.006

Huber, Martin and Giovanni Mellace, 2015. "Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints" *Review of Economics and Statistics* 97(2) 398-411.

https://doi.org/10.1162/REST_a_00450

Imbens, Guido, 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review" *Review of Economics and Statistics* 86(1) 4-29.
https://doi.org/10.1162/003465304323023651

Imbens, Guido, 2010. "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)" *Journal of Economic Literature* 48(2) 399-423.
https://doi.org/10.1257/jel.48.2.399

Imbens, Guido and Joshua Angrist, 1994. "Identification and Estimation of Local Average Treatment Effects" *Econometrica* 62(2) 467-475.
https://doi.org/10.2307/2951620

Imbens, Guido, and Thomas Lemieux, 2008. "Regression Discontinuity Designs: A Guide to Practice" *Journal of Econometrics* 142(2) 615-635.
https://doi.org/10.1016/j.jeconom.2007.05.001

Imbens, Guido, and Donald Rubin, 1997. "Estimating Outcome Distributions for Compliers in Instrumental Variables Models" *Review of Economic Studies* 64 555-574.
https://doi.org/10.2307/2971731

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

Kitagawa, Toru, 2015. "A Test for Instrument Validity" *Econometrica* 83(5) 2043-2063.
https://doi.org/10.3982/ECTA11974

Klein, Tobias, 2010. "Heterogeneous treatment Effects: Instrumental Variables without Monotonicity?" *Journal of Econometrics* 155 99-116.
https://doi.org/10.1016/j.jeconom.2009.08.006

Kowalski, Amanda, 2018a. "How to Examine External Validity Within an Experiment." NBER Working Paper No. 24834.
https://doi.org/10.1111/jems.12468

Kowalski, Amanda, 2018b. "Reconciling Seemingly Contradictory Results from the Oregon Health Insurance Experiment and the Massachusetts Health Reform." NBER Working Paper No. 24647.

Lee, David and Thomas Lemieux, 2010. "Regression Discontinuity Designs in Economics" *Journal of Economic Literature* 48(2) 281-355.
https://doi.org/10.1257/jel.48.2.281

List, John, Azeem Shaikh, and Yang Xu, 2019. "Multiple Hypothesis Testing in Experimen-

tal Economics" *Experimental Economics* 22(4) 773-793.
https://doi.org/10.1007/s10683-018-09597-5

Manski, Charles, 1990. "Nonparametric Bounds on Treatment Effects" *American Economic Review* 80(2) 319-323.
https://www.jstor.org/stable/2006592

Murphy, Kevin and Robert Topel, 2002. "Estimation and Inference in Two-step Models" *Journal of Business and Economic Statistics* 20(1) 88-97.
https://doi.org/10.2307/1391724

Nelson, Charles and Richard Startz, 1990. "Some Further Results on the Exact Small Sample Properties of the Instrumental Variables Estimator" *Econometrica* 58(4) 967-976.
https://doi.org/10.2307/2938359

Newey, Whitney, James Powell, and James Walker. 1990. "Semiparametric Estimation of Selection Models: Some Empirical Results" *American Economic Review* 80(2): 324-328.
https://www.jstor.org/stable/2006593

Newey, Whitney K., 2009. "Two-step Series Estimation of Sample Selection Models" *Econometrics Journal* 12, s217-29.
https://doi.org/10.1111/j.1368-423X.2008.00263.x

Oster, Emily, 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence" *Journal of Business and Economic Statistics* 37(2) 187-204.
https://doi.org/10.1080/07350015.2016.1227711

Poincair, Nenri. 1913. *The Foundation of Science* Translated by George Halsted. New York: The Science Press.

Romano, Joseph, and Azeem Shaikh, 2006. "Stepup Procedures for Control of Generalizations of the Familywise Error Rate" *Annals of Statistics* 34(4) 1850-73.
https://doi.org/10.1214/009053606000000461

Romano, Joseph, and Michael Wolf, 2005. "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing" *Journal of the American Statistical Association* 100(469) 94-108.
https://doi.org/10.1198/016214504000000539

Roy, Andrew, 1951. "Some Thoughts on the Distribution of Earnings" *Oxford Economics Papers* 3(2) 135-146.
https://doi.org/10.1093/oxfordjournals.oep.a041827

Schochet, Peter, 2008. *Technical Methods Report: Guidelines for Multiple Testing in Impact Evaluations* NCEE 2008-4018. Washington, DC: National Center for Education Evaluation

and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Sloczyński, Tymon, 2020. "Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights" *Review of Economics and Statistics*, 2022. https://doi.org/10.1162/rest_a_00953

Smith, Jeffrey and Petra Todd, 2005. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125(1) 305-53. https://doi.org/10.1016/j.jeconom.2004.04.011

Smith, Jeffrey and Alexander Whalley. 2021. "How Well Do We Measure Public Job Training?" Unpublished manuscript, University of Wisconsin.

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

Vytlacil, Edward, 2002. "Independence, Monotonicity, and Latent Index Models: An Equivalence Result" *Econometrica* 70(1) 331-41. https://doi.org/10.1111/1468-0262.00277

Wang, Xia, and Yongmiao Hong. 2018. "Characteristic Function Based Testing for Conditional Independence: A Nonparametric Regression Approach" *Econometric Theory* 34(4) 815-849. https://doi.org/10.1017/S026646661700010X

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. https://doi.org/10.1098/rspa.1907.0028

Table 1: Causal Impact of Having More than Three Children on Mothers Labor Supply, Angrist and Evans (1998)

	Parametric OLS	Parametric IV	Semiparametric model	Semiparametric IV model
Worked last year	–0.176***	–0.120***	–0.174***	–0.117***
	(0.002)	(0.025)	(0.002)	(0.025)
Weeks worked	–8.965***	–5.659***	–8.900***	–5.532***
	(0.071)	(1.108)	(0.073)	(1.109)
Hours worked	–6.656***	–4.595***	–6.587***	–4.447***
	(0.061)	(0.945)	(0.062)	(0.946)
Income ($1000)	–3.768***	–1.960***	–3.739	–1.915***
	(0.034)	(0.542)	(0.035)	(0.542)
First Stage:	—	0.062***	—	0.062***
Same sex coeff.		(0.001)		(0.001)
N	394,835	394,835	394,835	394,835

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$

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 Means, Angrist and Evans (1998) Data

	OLS		Semiparametric	
	(CMIA0) test	(CMIA1) test	(CMIA0) test	(CMIA1) test
Dependent variable	Y_0	Y_1	Y_0	Y_1
Worked last year				
Coefficient on instrument	0.005*	0.001	0.005**	0.002
(standard error)	(0.002)	(0.003)	(0.002)	(0.003)
[selection effect]	[0.047]	[0.010]	[0.052]	[0.012]
Weeks worked				
Coefficient on instrument	0.297***	0.053	0.315***	0.063
(standard error)	(0.090)	(0.104)	(0.090)	(0.105)
[selection effect]	[3.033]	[0.370]	[3.218]	[0.443]
Hours worked				
Coefficient on instrument	0.205**	–0.000	0.221**	0.016
(standard error)	(0.075)	(0.093)	(0.075)	(0.093)
[selection effect]	[2.086]	[-0.003]	[2.258]	[0.112]
Income				
Coefficient on instrument	0.188***	–0.007	0.194***	–0.005
(standard error)	(0.045)	(0.048)	(0.045)	(0.048)
[selection effect]	[1.920]	[–0.049]	[1.978]	[–0.039]
N	236,092	158,743	236,092	158,743

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$

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 model uses a fully saturated model in the covariates.

Table 3: Test of CIA using Percentiles and Moments, Angrist and Evans (1998) Data

	Weeks worked		Hours worked		Income	
	(CIA0) test	(CIA1) test	(CIA0) test	(CIA1) test	(CIA1) test	(CIA0) test
Dependent Variable	Y_0	Y_1	Y_0	Y_1	Y_0	Y_1
Percentiles						
Zero	–0.005**	–0.001	–0.005**	–0.001	–0.006**	–0.001
	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)
50th	–0.006***	–0.001	–0.007***	–0.002	–0.007**	–0.002
	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
60th	–0.007***	–0.001	–0.007**	–0.001	–0.005***	0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
70th	–0.006***	–0.002	–0.005**	–0.001	–0.006**	0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
80th	–0.004*	–0.001	–0.002**	0.001	–0.007***	–0.001
	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)
90th	—	—	—	—	–0.002	–0.000
					(0.001)	(0.001)
Moments						
1st moment	0.297***	0.053	0.205**	–0.000	0.188***	–0.007
	(0.090)	(0.104)	(0.075)	(0.093)	(0.045)	(0.048)
2nd moment	0.788	0.860	–0.498	–3.543	0.572*	–0.304
	(1.067)	(1.932)	(1.304)	(1.933)	(0.236)	(2.946)
3rd moment	0.017	0.005	0.004	–0.039	0.381	0.075
	(0.520)	(0.009)	(0.013)	(0.020)	(0.228)	(0.433)
4th moment	0.005	0.007	–0.048	–0.163*	4.271	2.196
	(0.004)	(0.013)	(0.049)	(0.078)	(3.529)	(6.923)
N	236,092	158,743	236,092	158,743	236,092	158,743

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$

Notes: Percentile regressions use a parametric model where the dependent variable is an indicator for being equal to or below a given percentile. All higher moments are centered about the mean. The third and fourth moments are normalized by the standard deviation. Fpr the second moment of income we divide the coefficient and standard error by one million. Covariates 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, and an indicator for whether the first child was a boy, and an indicator for whether the second child was a boy.

Table 4: Causal Impact of Class Size on Test Scores, Angrist and Lavy (1999)

	Reading comprehension	Math
Class size (continuous measure)	–0.098	0.095
	(0.090)	(0.113)
Class size less than 35	1.365	–1.335
	(1.888)	(2.382)
N	415	415

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$

Notes: Results from 2SLS regressions. Dependent variable is the average score of a class. All regressions control for percentage disadvantaged. Our replication of Angrist-Lavy uses their Discontinuity Sample and match results from their Table 5, columns (5) and (11) precisely, and using their continuous instrument. Our treatment indicator is defined as having a class size of less than 35. Our instrument is defined as having the Angrist-Lavy instrument taking on a value of less than 35. The F-statistic on the binary instrument equals 49.4. If we regress our instrument against class size, controlling percentage disadvantaged, the point estimate indicates our instrument results in a reduction of class size of 7.58.

Table 5: Test of CMIA, Angrist and Lavy (1999) Data

	(CMIA0) test Y_0	(CMIA1) test Y_1
Reading comprehension		
Coefficient on instrument	–0.673	1.959
(standard error)	(1.236)	(1.348)
Math		
Coefficient on instrument	–1.934	0.786
(standard error)	(1.762)	(1.567)
N	177	238

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$

Notes: All regressions control for percentage disadvantaged. Our treatment indicator is defined as having a class size of less than 35. Our instrument is defined as having the Angrist-Lavy instrument taking on a value of less than 35.

Table 6: Test of CIA using Percentiles and Moments, Angrist and Lavy (1999) Data

	Reading Comprehension		Math	
	(CIA0) test	(CIA1) test	(CIA0) test	(CIA1) test
Dependent Variable	Y_0	Y_1	Y_0	Y_1
Percentiles				
20th	–0.042	–0.087	–0.003	0.013
	(0.056)	(0.053)	(0.059)	(0.056)
40th	–0.047	–0.127*	0.192*	–0.038
	(0.073)	(0.062)	(0.084)	(0.064)
60th	0.131	–0.134*	0.178*	–0.093
	(0.084)	(0.061)	(0.084)	(0.064)
80th	0.026	–0.067	0.102	–0.075
	(0.071)	(0.054)	(0.075)	(0.052)
Moments				
1st moment	–0.673	1.959	–1.934	0.786
	(1.236)	(1.348)	(1.762)	(1.567)
2nd moment	1.784	11.111	–8.772	29.813
	(14.041)	(13.117)	(17.177)	(18.359)
3rd moment	0.027	0.097	0.348	–0.276
	(0.540)	(0.489)	(0.521)	(0.524)
4th moment	0.815	1.806	–0.672	2.237
	(1.219)	(1.111)	(1.022)	(1.239)
N	177	238	177	238

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$

Notes: Percentile regressions use a parametric model where the dependent variable is an indicator for being equal to or below given percentile. All higher moments are centered about the mean. The third and fourth moments are normalized by the standard deviation. All regressions control for percentage disadvantaged.

Table 7: Impact of Training for Adult Women Recommended for Classroom Training in the National JTPA Study

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)	(0.020)	(0.020)
(n=2,374)		
Bloom estimator		
First-stage treatment indicator	0.305***	0.305***
(standard error)	(0.019)	(0.019)
[F-statistic on instrument]	[246]	[246]
(n=2,374)		
Impact of classroom training on compliers	0.136**	0.122*
(standard error)	(0.067)	(0.065)
(n=2,374)		
Treatment group indicator for regression	0.006	0.005
(standard error)	(0.029)	(0.027)
[selection effect]	[0.013]	[0.011]
(n=1,233)		
Treatment group indicator for regression	0.068**	0.062**
(standard error)	(0.032)	(0.031)
[selection effect]	[0.145]	[0.132]
(n=1,501)		

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$

Notes: 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.

Table 8: Moments from the Angrist and Evans (1998) Data

	Worked last year	Weeks worked	Hours worked	Income
$\bar{Y}_{0,N}$	0.638	24.62	21.57	\$8,819
$\bar{Y}_{0,N\cup C}$	0.633	24.32	21.34	\$8,616
$\bar{Y}_{1,A}$	0.462	15.67	14.96	\$5,051
$\bar{Y}_{1,A\cup C}$	0.463	15.74	14.95	\$5,049
$\bar{Y}_{0,C}$	0.599	22.29	19.79	\$7,258
$\bar{Y}_{1,C}$	0.472	16.23	14.90	\$5,037
N	391,731	391,731	391,731	391,731

Notes: We reweight the covariates – the age of the mother (individual year indicators), the age of the mother at first birth (individual year indicators), 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 – so that each group has the same distribution of covariates as the aggregate data using inverse probability weighting. We drop 3,104 observations from Table 1 and 2 that violate the common support assumption. There are 1,418 cells.

9 Appendix A: Pathological Cases

This appendix describes the pathological cases mentioned in Section 4.1 of the main text. We continue to use the same notation and to assume a binary instrument Z. We omit the conditioning variables X for notational simplicity. The pathological case occurs because the VIA and CMIA assumptions do not imply the absence of selection, defined in our context as $E(Y_j|z) = E(Y_j)$, $j = 0, 1$. We call it pathological because it sits on a knife edge.

The VIA assumptions imply that:

$$E(Y_j) = Pr(A)\ E(Y_j|A) + Pr(C)\ E(Y_j|C) + Pr(N)\ E(Y_j|N) \quad j \in \{0, 1\}. \tag{A1}$$

The CMIA^j states that

$$E(Y_j|D = 1) = E(Y_j|D = 0). \tag{A2}$$

$D = 1$ occurs when an agent is an always-taker, with probability $Pr(A)$, or is a complier with $Z = 1$, with probability $\rho_z Pr(C)$, where $\rho_z = Pr(Z = 1)$, and $D = 0$ occurs when an agent is a never-taker, with probability $Pr(N)$, or is a complier with $Z = 0$, with probability $(1 - \rho_z)Pr(C)$. Rewriting equation (A2) using these probabilities we get

$$\frac{Pr(A)\ E(Y_j|A) + \rho_z Pr(C)E(Y_j|C)}{Pr(A) + \rho_z Pr(C)} = \frac{Pr(N)\ E(Y_j|N) + (1 - \rho_z)Pr(C)\ E(Y_j|C)}{Pr(N) + (1 - \rho_z)Pr(C)}. \tag{A3}$$

Solutions to equation (A3) may be divided into two classes. In the first class of solutions, the equation is solved when $E(Y_j|A) = E(Y_j|C) = E(Y_j|N)$. In this case, it holds for *any* values of ρ_z, $Pr(A)$, $Pr(C)$, and $Pr(N)$.

The second class of solutions allows the conditional expectations $E(Y_j|A), E(Y_j|C)$, and $E(Y_j|N)$ to differ. These solutions require the right combination of values of $E(Y_j|A)$, $E(Y_j|C)$, $E(Y_j|N)$, ρ_z, $Pr(A)$, $Pr(C)$, and $Pr(N)$. The data identify ρ_z, $Pr(A)$, $Pr(C)$, and $Pr(N)$ as well as two of the three conditional expectations. One can generally find a

value of the third, unobserved, expectation that solves equation (A3).

An example will clarify. Suppose researchers are testing CMIA^1 ($j = 1$) and find that $E(Y_1|A) > E(Y_1|C)$, i.e. $\alpha_1 < 0$ in equation (13). If Y_1 is continuously distributed on the real line, we may select a unique value of $E(Y_1|N)$ that satisfies equation (A3) (because the equation is linear in $E(Y_j|N)$), but for every other possible value of $E(Y_j|N)$ the CMIA^1 fails to hold. Moreover, as inspection of equation (A3) makes clear, the pathological value of $E(Y_1|N)$ is such that $E(Y_1|N) > E(Y_1|C)$ so that whatever unobserved variable shifts the conditional expectation $E(Y_1|A)$ also shifts $E(Y_1|N)$ above $E(Y_1|C)$. Thus, while our test will detect violations of the CMIA, it may also incorrectly reject the CMIA in this pathological knife-edge case. As Poincairé (1913, p. 435) notes, "Logic sometimes makes monsters." Fortunately, this monster exists at only a single point.

10 Appendix B: Monte Carlo Study

10.1 Data generating process

This appendix presents a Monte Carlo (MC) analysis that illustrates the finite sample behavior of the test described in the main text of the paper. In particular, we document how the performance of the test changes as we vary (i) instrument strength, (ii) sample size, and (iii) the extent of selection on unobserved variables. The reader will recall that our test involves estimating linear outcome equations that include the instrument as a covariate, with our test statistic consisting of the estimated coefficient on the instrument. As we estimate the linear outcome equation via OLS, the test statistic has all the usual statistical properties (e.g., unbiasedness and asymptotic normality), leading us to focus our analysis not on the sampling behavior of the test statistic itself but instead on the size and power of the test.

For our MC analyses, we specify a generalized Roy model as the population data generating process. In this model, participation status D_i may depend on the instrument Z_i and on an unobserved variable, which in turn may have a non-zero correlation with the unobserved

components of the outcome equations. Thus, it provides us with a rich structure that can embody varying amounts of selection on unobserved variables.

The data are generated from the following data generating process:

$$Y_{1i} = 3 + \epsilon_{1i} \tag{B1}$$

$$Y_{0i} = 1 + \epsilon_{0i} \tag{B2}$$

$$D_i^* = -\frac{\gamma^d}{3} + Z_i\gamma^d + \epsilon_{di} \tag{B3}$$

$$D_i = \begin{cases} 1 & \text{if } D_i^* \geq 0, \rightarrow \text{only } Y_{1i} \text{ is observed} \\ 0 & \text{if } D_i^* < 0, \rightarrow \text{only } Y_{0i} \text{ is observed} \end{cases}.$$

Equations (B1) and (B2) correspond to equations (8) and (9) in the main text. In this model, γ^d captures the strength of the instrument, with higher (absolute) values of γ^d indicating a substantively stronger instrument. We assume that $(\epsilon_{0i}, \epsilon_{1i}, \epsilon_{di}) \perp\!\!\!\perp Z_i$ and draw Z_i from a binomial distribution with $\Pr(Z_i = 0) = \Pr(Z_i = 1) = 0.5$. The error terms follow a joint normal distribution with the covariance structure:

$$\begin{pmatrix} \epsilon_{0i} \\ \epsilon_{1i} \\ \epsilon_{di} \end{pmatrix} \sim iid\,\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \qquad \mathbf{\Sigma} = \begin{pmatrix} \sigma_0^2 & -\rho\sigma_0\sigma_1 & -\rho\sigma_0\sigma_d \\ -\rho\sigma_0\sigma_1 & \sigma_1^2 & \rho\sigma_1\sigma_d \\ -\rho\sigma_0\sigma_d & \rho\sigma_1\sigma_d & \sigma_d^2 \end{pmatrix}.$$

We set $\sigma_0 = \sigma_1 = 2$. Following standard practice in probit models, we normalize $\sigma_d = 1$.

The parameter ρ, which we vary in our MC simulations, determines the correlations among the error terms, with $corr(\epsilon_{1i}, \epsilon_{di}) = \rho$ and $corr(\epsilon_{0i}, \epsilon_{di}) = corr(\epsilon_{1i}, \epsilon_{0i}) = -\rho$. Implicitly, ρ governs the degree of selection on unobserved variables. Setting $\rho = 0$ implies no selection on unobserved variables, so that $\alpha_0 = 0$ and $\alpha_1 = 0$. In contrast, a $\rho \neq 0$ implies that $\alpha_0 \neq 0$ and $\alpha_1 \neq 0$. Put differently, $\rho = 0$ implies that the null hypotheses tested by our test hold in the population, while $\rho \neq 0$ implies that those null hypotheses do not hold in the population. A larger (absolute) value of ρ implies substantively stronger selection on

unobserved variables.

To see how the sign of the selection effect plays out, assume $\rho > 0$ so that $corr(\epsilon_{1i}, \epsilon_{di}) > 0$. Now observe that in equation (B3), $Z_i = 1$ implies that, at the margin, $D_i = 1$ for observations with lower values of ϵ_{di}. In notation, $E[\epsilon_{1i}|Z_i = 1] < 0$ in this case, indicating that the instrument induces negative selection into treatment on (the unobserved component of) Y_{1i} in equation (B1). A parallel argument, but with opposite signs, holds for selection on the unobserved component of Y_{0i} in equation (B2). Note that having a single ρ parameter, rather than positing separate correlations between ϵ_{di} and each of ϵ_{1i} and ϵ_{0i}, imposes important substantive restrictions on the model. In particular, it rules out patterns like the one we find in our analysis of the data from Angrist and Evans (1998) in Section 5.1, with selection into treatment based on (the unobserved component of) Y_{0i} but not (the unobserved component of) Y_{1i}. We make this restriction solely to reduce the number of parameters we have to keep track of in the MC analysis.

We do not include exogenous covariates X_i like those in equations (13) and (14) in the main text in our data generating process because we do not need them to make the points we want to make with the MC analysis given our focus on the size and power of our test. Indeed, given our assumptions on Z_i, including it would not affect the asymptotic distribution of the estimates of the auxilliary functions $g_0(X_i)$ and $g_1(X_i)$. In short, omitting covariates keeps things simple without losing any substance. More broadly, we chose not to undertake an "empirical Monte Carlo" analysis like that in Huber, Lechner and Wunsch (2013) as we want to emphasize the applicability of our test in many quite distinct empirical contexts.

10.2 Monte Carlo details

We conduct two sets of MC analyses to examine the performance of our proposed test. We set the nominal size of the test to 0.05 throughout, which is to say that we reject the null in MC samples in which the estimated test statistic exceeds 1.96 in absolute value. For each set of parameter values and sample size, we draw 1000 Monte Carlo samples from an artificial

population of 100,000,000 observations created based on the DGP described above. In each Monte Carlo sample, we estimate equations (8) and (9) and perform our test as described in the main text. We also estimate the linear probability model analogue to our participation probit, i.e., a linear regression of D_i on Z_i, which provides us with first-stage F-statistics.

In the first set of MC analyes, we fix ρ and vary the sample size and the strength of the instrument as embodied in γ_d. Specifically, we fix $\rho = 0.6$, implying quite strong selection on unobserved variables. Of course, in most actual empirical exercises the researcher would reduce the extent of selection on unobserved variables by including observed variables, as we do not for the reasons noted above; see also the related discussion around equation (36) in Section 6.2 of the main text. Setting $\rho = 0.6$ implies that our nulls that $\alpha_0 = 0$ and $\alpha_1 = 0$ do not hold (i.e. are false) in the population because $\rho \neq 0$. We then vary $\gamma^d \in \{0.00, 0.05, 0.15, 0.25, 0.50, 1.50, 2.50\}$, thereby varying instrument strength from none—i.e., abject failure of the second part of the (EI) assumption in Section 3—to levels of instrument strength rarely seen outside of randomized control trials. To provide a scaling more familiar to applied researchers, we present means of realized first-stage F statistics among our MC results. Finally, we repeat our analyses for sample sizes in $N \in \{100, 500, 1{,}000, 5{,}000, 10{,}000, 50{,}000\}$, which captures the range of samples sizes used in most social science empirical work.

In the second set of MC analyses, we fix the sample size N and vary the instrument strength as embodied in γ_d and the substantive importance of selection on observed variables as embodied in ρ. We fix the sample size first at $N = 1,000$ and then at $N = 10,000$. We vary $\gamma^d \in \{0.00, 0.05, 0.15, 0.25, 0.50, 1.50, 2.50\}$, just as in the first set of simulations. We examine $\rho \in \{0.0, 0.2, 0.4, 0.6\}$. This varies the degree of selection on unobserved variables from none, which implies that the nulls $\alpha_0 = 0$ and $\alpha_1 = 0$ hold in the population, to the quite substantial amount assumed in the first set of MC analyses.

10.3 Results

Tables B1 and B2 report the results of the MC analyses. Each block of each table consists of nine rows, with the values in each row calculated using the estimates from the 1,000 independent Monte Carlo samples for that block. The first row, labeled "Mean $\hat{\alpha}_0$," gives the mean of the estimated coefficients on Z_i or, equivalently, the mean of the estimated test statistic. We include this row mainly to confirm that our simulations work as intended. The second row, labeled "Std. Dev. $\hat{\alpha}_0$," presents the standard deviation of the estimated coefficients on Z_i. It gives a more continuous sense of the variability in the test statistic across MC samples than the rejection probability provides. The third row, labeled "Mean N_0," gives the average number of untreated units in the Monte Carlo samples for each block. We picked the parameters of our DGP to generate treatment probabilities around 0.5, so this row again serves mainly to confirm the good behavior of our code.

The fourth row, labeled "Prob. Rej. $H_0 : \alpha_0 = 0$" indicates the fraction of MC samples in the block for which the null is rejected at the 0.05 level. In blocks wherein $\rho = 0.0$, so that the null that $\alpha_0 = 0$ holds in the population, this rejection probability represents the realized size of the test, i.e., the realized probability of a Type I error. For blocks wherein $\rho > 0$, so that the null that $\alpha_0 = 0$ does not hold in the population, this probability represents the realized power of the test, i.e., one minus the probability of a Type II error.

The fifth through eighth rows in each block repeat the same statistics as in the first four rows but for the $D_i = 1$ units in each MC sample. Finally, The ninth and final row in each block, labeled "Mean 1st Stage F-stat." offers exactly that, namely the mean of the F-statistics obtained from estimating linear probability models of D_i on Z_i using the MC samples for the block. The first-stage F-statistic is common to the treated and untreated units and so appears only once.

Table B1 reports the results for the first set of Monte Carlo analyses, while Table B2 reports the results from the second set. Taking the tables together reveals several general findings. First, the realized size of the test under the null of no selection on unobserved vari-

ables equals approximately 0.05. Second, the power of the test increases in the sample size, holding constant instrument strength and the amount of selection on unobserved variables. Third, the power of the test increases in instrument strength holding constant sample size and the amount of selection on unobserved variables. Fourth, and finally, the power of the test increases in the degree of selection on unobserved variables, holding constant sample size and instrument strength.

In addition to the completely unsurprising qualitative findings just described, we can also say a bit about the quantitative levels of power, keeping in mind our decision to adopt a very simple DGP that does not closely correspond to any particular empirical context. We frame this discussion in part around the mean first-stage F-statistic, keeping in mind that it increases in both the sample size (and thus across columns in Table B1 and across panels in Table B2) and in the value of γ_d, and thus with blocks in both Tables B1 and B2.

In Table B1, a mean first-stage F-statistic over 100, as demanded by Lee, McCrary, Moreira, and Porter (2021), ensures statistical power close to 1.0. A comparison with Table B2 shows that this relationship need not hold when $\rho < 0.6$, as in the case with $N = 1,000$, $\gamma_d = 1.5$ and $\rho = 0.2$. Even in this case, though, the power reaches a respectable, if not overwhelming, level of 0.47.

First-stage F-statistics exceeding 100 are thin on the ground outside of randomized control trials and nearly sharp regression discontinuity designs. What happens in cases where the mean of the first-stage F-statistics lies above the traditional rule-of-thumb level of 10.0 but below 100.0? Here the realized power depends very much on the extent of selection on unobserved variables and on just where the realized first-stage F-statistic lies within the range from 10 to 100. To see the first point, consider the case in Table B2 with $N = 1,000$ and $\gamma_d = 0.5$. Here the mean first-stage F-statistic equals around 41. With $\rho = 0.2$, this implies a realized power of only about 0.11 relative to both nulls, while when ρ rises to 0.6, the realized power rises to around 0.67. To see the second point, compare the block just considered to the one that precedes it in Table B2, wherein $\gamma_d = 0.25$ rather than $\gamma_d = 0.5$.

In that block, the mean first-stage F-statistic equals about 11 and the realized power ranges from 0.07 when $\rho = 0.2$ to 0.21 when $\rho = 0.6$.

Keeping in mind that we have not specialized our Monte Carlo analysis to any specific empirical context nor have we included any covariates in it, we view it as sending a relatively clear message. In cases with a strong first-stage and substantial amounts of selection on unobserved variables, our test does really well in terms of statistical power. In cases with a meaningful first stage but not an overpowering one, the power of the test increases in the degree of selection on unobserved variables. While generally not close to one, in many such cases it is also quite far from 0.05, suggesting that the benefits of our inexpensive-to-implement test will still exceed its costs for most researchers. Not unrelated, in these intermediate cases rejections are, in an important sense, more informative than failures to reject. Put differently, our test will yield more false negatives than false positives. Finally, in cases with a weak first stage, the researcher has bigger troubles to worry about than the low power of our test.

Table B1(a): Monte Carlo Analyses on Varying Degree of Instrument Strengths and Different Sample Sizes (1/3)

γ^d			N =100	N =500	N =1,000	N =5,000	N =10,000	N =50,000
0	Y_0	Mean $\hat{\alpha}_0$	-0.02	0.00	0.00	0.00	0.00	0.00
		Std. Dev. $\hat{\alpha}_0$	(0.50)	(0.22)	(0.16)	(0.07)	(0.05)	(0.02)
		Mean N_0	49.89	250.18	500.82	2,498.77	5,000.55	25,000.99
		Prob. Rej. $H_0 : \alpha_0 = 0$	0.046	0.049	0.051	0.071	0.050	0.055
	Y_1	Mean $\hat{\alpha}_1$	-0.01	0.00	0.00	0.00	0.00	0.00
		Std. Dev. $\hat{\alpha}_1$	(0.49)	(0.22)	(0.16)	(0.07)	(0.05)	(0.02)
		Mean N_1	50.11	249.82	499.18	2,501.23	4,999.45	24,999.01
		Prob. Rej. $H_0 : \alpha_1 = 0$	0.044	0.053	0.051	0.052	0.048	0.058
	Mean 1st Stage F-stats		1.07	0.98	1.07	0.97	1.05	1.01
0.05	Y_0	Mean $\hat{\alpha}_0$	0.07	0.04	0.04	0.04	0.04	0.04
		Std. Dev. $\hat{\alpha}_0$	(0.51)	(0.22)	(0.16)	(0.07)	(0.05)	(0.02)
		Mean N_0	49.2	248.2	497.18	2,483.55	4,968.35	24,837.09
		Prob. Rej. $H_0 : \alpha_0 = 0$	0.044	0.054	0.059	0.083	0.117	0.390
	Y_1	Mean $\hat{\alpha}_1$	-0.03	-0.03	-0.04	-0.04	-0.04	-0.04
		Std. Dev. $\hat{\alpha}_1$	(0.50)	(0.23)	(0.15)	(0.07)	(0.05)	(0.02)
		Mean N_1	50.8	251.8	502.82	2,516.45	5,031.65	25,162.91
		Prob. Rej. $H_0 : \alpha_1 = 0$	0.045	0.059	0.056	0.071	0.133	0.418
	Mean 1st Stage F-stats		0.99	1.25	1.34	3.02	4.91	21.15
0.15	Y_0	Mean $\hat{\alpha}_0$	0.10	0.12	0.13	0.12	0.11	0.12
		Std. Dev. $\hat{\alpha}_0$	(0.51)	(0.22)	(0.17)	(0.07)	(0.05)	(0.02)
		Mean N_0	49.08	244.83	490.6	2,448.87	4,899.37	24,505.76
		Prob. Rej. $H_0 : \alpha_0 = 0$	0.061	0.079	0.134	0.377	0.632	0.999
	Y_1	Mean $\hat{\alpha}_1$	-0.11	-0.12	-0.12	-0.11	-0.11	-0.11
		Std. Dev. $\hat{\alpha}_1$	(0.51)	(0.23)	(0.16)	(0.07)	(0.05)	(0.02)
		Mean N_1	50.92	255.17	509.4	2,551.13	5,100.63	25,494.24
		Prob. Rej. $H_0 : \alpha_1 = 0$	0.055	0.094	0.120	0.354	0.594	1.000
	Mean 1st Stage F-stats		1.30	2.81	4.62	18.86	36.87	180.62

Table B1(b): Monte Carlo Analyses on Varying Degree of Instrument Strengths and Different Sample Sizes (2/3)

γ^d			N =100	N =500	N =1,000	N =5,000	N =10,000	N =50,000
0.25	Y_0	Mean $\hat{\alpha}_0$	0.22	0.18	0.20	0.19	0.20	0.19
		Std. Dev. $\hat{\alpha}_0$	(0.50)	(0.22)	(0.16)	(0.07)	(0.05)	(0.02)
		Mean N_0	48.15	241.11	482.33	2,416.13	4,834.96	24,168.62
		Prob. Rej. $H_0 : \alpha_0 = 0$	0.063	0.122	0.216	0.756	0.975	1.000
	Y_1	Mean $\hat{\alpha}_1$	-0.17	-0.19	-0.19	-0.19	-0.19	-0.19
		Std. Dev. $\hat{\alpha}_1$	(0.48)	(0.22)	(0.15)	(0.07)	(0.05)	(0.02)
		Mean N_1	51.85	258.89	517.67	2,583.87	5,165.04	25,831.38
		Prob. Rej. $H_0 : \alpha_1 = 0$	0.055	0.126	0.219	0.784	0.959	1.000
	Mean 1st Stage F-stats		2.10	6.00	11.31	51.28	102.22	497.11
0.5	Y_0	Mean $\hat{\alpha}_0$	0.40	0.39	0.39	0.39	0.39	0.39
		Std. Dev. $\hat{\alpha}_0$	(0.51)	(0.22)	(0.16)	(0.08)	(0.05)	(0.02)
		Mean N_0	47.06	234.2	468.45	2,340.41	4,676.3	23,387.59
		Prob. Rej. $H_0 : \alpha_0 = 0$	0.096	0.388	0.652	1.000	1.000	1.000
	Y_1	Mean $\hat{\alpha}_1$	-0.35	-0.38	-0.38	-0.37	-0.37	-0.37
		Std. Dev. $\hat{\alpha}_1$	(0.49)	(0.21)	(0.16)	(0.07)	(0.05)	(0.02)
		Mean N_1	52.94	265.8	531.55	2,659.59	5,323.7	26,612.41
		Prob. Rej. $H_0 : \alpha_1 = 0$	0.108	0.391	0.636	1.000	1.000	1.000
	Mean 1st Stage F-stats		5.20	21.44	41.83	203.52	406.88	2,021.02

Table B1(c): Monte Carlo Analyses on Varying Degree of Instrument Strengths and Different Sample Sizes (3/3)

γ^d			N =100	N =500	N =1,000	N =5,000	N =10,000	N =50,000
1.5	Y_0	Mean $\hat{\alpha}_0$	1.23	1.22	1.23	1.23	1.22	1.22
		Std. Dev. $\hat{\alpha}_0$	(0.69)	(0.29)	(0.21)	(0.09)	(0.07)	(0.03)
		Mean N_0	42.78	213.11	425.57	2,126.19	4,250	21,259.33
		Prob. Rej. $H_0 : \alpha_0 = 0$	0.388	0.978	1.000	1.000	1.000	1.000
	Y_1	Mean $\hat{\alpha}_1$	-1.02	-1.01	-1.03	-1.02	-1.02	-1.02
		Std. Dev. $\hat{\alpha}_1$	(0.57)	(0.23)	(0.16)	(0.07)	(0.05)	(0.02)
		Mean N_1	57.22	286.89	574.43	2,873.81	5,750	28,740.67
		Prob. Rej. $H_0 : \alpha_1 = 0$	0.452	0.991	1.000	1.000	1.000	1.000
	Mean 1st Stage F-stats		43.60	206.60	409.95	2,045.60	4,091.17	20,458.59
2.5	Y_0	Mean $\hat{\alpha}_0$	2.05	2.08	2.08	2.09	2.08	2.08
		Std. Dev. $\hat{\alpha}_0$	(1.19)	(0.52)	(0.36)	(0.16)	(0.11)	(0.05)
		Mean N_0	42.12	211.08	422.32	2,113.3	4,230.46	21,138.08
		Prob. Rej. $H_0 : \alpha_0 = 0$	0.341	0.953	0.998	1.000	1.000	1.000
	Y_1	Mean $\hat{\alpha}_1$	-1.52	-1.54	-1.54	-1.55	-1.54	-1.55
		Std. Dev. $\hat{\alpha}_1$	(0.66)	(0.27)	(0.19)	(0.09)	(0.06)	(0.03)
		Mean N_1	57.88	288.92	577.68	2,886.7	5,769.54	28,861.92
		Prob. Rej. $H_0 : \alpha_1 = 0$	0.616	1.000	1.000	1.000	1.000	1.000
	Mean 1st Stage F-stats		146.05	692.68	1,372.45	6,801.57	13,623.12	68,008.21

Notes. *: p¡0.05, **: p¡0.01, ***: p¡0.001. This table reports the estimation results for the first set of Monte Carlo analyses, where we fix the degree of selection on unobservables at $\rho = 0.6$. Y_0 and Y_1 refer to the dependent variable in the auxiliary regression. N refers to the sample size. γ^d represents the strength of the instrument in the data generating process. The "Mean $\hat{\alpha}_0$" and "Mean $\hat{\alpha}_1$" rows report the average coefficient estimates on the auxiliary regressor Z_i, which is our test statistic. The "Std. Dev. $\hat{\alpha}_0$" and "Std. Dev. $\hat{\alpha}_1$" rows report its standard deviations. The "Mean N_0" and "Mean N_1" rows provide the average number of $D_i = 0$ and $D_i = 1$ units, respectively. The "Prob. Rej. $H_0 : \alpha_0 = 0$" and "Prob. Rej. $H_0 : \alpha_1 = 0$" rows report the proportion of samples in which the data reject the corresponding null when the level of the test is 0.05. The "Mean 1st Stage F-stats" rows reports the average first-stage F-statistics. All reported means and standard deviations refer to the 1,000 MC samples.

Table B2(a): Monte Carlo Analyses on Varying Degree of Instrument Strengths and Selection on Unobservables (N =1,000) (1/3)

γ^d			ρ =0	ρ =0.2	ρ =0.4	ρ =0.6
0	Y_0	Mean $\hat{\alpha}_0$	0.00	0.00	-0.01	0.00
		Std. Dev. $\hat{\alpha}_0$	(0.18)	(0.17)	(0.17)	(0.15)
		Mean N_0	499.98	499.71	499.21	499.16
		Prob. Rej. $H_0 : \alpha_0 = 0$	0.057	0.037	0.046	0.040
	Y_1	Mean $\hat{\alpha}_1$	-0.01	0.00	0.00	0.00
		Std. Dev. $\hat{\alpha}_1$	(0.17)	(0.18)	(0.17)	(0.15)
		Mean N_1	500.02	500.29	500.79	500.84
		Prob. Rej. $H_0 : \alpha_1 = 0$	0.044	0.056	0.039	0.050
	Mean 1st Stage F-stats		1.01	1.12	1.03	1.02
0.05	Y_0	Mean $\hat{\alpha}_0$	0.01	0.01	0.03	0.05
		Std. Dev. $\hat{\alpha}_0$	(0.18)	(0.18)	(0.17)	(0.16)
		Mean N_0	496.96	497.11	496.12	495.44
		Prob. Rej. $H_0 : \alpha_0 = 0$	0.046	0.051	0.050	0.058
	Y_1	Mean $\hat{\alpha}_1$	0.00	-0.01	-0.03	-0.02
		Std. Dev. $\hat{\alpha}_1$	(0.18)	(0.18)	(0.17)	(0.16)
		Mean N_1	503.04	502.89	503.88	504.56
		Prob. Rej. $H_0 : \alpha_1 = 0$	0.047	0.059	0.053	0.061
	Mean 1st Stage F-stats		1.34	1.33	1.50	1.39
0.15	Y_0	Mean $\hat{\alpha}_0$	0.00	0.04	0.07	0.12
		Std. Dev. $\hat{\alpha}_0$	(0.18)	(0.17)	(0.17)	(0.16)
		Mean N_0	489.97	489.57	490.26	490.06
		Prob. Rej. $H_0 : \alpha_0 = 0$	0.049	0.042	0.061	0.103
	Y_1	Mean $\hat{\alpha}_1$	0.01	-0.04	-0.08	-0.11
		Std. Dev. $\hat{\alpha}_1$	(0.17)	(0.18)	(0.18)	(0.16)
		Mean N_1	510.03	510.43	509.74	509.94
		Prob. Rej. $H_0 : \alpha_1 = 0$	0.044	0.056	0.096	0.104
	Mean 1st Stage F-stats		4.73	4.70	4.45	4.52

Table B2(b): Monte Carlo Analyses on Varying Degree of Instrument Strengths and Selection on Unobservables (N =1,000) (2/3)

γ^d			ρ =0	ρ =0.2	ρ =0.4	ρ =0.6
0.25	Y_0	Mean $\hat{\alpha}_0$	-0.01	0.06	0.13	0.19
		Std. Dev. $\hat{\alpha}_0$	(0.19)	(0.19)	(0.18)	(0.17)
		Mean N_0	484.46	483.00	483.19	483.43
		Prob. Rej. $H_0 : \alpha_0 = 0$	0.057	0.074	0.123	0.215
	Y_1	Mean $\hat{\alpha}_1$	0.00	-0.06	-0.13	-0.19
		Std. Dev. $\hat{\alpha}_1$	(0.17)	(0.18)	(0.17)	(0.15)
		Mean N_1	515.54	517.00	516.81	516.57
		Prob. Rej. $H_0 : \alpha_1 = 0$	0.039	0.069	0.117	0.213
	Mean 1st Stage F-stats		10.97	10.87	10.82	11.32
0.5	Y_0	Mean $\hat{\alpha}_0$	0.00	0.14	0.26	0.40
		Std. Dev. $\hat{\alpha}_0$	(0.20)	(0.19)	(0.18)	(0.17)
		Mean N_0	467.71	467.45	467.4	467.73
		Prob. Rej. $H_0 : \alpha_0 = 0$	0.053	0.116	0.289	0.654
	Y_1	Mean $\hat{\alpha}_1$	-0.01	-0.12	-0.25	-0.38
		Std. Dev. $\hat{\alpha}_1$	(0.18)	(0.18)	(0.17)	(0.15)
		Mean N_1	532.29	532.55	532.6	532.27
		Prob. Rej. $H_0 : \alpha_1 = 0$	0.055	0.103	0.321	0.680
	Mean 1st Stage F-stats		42.06	40.94	41.36	42.61

Table B2(c): Monte Carlo Analyses on Varying Degree of Instrument Strengths and Selection on Unobservables (N =1,000) (3/3)

γ^d			ρ =0	ρ =0.2	ρ =0.4	ρ =0.6
1.5	Y_0	Mean $\hat{\alpha}_0$	0.00	0.40	0.81	1.22
		Std. Dev. $\hat{\alpha}_0$	(0.24)	(0.25)	(0.24)	(0.21)
		Mean N_0	425.67	424.54	423.89	424.25
		Prob. Rej. $H_0 : \alpha_0 = 0$	0.051	0.374	0.926	1.00
	Y_1	Mean $\hat{\alpha}_1$	-0.01	-0.35	-0.68	-1.02
		Std. Dev. $\hat{\alpha}_1$	(0.20)	(0.19)	(0.18)	(0.16)
		Mean N_1	574.33	575.46	576.11	575.75
		Prob. Rej. $H_0 : \alpha_1 = 0$	0.061	0.469	0.959	1.00
	Mean 1st Stage F-stats		411.36	410.20	413.54	416.12
2.5	Y_0	Mean $\hat{\alpha}_0$	0.02	0.69	1.40	2.10
		Std. Dev. $\hat{\alpha}_0$	(0.45)	(0.42)	(0.40)	(0.36)
		Mean N_0	422.84	422.76	422.5	422.17
		Prob. Rej. $H_0 : \alpha_0 = 0$	0.061	0.388	0.914	0.999
	Y_1	Mean $\hat{\alpha}_1$	0.00	-0.52	-1.04	-1.55
		Std. Dev. $\hat{\alpha}_1$	(0.22)	(0.22)	(0.21)	(0.19)
		Mean N_1	577.16	577.24	577.5	577.83
		Prob. Rej. $H_0 : \alpha_1 = 0$	0.049	0.644	0.998	1.000
	Mean 1st Stage F-stats		1,373.49	1,376.46	1,373.56	1,366.90

Notes. *: p¡0.05, **: p¡0.01, ***: p¡0.001. This table reports the estimation results for the second set of Monte Carlo analyses, where we fix the sample size at $N = 1,000$. Y_0 and Y_1 refer to the dependent variable in the auxiliary regression. N refers to the sample size. γ^d represents the strength of the instrument in the data generating process. The "Mean $\hat{\alpha}_0$" and "Mean $\hat{\alpha}_1$" rows report the average coefficient estimates on the auxiliary regressor Z_i, which is our test statistic. The "Std. Dev. $\hat{\alpha}_0$" and "Std. Dev. $\hat{\alpha}_1$" rows report its standard deviations. The "Mean N_0" and "Mean N_1" rows provide the average number of $D_i = 0$ and $D_i = 1$ units, respectively. The "Prob. Rej. $H_0 : \alpha_0 = 0$" and "Prob. Rej. $H_0 : \alpha_1 = 0$" rows report the proportion of samples in which the data reject the corresponding null when the level of the test is 0.05. The "Mean 1st Stage F-stats" rows reports the average first-stage F-statistics. All reported means and standard deviations refer to the 1,000 MC samples.

Table B2(d): Monte Carlo Analyses on Varying Degree of Instrument Strengths and Selection on Unobservables (N =10,000) (1/3)

γ^d			ρ =0	ρ =0.2	ρ =0.4	ρ =0.6
0	Y_0	Mean $\hat{\alpha}_0$	0.00	0.00	0.01	0.00
		Std. Dev. $\hat{\alpha}_0$	(0.06)	(0.06)	(0.05)	(0.05)
		Mean N_0	4,997.33	4,997.92	4,999.98	4,997.98
		Prob. Rej. $H_0 : \alpha_0 = 0$	0.056	0.049	0.052	0.052
	Y_1	Mean $\hat{\alpha}_1$	0.00	0.00	0.00	0.00
		Std. Dev. $\hat{\alpha}_1$	(0.05)	(0.05)	(0.05)	(0.05)
		Mean N_1	5,002.67	5,002.08	5,000.02	5,002.02
		Prob. Rej. $H_0 : \alpha_1 = 0$	0.054	0.042	0.046	0.041
	Mean 1st Stage F-stats		1.02	0.99	1.00	1.01
0.05	Y_0	Mean $\hat{\alpha}_0$	0.00	0.01	0.02	0.04
		Std. Dev. $\hat{\alpha}_0$	(0.06)	(0.06)	(0.05)	(0.05)
		Mean N_0	4,964.93	4,968.46	4,969.12	4,965.2
		Prob. Rej. $H_0 : \alpha_0 = 0$	0.057	0.055	0.069	0.113
	Y_1	Mean $\hat{\alpha}_1$	0.00	-0.01	-0.03	-0.04
		Std. Dev. $\hat{\alpha}_1$	(0.06)	(0.06)	(0.05)	(0.05)
		Mean N_1	5,035.07	5,031.54	5,030.88	5,034.8
		Prob. Rej. $H_0 : \alpha_1 = 0$	0.047	0.047	0.088	0.118
	Mean 1st Stage F-stats		4.83	5.04	4.81	4.79
0.15	Y_0	Mean $\hat{\alpha}_0$	0.00	0.04	0.08	0.12
		Std. Dev. $\hat{\alpha}_0$	(0.06)	(0.06)	(0.05)	(0.05)
		Mean N_0	4,898.89	4,901.47	4,899.88	4,902.14
		Prob. Rej. $H_0 : \alpha_0 = 0$	0.046	0.115	0.328	0.663
	Y_1	Mean $\hat{\alpha}_1$	0.00	-0.04	-0.08	-0.12
		Std. Dev. $\hat{\alpha}_1$	(0.06)	(0.05)	(0.05)	(0.05)
		Mean N_1	5,101.11	5,098.53	5,100.12	5,097.86
		Prob. Rej. $H_0 : \alpha_1 = 0$	0.051	0.103	0.295	0.635
	Mean 1st Stage F-stats		36.95	36.34	37.39	36.81

Table B2(e): Monte Carlo Analyses on Varying Degree of Instrument Strengths and Selection on Unobservables (N =10,000) (2/3)

γ^d			ρ =0	ρ =0.2	ρ =0.4	ρ =0.6
0.25	Y_0	Mean $\hat{\alpha}_0$	0.00	0.06	0.13	0.20
		Std. Dev. $\hat{\alpha}_0$	(0.06)	(0.06)	(0.05)	(0.05)
		Mean N_0	4,834.28	4,834.52	4,834.20	4,835.94
		Prob. Rej. $H_0 : \alpha_0 = 0$	0.047	0.210	0.657	0.975
	Y_1	Mean $\hat{\alpha}_1$	0.00	-0.06	-0.13	-0.19
		Std. Dev. $\hat{\alpha}_1$	(0.06)	(0.05)	(0.05)	(0.05)
		Mean N_1	5,165.72	5,165.48	5,165.80	5,164.06
		Prob. Rej. $H_0 : \alpha_1 = 0$	0.051	0.189	0.666	0.962
	Mean 1st Stage F-stats		102.23	101.65	101.06	100.94
0.5	Y_0	Mean $\hat{\alpha}_0$	0.00	0.13	0.26	0.39
		Std. Dev. $\hat{\alpha}_0$	(0.06)	(0.06)	(0.06)	(0.05)
		Mean N_0	4,678.93	4,678.31	4,677.16	4,678.46
		Prob. Rej. $H_0 : \alpha_0 = 0$	0.055	0.610	0.997	1.00
	Y_1	Mean $\hat{\alpha}_1$	0.00	-0.12	-0.25	-0.37
		Std. Dev. $\hat{\alpha}_1$	(0.06)	(0.06)	(0.05)	(0.05)
		Mean N_1	5,321.07	5,321.69	5,322.84	5,321.54
		Prob. Rej. $H_0 : \alpha_1 = 0$	0.052	0.620	0.998	1.000
	Mean 1st Stage F-stats		405.87	403.92	405.56	407.14

Table B2(f): Monte Carlo Analyses on Varying Degree of Instrument Strengths and Selection on Unobservables (N =10,000) (3/3)

γ^d			ρ =0	ρ =0.2	ρ =0.4	ρ =0.6
1.5	Y_0	Mean $\hat{\alpha}_0$	0.00	0.41	0.81	1.22
		Std. Dev. $\hat{\alpha}_0$	(0.08)	(0.07)	(0.07)	(0.07)
		Mean N_0	4,251.75	4,253.15	4,248.74	4,251.16
		Prob. Rej. $H_0 : \alpha_0 = 0$	0.058	0.999	1.000	1.000
	Y_1	Mean $\hat{\alpha}_1$	0.00	-0.34	-0.68	-1.02
		Std. Dev. $\hat{\alpha}_1$	(0.06)	(0.06)	(0.06)	(0.05)
		Mean N_1	5,748.25	5,746.85	5,751.26	5,748.84
		Prob. Rej. $H_0 : \alpha_1 = 0$	0.038	1.000	1.000	1.000
	Mean 1st Stage F-stats		4,104.46	4,091.92	4,093.64	4,098.88
2.5	Y_0	Mean $\hat{\alpha}_0$	0.00	0.69	1.39	2.07
		Std. Dev. $\hat{\alpha}_0$	(0.13)	(0.13)	(0.12)	(0.11)
		Mean N_0	4,228.44	4,226.26	4,229.89	4,226.3
		Prob. Rej. $H_0 : \alpha_0 = 0$	0.048	0.999	1.000	1.000
	Y_1	Mean $\hat{\alpha}_1$	0.00	-0.52	-1.03	-1.55
		Std. Dev. $\hat{\alpha}_1$	(0.07)	(0.07)	(0.06)	(0.06)
		Mean N_1	5,771.56	5,773.74	5,770.11	5,773.7
		Prob. Rej. $H_0 : \alpha_1 = 0$	0.045	1.000	1.000	1.000
	Mean 1st Stage F-stats		13,607.83	13,601.76	13,616.54	13,583.13

Notes. *: p¡0.05, **: p¡0.01, ***: p¡0.001. This table reports the estimation results for the second set of Monte Carlo analyses, where we fix the sample size at $N = 10,000$. Y_0 and Y_1 refer to the dependent variable in the auxiliary regression. N refers to the sample size. γ^d represents the strength of the instrument in the data generating process. The "Mean $\hat{\alpha}_0$" and "Mean $\hat{\alpha}_1$" rows report the average coefficient estimates on the auxiliary regressor Z_i, which is our test statistic. The "Std. Dev. $\hat{\alpha}_0$" and "Std. Dev. $\hat{\alpha}_1$" rows report its standard deviations. The "Mean N_0" and "Mean N_1" rows provide the average number of $D_i = 0$ and $D_i = 1$ units, respectively. The "Prob. Rej. $H_0 : \alpha_0 = 0$" and "Prob. Rej. $H_0 : \alpha_1 = 0$" rows report the proportion of samples in which the data reject the corresponding null when the level of the test is 0.05. The "Mean 1st Stage F-stats" rows reports the average first-stage F-statistics. All reported means and standard deviations refer to the 1,000 MC samples.