Advances in Economics and Econometrics

Tenth World Congress

Edited by Daron Acemoglu, Manuel Arellano, Eddie Dekel

Book DOI: http://dx.doi.org/10.1017/CBO9781139060035

Online ISBN: 9781139060035

Hardback ISBN: 9781107016064

Paperback ISBN: 9781107627314

Chapter

11 - ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework pp. 401-434

Chapter DOI: http://dx.doi.org/10.1017/CBO9781139060035.012

Cambridge University Press

ELEVEN

ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework

Joshua D. Angrist and Iván Fernández-Val

1.0 Introduction

Local Average Treatment Effects (LATE) capture the causal effect of an instrument-induced shift in treatment. This effect is necessarily tied to the instrument that generates the shift. The interpretation of instrumental variable (IV) estimates as instrument-specific should not be surprising or troubling – when this point is cast in terms of specific examples, we wonder how it could be otherwise. Quarter-of-birth instruments for a wage equation reveal the payoff to schooling induced by compulsory-attendance laws and not the value of a bought-and-paid-for MBA. Still, a clear statement of the nature of causal effects revealed by any instrument raises questions about the external validity of this estimate. Can we use a given IV estimate to identify the effects induced by another source of variation? What about an unconditional average effect? Can we go from average effects on compliers to average effects on the entire treated population?

The usual answer to these questions is "no." Except in special cases, we cannot go farther, at least not without additional assumptions. As described by Angrist, Imbens, and Rubin (1996), the treated population includes two groups: (1) compliers whose behavior is affected by the instrument at hand, and (2) always-takers who are treated irrespective of whether a Bernoulli instrument is switched off or on. The nontreated are likewise composed of compliers and never-takers; the latter group avoids treatment

Our thanks to Alberto Abadie, Manuel Arellano, Gary Chamberlain, Victor Chernozhukov, Guido Imbens, Frank Vella, participants in the 2010 Econometric Society World Congress, and seminar participants at Boston College, Boston University, Carnegie Mellon, Georgetown, Michigan, and the Harvard–MIT Econometrics Workshops for helpful discussions and comments.

no matter what. In the absence of strong homogeneity or distributional assumptions, the data are uninformative for always-takers and never-takers. Moreover, each instrument typically generates its own compliant subpopulation. Effects for one group of compliers need not generalize to another.

One route to external validity is structural. A latent-index choice framework sometimes allows us to fill gaps in the data. For example, Heckman, Tobias, and Vytlacil (2001, 2003) and Angrist (2004) used parametric latent-index models to identify and compare alternative causal effects. Chamberlain (2010) developed a Bayesian semiparametric procedure for extrapolation that relies on models for variation in outcome distributions as a function of the first stage. In a recent paper, Heckman (2010) summarized a literature on IV that establishes theoretical links between parameters such as LATE and effects on the treated. There is no free lunch, however; these links provide *nonparametric* identification of effects other than LATE only with instruments that drive the probability of treatment over a wide range (in fact, from zero to one if we hope to recover the population average causal effect). Such "super-instruments" are rare if not unknown in applied work. In practice, most instruments that identify casual effects are discrete with finite support, and many are Bernoulli.

These theoretical challenges notwithstanding, the predictive value of a particular set of IV estimates may be revealed empirically when a researcher succeeds in isolating multiple instruments for the same underlying causal relation. A pioneering effort in this direction is the Oreopoulos (2006) study of the economic returns to schooling. Oreopoulos compared IV estimates of the returns to schooling across instruments of different strengths. Some of Oreopoulos's instruments are derived from compulsory-attendance policies that had modest effects on schooling. However, two of his policy experiments generate instruments with large first-stage effects, close to a half-year increase in schooling. Moreover, in these examples, there are few never-takers, so LATE is the average effect of treatment on the nontreated. As it turns out, Oreopoulos's IV estimates of the returns to schooling using marginal and full-bore instruments are similar, suggesting a robust causal effect that is likely to have considerable predictive value. Angrist, Lavy, and Schlosser (2010) made a similar homogeneity argument for IV estimates of the causal effects of family size on human capital (a relationship known as the "quantity–quality trade-off"). Sex-composition instruments, which have a modest first stage, generate causal effects similar to those found

using twins instruments, for which the first stage is larger by an order of magnitude, with no never-takers.[1]

These examples are encouraging because they suggest that in a number of important applications, IV estimates are reasonably stable across instruments. In many applications, however, heterogeneous effects are likely to be important. For example, Ichino and Winter-Ebmer (1999) documented substantial heterogeneity in alternative IV estimates of the returns to schooling in Germany. The range of variation in this case far exceeds that which can be attributed to sampling variance.

In this chapter, we ask whether instrumental variable estimates using different instruments, possibly with very different compliant subpopulations, can be reconciled solely by differences in the observed characteristics of compliers. An important consequence of the Abadie (2003) weighting theorem is that the distribution of complier characteristics is identified and easy to describe empirically. A natural first step when comparing alternative IV estimates is to compare and contrast the observed characteristics of compliers. Assuming that treatment-effect heterogeneity is limited in a way that we make precise below, we can use the distribution of complier characteristics to construct an estimator that converts covariate-specific LATEs into broader parameters such as the effect of treatment on the treated and LATE using alternative instruments. Our first contribution is to explain and illustrate this approach to external validity.[2]

The second item on our agenda is the use of overidentification tests in pursuit of external validity. In the classical simultaneous equation framework, statistically significant differences between alternative IV estimates signal a failure of internal validity, perhaps due to violations of the exclusion restriction. In the LATE framework, by contrast, different (internally valid) instruments capture different causal effects. At the same time, covariate-specific overidentification tests and summary conditional tests weighted across covariate cells indicate whether differences in the observed characteristics of compliant subpopulations are enough to explain differences in

[1] See also Ebenstein (2009), who compared LATEs generated by first stages of varying strength for the effect of fertility on labor supply in the United States and Taiwan; and Cruces (2005) and Cruces and Galiani (2007), both of which explored the external validity of the fertility estimates generated by the Angrist and Evans (1998) sex composition instruments using data from Latin America.

[2] Other efforts in this direction include Angrist (2004) and Aronow and Sovey (2010), both of which focused on the possibility of using LATE to learn about unconditional average treatment effects.

unconditional effects. If so, it seems fair to say that the IV estimates at hand meet an empirically useful standard of external validity.

In practice, the question of whether covariates explain the difference between two sets of IV estimates need not have a simple answer. For some covariate values, or perhaps over a certain range, there may be a good match. In other cases, the match will be poor and the underlying estimates essentially unreconciled. We therefore use overidentification test statistics to design a hybrid testing-and-weighting scheme that isolates covariate-defined subsamples for which alternative IV estimates can be reconciled. We think of these samples as coming from a subpopulation for which heterogeneity in treatment effects is a function solely of observed characteristics. For this subpopulation, the predictive value of IV estimates is likely to be especially high.

The ideas in this chapter are illustrated through a comparison of alternative IV estimates of the labor supply consequences of childbearing. As in the study by Angrist and Evans (1998), the instruments are constructed from twin births and sibling sex composition. These instruments have very different first stages and produce significantly different estimates of the causal effect of a third birth. We show here that differences in the characteristics of instrument-specific complier subpopulations can account for most of the difference between the two sets of IV estimates.

2.0 Framework

We imagine that each individual is associated with two potential outcomes, Y_0 and Y_1, which describe outcomes realized under alternative assignments of a Bernoulli treatment, D. The observed outcome, Y, is linked to potential outcomes as follows:

$$\mathrm{Y} = \mathrm{Y}_0 + (\mathrm{Y}_1 - \mathrm{Y}_0)\mathrm{D}. \tag{1}$$

A random-coefficients notation for this is:

$$\mathrm{Y} = \alpha + r\mathrm{D} + \eta$$

where $\alpha = \mathbb{E}[\mathrm{Y}_0]$, $\eta = \mathrm{Y}_0 - \alpha$, and $r = \mathrm{Y}_1 - \mathrm{Y}_0$ is an individual-level causal effect.

We also define potential treatment status indexed against a Bernoulli instrument, z. Potential treatment status is $\mathrm{D}_1^{\mathrm{z}}$ when the instrument is switched on and $\mathrm{D}_0^{\mathrm{z}}$ when the instrument is switched off. The variables $\mathrm{D}_1^{\mathrm{z}}$ and $\mathrm{D}_0^{\mathrm{z}}$ are superscripted to signal the fact that they are tied to z. Observed

treatment status is:

$$\mathrm{D} = \mathrm{D}_0^{\mathrm{Z}} + (\mathrm{D}_1^{\mathrm{Z}} - \mathrm{D}_0^{\mathrm{Z}})\mathrm{Z}$$

or, in random-coefficients notation,

$$\mathrm{D} = \gamma + p\mathrm{Z} + \upsilon$$

where $\gamma = \mathbb{E}[\mathrm{D}_0^{\mathrm{Z}}]$, $\upsilon = \mathrm{D}_0^{\mathrm{Z}} - \gamma$, and $p = \mathrm{D}_1^{\mathrm{Z}} - \mathrm{D}_0^{\mathrm{Z}}$.

IV using Z as an instrument for the effect of D on Y with no covariates is the Wald (1940) estimator. The Wald estimand can be interpreted as the effect of D on those whose treatment status can be changed by the instrument. Assuming, as we do here, that the instrument can make treatment move only in one direction (e.g., make treatment more likely), those whose treatment status is changed by Z have $\mathrm{D}_1^{\mathrm{Z}} = 1$ and $\mathrm{D}_0^{\mathrm{Z}} = 0$. The causal effect on this group is LATE (Imbens and Angrist 1994). Formally, we have:

Assumption 1 (LATE):

(a) Independence and Exclusion: $(\mathrm{Y}_1, \mathrm{Y}_0, \mathrm{D}_1^{\mathrm{Z}}, \mathrm{D}_0^{\mathrm{Z}}) \perp\!\!\!\perp \mathrm{Z}$.
(b) First-stage: $\mathbb{E}[\mathrm{D}_1^{\mathrm{Z}} - \mathrm{D}_0^{\mathrm{Z}}] \neq 0$ *and* $0 < \mathbb{P}[\mathrm{Z} = 1] < 1$.
(d) Monotonicity: $\mathrm{D}_1^{\mathrm{Z}} \geq \mathrm{D}_0^{\mathrm{Z}}$ *a.s., or vice versa.*

Theorem 1 (LATE): *Under Assumption 1:*

$$\frac{\mathbb{E}[\mathrm{Y} \mid \mathrm{Z} = 1] - \mathbb{E}[\mathrm{Y} \mid \mathrm{Z} = 0]}{\mathbb{E}[\mathrm{D} \mid \mathrm{Z} = 1] - \mathbb{E}[\mathrm{D} \mid \mathrm{Z} = 0]} = \mathbb{E}[\mathrm{Y}_1 - \mathrm{Y}_0 \mid \mathrm{D}_1^{\mathrm{Z}} > \mathrm{D}_0^{\mathrm{Z}}] = \mathbb{E}[r \mid p > 0] =: \Delta^{\mathrm{Z}}.$$

Proof: See Imbens and Angrist (1994). ■

As noted by Angrist, Imbens, and Rubin (1996), the LATE framework partitions the population exposed to an instrument into three instrument-dependent groups. These groups are defined by the way people react to the instrument.

Definition 1 (Groups Defined by Instrument z):

(a) z-Compliers: The subpopulation with $\mathrm{D}_1^{\mathrm{Z}} = 1$ and $\mathrm{D}_0^{\mathrm{Z}} = 0$.
(b) z-Always-takers: The subpopulation with $\mathrm{D}_1^{\mathrm{Z}} = \mathrm{D}_0^{\mathrm{Z}} = 1$.
(c) z-Never-takers: The subpopulation with $\mathrm{D}_1^{\mathrm{Z}} = \mathrm{D}_0^{\mathrm{Z}} = 0$.

LATE using Z as an instrument is the effect of treatment on the group of z-compliers.

Table 1. *Wald Estimates of the Effects of Family Size on Labor Supply*

			Twins Instrument		Same-Sex Instrument		Both
Dependent Variable	Mean	OLS (1)	First Stage (2)	Wald Estimates (3)	First Stage (4)	Wald Estimates (5)	2SLS Estimates (6)
Weeks worked	20.83	−8.98	0.603	−3.28	0.060	−6.36	−3.97
		(0.072)	(0.008)	(0.634)	(0.002)	(1.18)	(0.558)
	Overid: $\chi^2(1)$ (p-value)	–	–	–	–	–	5.3 (0.02)
Employment	0.565	−0.176		−0.076		−0.132	−0.088
		(0.002)		(0.014)		(0.026)	(0.012)
	Overid: $\chi^2(1)$ (p-value)	–	–	–	–	–	3.5 (0.06)

Note: The table reports OLS, Wald, and 2SLS estimates of the effects of a third birth on labor supply using twins and sex composition instruments. Data are from the Angrist and Evans (1998) extract from the 1980 U.S. census 5 percent sample, including women aged 21–35 with at least two children. OLS models include controls for mother's age, age at first birth, ages of the first two children, and dummies for race. The sample size is 394,840.

3.0 Covariate-Driven Heterogeneity

3.1 Two Instruments for One Effect

The case for omitted variables bias in the relationship between fertility and labor supply is clear: Mothers with weak labor force attachment or low earnings potential may be more likely to have children than mothers with strong labor force attachment or high earnings potential. This makes the observed association between family size and employment difficult to interpret because mothers who have big families work less anyway. Angrist and Evans (1998) solved this omitted-variables problem using two instruments, both of which lend themselves to Wald-type estimation strategies.

The first Wald estimator uses twin births, an instrument for the effects of family size introduced by Rosenzweig and Wolpin (1980). The twins instrument in Angrist and Evans (1998) is a dummy indicating multiple second births in a sample of mothers with at least two children. The twins first stage is about 0.6, an estimate reported in Column 2 of Table 1. This means that 40 percent of mothers with two or more children would have had

a third birth anyway; a multiple third birth increases this proportion to 1. The twins instrument rests on the idea that the occurrence of a multiple birth is essentially random – unrelated to potential outcomes or demographic characteristics – and that a multiple birth affects labor supply solely by increasing fertility.

The second Wald estimator in Table 1 uses a dummy for same-sex sibling pairs as an instrument. This is motivated by the fact that American parents with two children are more likely to have a third child if the first two are same-sex than if sex composition is mixed. This is illustrated in Column 4 of Table 1, which shows that parents of same-sex siblings are about 6 percentage points more likely to have a third birth than those with a mixed-sex sibship (the probability of a third birth among parents with a mixed-sex sibship is 0.38). Internal validity of the same-sex instrument rests on the claim that sibling sex composition is essentially random and affects mothers' labor supply solely by increasing fertility.

Twins and sex-composition IV estimates both suggest that the birth of a third child substantially reduces weeks worked and employment. Wald estimates using twins instruments show a precisely estimated reduction in weeks worked of a little more than three weeks, with an employment reduction of about 0.8. These results, which appear in Column 3 of Table 1, are smaller in absolute value than the corresponding ordinary least squares (OLS) estimates reported in Column 1 (the latter include a set of controls listed in the table). This suggests that the OLS estimates are exaggerated by selection bias. It is interesting and perhaps surprising that Wald estimates constructed using a same-sex dummy, reported in Column 5, are larger in magnitude than the twins estimates. The juxtaposition of twins and sex-composition instruments suggests that different instruments need not generate similar estimates of causal effects even if both instruments are valid.

The last column of Table 1 reports 2SLS estimates of childbearing using both twins and same-sex instruments, along with the associated overidentification test statistic. The overidentification test statistic generates p-values of 0.02 and 0.06, implying that the twins and sex-composition IV estimates are at least marginally significantly different from one another.

Twins and same-sex IV estimates reflect behavior in different compliant subpopulations. To show this, we let x be a Bernoulli-distributed characteristic – for example, a dummy indicating college graduates. Are sex-composition compliers more or less likely to be college graduates than other women with two children? This question is answered by the following

complier characteristics ratio:

$$\frac{\mathbb{P}[\mathrm{x}=1 \mid \mathrm{d}_1^z > \mathrm{d}_0^z]}{\mathbb{P}[\mathrm{x}=1]} = \frac{\mathbb{P}[\mathrm{d}_1^z > \mathrm{d}_0^z \mid \mathrm{x}=1]}{\mathbb{P}[\mathrm{d}_1^z > \mathrm{d}_0^z]} = \frac{\mathbb{E}[\mathrm{d} \mid \mathrm{z}=1, \mathrm{x}=1] - \mathbb{E}[\mathrm{d} \mid \mathrm{z}=0, \mathrm{x}=1]}{\mathbb{E}[\mathrm{d} \mid \mathrm{z}=1] - \mathbb{E}[\mathrm{d} \mid \mathrm{z}=0]}$$

where the first equality follows by Bayes rule. This second equality shows that the relative likelihood that a z-complier is a college graduate is given by the ratio of the first stage for college graduates to the overall first stage.

This calculation is illustrated in Table 2, which reports compliers' characteristics ratios for the age of the second-born and mother's schooling as described by dummies for high school graduates, some college, and college graduates. Twins compliers have younger second-born children, reflecting the fact that few women who recently had their second child will have had time to have a third child. The birth of a third child in this group is therefore especially likely to have been caused by a multiple pregnancy. (Among second-borns who are less than about a year old, the only explanation for a third birth in the family is a multiple birth.) This is important because the birth of a third child may matter less if the second child is young, helping explain the finding that twins-IV estimates are smaller than same-sex estimates (Gelbach 2002 showed that the presence of a child younger than age 5 in the household is a key labor-supply mediator).

Twins compliers are more likely to be college graduates than the average mother, whereas sex-composition compliers are less educated. This fact also helps explain the smaller Wald estimates generated by twins instruments because Angrist and Evans (1998) showed that the labor-supply consequences of childbearing decline with mother's schooling.

A general method for constructing the mean or other features of the distribution of covariates for compliers uses Abadie's (2003) kappa-weighting scheme. A consequence of Theorem 3.1 in Abadie (2003) is that:

$$\mathbb{E}[\mathrm{x} \mid \mathrm{d}_1 > \mathrm{d}_0] = \frac{\mathbb{E}[\kappa^z(\mathrm{x})\,\mathrm{x}]}{\mathbb{E}[\kappa^z(\mathrm{x})]} \tag{2}$$

where:

$$\kappa^z(x) = 1 - \frac{\mathrm{d}(1-\mathrm{z})}{1-\mathbb{P}[\mathrm{z}=1 \mid \mathrm{x}=x]} - \frac{(1-\mathrm{d})\mathrm{z}}{\mathbb{P}[\mathrm{z}=1 \mid \mathrm{x}=x]}.$$

Table 2. *Complier Characteristics for Twins and Sex Composition Instruments*

Variable	Population Mean	Mean for Twins Compliers		Mean for Same-Sex Compliers	
	$E[x_{1i}]$	$E[x_{1i}\|D_{1i} > D_{0i}]$	$E[x_{1i}\|D_{1i} > D_{0i}]/E[x_{1i}]$	$E[x_{1i}\|D_{1i} > D_{0i}]$	$E[x_{1i}\|D_{1i} > D_{0i}]/E[x_{1i}]$
	(1)	(2)	(3)	(4)	(5)
			A. Dummy characteristics		
Age of second child is less than or equal to 4 years	0.343	0.449	1.31	0.194	0.565
High school graduate	0.488	0.498	1.02	0.515	1.06
Some college	0.202	0.212	1.05	0.212	1.05
College graduate	0.132	0.151	1.14	0.092	0.702
			B. Discrete, ordered characteristics		
Age of second child	6.59	5.51	0.835	7.14	1.08
Mother's schooling	12.13	12.43	1.03	12.09	1.00

Note: The table reports an analysis of complier characteristics for twins and sex composition instruments. The ratios in columns 3 and 5 in Panel A give the relative likelihood that compliers have the characteristic indicated at left. The values in columns 2 and 4 in Panel B represent Abadie's (2003) kappa-weighted means. Data are from the 1980 census 5 percent sample including mothers aged 21–35 with at least two children, as in Angrist and Evans (1998). The sample size is 394,840.

Intuitively, this works because, as Abadie showed, the weighting function, $\kappa^z(x)$, "finds compliers," even though it is not a simple indicator for compliers. Estimates of $\mathbb{E}[\mathrm{x} \mid \mathrm{d}_1 > \mathrm{d}_0]$ for age of second child and mother's education are reported in the last two rows of Table 2. These estimates show a marked difference in the average second-child age and a smaller difference in schooling. The main difference between the schooling of twins and same-sex compliers is in the proportion of college graduates.

3.2 Covariates and Extrapolation

Covariates play two roles in our analysis. First, they may be necessary for identification. For example, we might want to control for race and maternal age when using twins instruments because the probability of multiple births varies by race and increases with maternal age. Second, we use covariates for extrapolation. Specifically, we argue that in some cases, including the twins and same-sex comparisons, variation in causal effects across covariate cells is sufficient to explain differences between IV estimates.

The foundation of our analysis with covariates is a *conditional* independence assumption. This assumption expresses the idea that we think of the instruments as being "as good as randomly assigned," conditional on covariates, x. Generalizing Assumption 1, we have:

Assumption 2 (Conditional LATE):

(a) Independence and Exclusion: $(\mathrm{y}_1, \mathrm{y}_0, \mathrm{d}_1^z, \mathrm{d}_0^z) \amalg \mathrm{z} \mid \mathrm{x}$ *a.s.:*
(b) First-stage: $\mathbb{E}[\mathrm{d}_1^z - \mathrm{d}_0^z \mid \mathrm{x}] \neq 0$ *and* $0 < \mathbb{P}[\mathrm{z} = 1 \mid \mathrm{x}] < 1$ *a.s.*
(c) Monotonicity: $\mathbb{P}[\mathrm{d}_1^z \geq \mathrm{d}_0^z \mid \mathrm{x}] = 1$ *a.s., or* $\mathbb{P}[\mathrm{d}_1^z \leq \mathrm{d}_0^z \mid \mathrm{x}] = 1$ *a.s.*

For each value of x, we define covariate-specific LATE using z as an instrument:

$$\Delta^z(x) := \mathbb{E}[\mathrm{y}_1 - \mathrm{y}_0 \mid \mathrm{d}_1^z > \mathrm{d}_0^z, \mathrm{x} = x]. \tag{3}$$

As noted by Frolich (2007), when conditioning is required for identification, unconditional LATE can be constructed by averaging $\Delta^z(x)$:

$$\begin{aligned}\Delta^z &= \mathbb{E}[\Delta^z(\mathrm{x}) \mid \mathrm{d}_1^z > \mathrm{d}_0^z] = \int \Delta^z(x) dF_{\mathrm{x}}(x \mid \mathrm{d}_1^z > \mathrm{d}_0^z) \\ &= \int \Delta^z(x) \frac{\mathbb{E}[\mathrm{d} \mid \mathrm{z} = 1, \mathrm{x} = x] - \mathbb{E}[\mathrm{d} \mid \mathrm{z} = 0, \mathrm{x} = x]}{\mathbb{E}[\mathrm{d} \mid \mathrm{z} = 1] - \mathbb{E}[\mathrm{d} \mid \mathrm{z} = 0]} dF_{\mathrm{x}}(x) \end{aligned} \tag{4}$$

where $F_{\mathrm{X}}(\cdot \mid \mathrm{D}_1^{\mathrm{Z}} > \mathrm{D}_0^{\mathrm{Z}})$ is the distribution of X for Z-compliers and $F_{\mathrm{X}}(\cdot)$ is the distribution of X in the population.

We first show how to construct average causal effects such as the effect on the treated, $\mathbb{E}[\mathrm{Y}_1 - \mathrm{Y}_0 \mid \mathrm{D} = 1]$, from $\Delta^{\mathrm{Z}}(x)$. This is possible because we assume that heterogeneity in causal effects across instruments is entirely due to changes in the observable characteristics of compliers. Specifically, we start with:

Assumption 3 (CEI): *Conditional Effect Ignorability for an instrument* Z*:*

$$\mathbb{E}[\mathrm{Y}_1 - \mathrm{Y}_0 \mid \mathrm{D}_1^{\mathrm{Z}}, \mathrm{D}_0^{\mathrm{Z}}, \mathrm{X}] = \mathbb{E}[\mathrm{Y}_1 - \mathrm{Y}_0 | \mathrm{X}] \quad a.s.$$

A sufficient condition for CEI is:

$$\mathrm{Y}_1 = \mathrm{Y}_0 + g(\mathrm{X}) + \nu$$

where $g(\mathrm{X})$ is any function and ν is mean-independent of $(\mathrm{D}_1^{\mathrm{Z}}, \mathrm{D}_0^{\mathrm{Z}})$ conditional on X. In other words, heterogeneity in average causal effects is solely due to observed covariates.[3]

To see what this assumption means in a latent-index specification, suppose that:

$$\mathrm{D} = 1[h(\mathrm{X}, \mathrm{Z}) > \eta]$$

where η is a random factor involving unobserved costs and benefits of D assumed to be independent of Z conditional on X. This latent-index model characterizes potential treatment assignments as:

$$\mathrm{D}_0^{\mathrm{Z}} = 1[h(\mathrm{X}, 0) > \eta] \text{ and } \mathrm{D}_1^{\mathrm{Z}} = 1[h(\mathrm{X}, 1) > \eta].$$

The associated model for potential outcomes is:

$$\mathrm{Y}_0 = g_0(\mathrm{X}) + \epsilon_0$$
$$\mathrm{Y}_1 = g_1(\mathrm{X}) + \epsilon_1$$

where the errors here have mean zero conditional on the covariates and instrument. Assuming $h(\mathrm{X}, 1) \geq h(\mathrm{X}, 0)$ a.s., conditional LATE can be written:

$$\begin{aligned}\Delta^{\mathrm{Z}}(\mathrm{X}) &= \mathbb{E}[\mathrm{Y}_1 - \mathrm{Y}_0 \mid \mathrm{X}, h(\mathrm{X}, 1) > \eta > h(\mathrm{X}, 0)] \\ &= g_1(\mathrm{X}) - g_0(\mathrm{X}) + \mathbb{E}[\epsilon_1 - \epsilon_0 \mid \mathrm{X}, \mathrm{D}_1^{\mathrm{Z}} > \mathrm{D}_0^{\mathrm{Z}}] \qquad (5)\end{aligned}$$

[3] This is a conditional-on-covariates version of Restriction 2 in Angrist (2004) and is similar to the Frangakis and Rubin (2002) notion of principal stratification, which isolates covariate-defined subpopulations in which selection bias is likely to be minimal.

The CEI assumption implies that $\epsilon_1 - \epsilon_0$ is mean independent of $(\mathrm{D}_1^{\mathrm{z}}, \mathrm{D}_0^{\mathrm{z}})$ conditional on x, so that Equation (5) simplifies to:

$$\Delta^{\mathrm{z}}(\mathrm{x}) = \mathbb{E}[\mathrm{Y}_1 - \mathrm{Y}_0 \mid \mathrm{x}] = g_1(\mathrm{x}) - g_0(\mathrm{x}).$$

In the language of Rubin (1977), CEI is a type of ignorability assumption for treatment effects. Given this ignorability, we might wonder whether we need to be concerned about selection bias in the first place. Under CEI, selection bias arises due to correlation between η and ϵ_0. For example, a latent-index specification compatible with the CEI sets $\epsilon_j = \theta + \xi_j$, $j = 0, 1$, where θ is correlated with η but ξ_j is not.[4] CEI rules out Roy (1951)–type selection into treatment. In other words, η is assumed to be independent of unobserved gains, denoted by $\epsilon_1 - \epsilon_0$ in the latent-index specification. CEI does not rule out selection bias, but it eliminates an important source of heterogeneity in average causal effects.[5] Although the empirical importance of Roy selection has yet to be established, the Roy model is an important econometric benchmark. Here, however, we focus on an effort to manage the treatment-effect heterogeneity driven by observable characteristics.

The latent-index specification can be used to formulate a structural justification for the CEI assumption in our empirical application. We start by combining elements of selection models in Olsen (1980) and Vytlacil (2002):

$$\mathbb{E}[\epsilon_1 - \epsilon_0 \mid \eta, \mathrm{x}] = \rho(\mathrm{x})(\eta - 1/2) \text{ and } \eta \mid \mathrm{x}, \mathrm{z} \sim U(0, 1).$$

Conditional LATE then becomes:

$$\begin{aligned}\Delta^{\mathrm{z}}(\mathrm{x}) &= g_1(\mathrm{x}) - g_0(\mathrm{x}) + \rho(\mathrm{x})\mathbb{E}[\eta - 1/2 \mid h(\mathrm{x}, 1) > \eta > h(\mathrm{x}, 0), \mathrm{x}] \\ &= g_1(\mathrm{x}) - g_0(\mathrm{x}) + \rho(\mathrm{x})[h(\mathrm{x}, 1) + h(\mathrm{x}, 0) - 1]/2.\end{aligned}$$

For each $\mathrm{x} = x$, CEI turns on whether $\rho(x) = 0$.[6] Following Imbens and Newey (2009), we assume that treatment (i.e., fertility) decisions are based on a comparison of predicted benefits and costs of childbearing. Specifically, women choose to have a third child if:

$$1[h(\mathrm{x}, \mathrm{z}) > \eta] = 1\{\lambda(\mathrm{x})\mathbb{E}[\mathrm{Y}_1 - \mathrm{Y}_0 \mid \mathrm{x}, \eta] > c(\mathrm{x}, \mathrm{z})\}$$

[4] See Appendix A for an illustrative limited dependent-variable model with selection bias that satisfies CEI.

[5] This has been noted by others working with models of this type; see, e.g., Vella and Verbeek (1999).

[6] Interestingly, this example has the property that LATE is the unconditional average treatment effect when the first stage is symmetric (i.e., $h(\mathrm{x}, 1) = 1 - h(\mathrm{x}, 0)$). See Angrist (2004) for other selection models with this property.

where $\lambda(x)$ is the weight given to outcome gaps; $c(x, z)$ is the expected cost of having a third child; the instrument, z, is a cost-shifter independent of potential outcomes; and η is private information about $Y_1 - Y_0$, orthogonal to x. Then, $\rho(x)$ is close to zero when either $\lambda(x)$ is small (i.e., labor-supply consequences are of little import), or η matters little given x (e.g., for women with a young second-born who are already at home or for relatively educated women who more easily can afford to pay for child care). Variation in $\rho(x)$ also offers a possible explanation for why CEI might be satisfied for some women but not for others. We consider econometric models in which CEI is partially satisfied in Section 3.4.

3.3 Reweighting LATE

Our covariate-based strategy reweights conditional treatment effects across covariate cells. This is similar to matching estimators designed to control for selection bias; such estimators reweight the conditional-mean function for outcome variables, capturing causal effects when identification is based on a selection-on-observables story (see, e.g., Hahn 1998). In this case, however, we rely on instrumental variables to control for selection bias while using covariates to manage treatment-effect heterogeneity.

Theorem 2 (LATE-Reweight): *Let* z *be an instrument that satisfies Assumption 3 and let* $S^z = s(D_0^z, D_1^z, z)$ *be an indicator for any group defined by* z. *For example, for* z*-compliers, we have* $S^z = D_1^z - D_0^z$*; for the treated,* $S^z = (1 - z)D_0^z + zD_1^z = D$*; for the nontreated,* $S^z = (1 - z)(1 - D_0^z) + z(1 - D_1^z) = 1 - D$*; and for the entire population,* $S^z = 1$. *Under Assumption 2 and* $\mathbb{E}[|Y|] < \infty$*:*

$$\mathbb{E}[Y_1 - Y_0 \mid S^z = 1] = \mathbb{E}[\Delta^z(x) \mid S^z = 1] = \int \Delta^z(x) dF_x(x \mid S^z = 1)$$
$$= \int \Delta^z(x)\omega_s^z(x) dF_x(x)$$

where $\omega_s^z(x) = \mathbb{P}[S^z = 1 \mid x = x]/\mathbb{P}[S^z = 1]$ *and* $\int \omega_s^z(x) dF_x(x) = 1$.

Proof: By the law of iterated expectations:

$$\mathbb{E}[Y_1 - Y_0 \mid S^z = 1] = \mathbb{E}[\mathbb{E}[Y_1 - Y_0 \mid S^z = 1, x] \mid S^z = 1]$$
$$= \mathbb{E}[\mathbb{E}[Y_1 - Y_0 \mid x] \mid S^z = 1] = \mathbb{E}[\Delta^z(x) \mid S^z = 1]$$

where the second and third equalities follow from CEI. By Bayes rule

$$\int \Delta^{z}(x) dF_{x}(x \mid s^{z} = 1) = \int \Delta^{z}(x)\omega_{s}^{z}(x) dF_{x}(x)$$

where $\omega_{s}^{z}(x) = \mathbb{P}[s^{z} = 1 \mid x = x]/\mathbb{P}[s^{z} = 1]$ and

$$\int \omega_{s}^{z}(x) dF_{x}(x) = \mathbb{E}\{\mathbb{P}[s^{z} = 1 \mid x = x]\}/\mathbb{P}[s^{z} = 1] = 1$$

by the law of iterated expectations. ■

The LATE-Reweight theorem allows us to go from LATE to the population average treatment effect (ATE), the effect of treatment on the treated (TOT), and the effect of treatment on the nontreated (TNT). The relevant weighting functions, $\omega_{s}^{z}(x)$, can be written in terms of observed variables as:

$$\omega_{\Delta}^{z}(x) = \frac{\mathbb{E}[d \mid z = 1, x = x] - \mathbb{E}[d \mid z = 0, x = x]}{\mathbb{E}[d \mid z = 1] - \mathbb{E}[d \mid z = 0]} \tag{6}$$

for effects on z-compliers;

$$\omega_{TOT}^{z}(x) = \mathbb{E}[d \mid x = x]/\mathbb{E}[d] \tag{7}$$

for effects on the treated;

$$\omega_{TNT}^{z}(x) = \mathbb{E}[1 - d \mid x = x]/\mathbb{E}[1 - d] \tag{8}$$

for effects on the nontreated; and

$$\omega_{ATE}^{z}(x) = 1 \tag{9}$$

for the population.

3.4 Overidentification

Differences in the observable characteristics of twins and same-sex compliers may explain the difference between the Wald estimates constructed using these two instruments. If so, we can reweight covariate-specific LATEs to go from one to the other. We show this for two instruments, z and w. The difference between the LATE generated by each can be decomposed as:

$$\begin{aligned}\Delta^{z} - \Delta^{w} = &\int [\Delta^{z}(x) - \Delta^{w}(x)]\omega_{\Delta}^{z}(x) dF_{x}(x) \\ &+ \int \Delta^{w}(x)[\omega_{\Delta}^{z}(x) - \omega_{\Delta}^{w}(x)] dF_{x}(x).\end{aligned} \tag{10}$$

The first term reflects differences in conditional LATEs between z-compliers and w-compliers, while the second term captures differences in complier characteristics. If z and w satisfy CEI, the following *compatibility condition* holds:

$$\Delta^{w}(x) = \Delta^{z}(x) \text{ a.s.} \tag{11}$$

and the first term of the decomposition (10) is zero.[7] In other words, CEI for both instruments implies that the instruments w and z satisfy overidentifying restrictions conditional on x. The following theorem shows that in such cases, we can use the distribution of compliers for a hypothetical instrument to construct the treatment effects that might be generated by instruments other than those we have.

Theorem 3 (LATE-Overid): *Let* z *and* w *be two instruments that satisfy Assumptions 2 and 3. Let* $\Delta^{w}(x)$ *be defined as in Equation (3) using the instrument* w. *Let* $s^{w} = s(D_0^{w}, D_1^{w}, w)$ *be an indicator for any group defined by the instrument* w *with corresponding potential treatment assignments* (D_0^{w}, D_1^{w}). *If* $\mathbb{E}[|Y|] < \infty$*:*

$$\mathbb{E}[Y_1 - Y_0 \mid s^{w} = 1] = \int \Delta^{w}(x)\omega_s^{w}(x)dF_x(x) = \int \Delta^{z}(x)\omega_s^{w}(x)dF_x(x). \tag{12}$$

Proof: The first equality in Equation (12) follows from Theorem 2 applied to w. To establish the second equality, note that the law of iterated expectations and CEI for z give

$$\mathbb{E}[Y_1 - Y_0 \mid s^{w} = 1] = \mathbb{E}[\mathbb{E}[Y_1 - Y_0 \mid x] \mid s^{w} = 1] = \mathbb{E}[\Delta^{z}(x) \mid s^{w} = 1].$$

The result then follows by Bayes rule. ■

The LATE-Overid theorem allows us to determine whether differences in the distribution of complier covariates are enough to explain differences in IV estimates across instruments. If so, it seems fair to say that the underlying covariate-specific results have predictive value for subpopulations defined by these covariate values and therefore some claim to external validity.

[7] A weaker alternative to CEI that implies the compatibility condition is $\mathbb{E}[Y_1 - Y_0 \mid D_1^{z} > D_0^{z}, x] = \mathbb{E}[Y_1 - Y_0 \mid D_1^{w} > D_0^{w}, x]$ a.s.; in other words, equality of conditional LATEs only for compliers. We find the stronger CEI assumption more appealing because it seems hard to imagine a mechanism or model that generates this more limited sort of independence without generating full CEI.

CEI also implies that the conditional LATEs, $\Delta^{z}(x)$, are overidentified if we observe both z and w. We can therefore construct more precise estimators for conditional LATEs using cell-by-cell GMM procedures that use both z and w to form moment conditions. Moreover, GMM overidentification tests can be used to statistically assess the LATE compatibility condition, (11). Comparisons of estimates of the two expressions on the right-hand side of Equation (12) also serve as a test of compatibility.

In practice, of course, covariates need not account fully for the difference between two LATEs. For some covariate values, there may be a good match, while for others, CEI fails. A rationale for partial fulfillment of CEI emerges in our latent-index example. In this context, partial CEI is like saying that some types of women select on gains while others do not. The values for which CEI is satisfied define a subpopulation for which heterogeneous treatment effects can be understood to be solely a function of observable characteristics. For this subpopulation, we can define an average causal effect for which the predictive value of IV estimates is likely to be especially high.

Definition 2 (CATE): The Compatible Average Treatment effect is:

$$\Delta^{z,w} := \int \Delta^{z}(x) dF_{x}(x \mid \Delta^{z}(x) = \Delta^{w}(x)).$$

If the compatibility condition, $\Delta^{z}(x) = \Delta^{w}(x)$, holds for all values of x, CATE is ATE. However, if compatibility holds for only some values of x, CATE is ATE for the subpopulation defined by these values.

The compatible subpopulation may be of interest for a number of reasons. First, we might be looking for the largest and most representative subpopulation for which a given set of IV estimates has predictive value. This might be, for example, a compatible subset of the treated. Second, we may be interested in constructing a precise estimate of covariate-specific treatment effects. CATE difines a subpopulation where it is possible to use two instruments to construct more precise estimates.

4.0 Estimation and Inference

We assume that the effects of interest are to be estimated in a random sample of size n.

Assumption 4 (Sampling): $\{R_i = (Y_i, D_i, X_i, Z_i, W_i), i = 1, \ldots, n\}$ *are i.i.d. observations of the random vector* $R = (Y, D, X, Z, W)$.

We also assume that the covariates, x, take on a finite and fixed number of values. The education and age covariates in the empirical example satisfy this condition. Generalization to continuous covariates seems straightforward but requires additional technical machinery. For example, it seems likely that with continuous covariates, we should allow for a gradual failure of CEI as opposed to discrete cutoffs. We therefore leave this extension for future work.

Assumption 5 (Discrete Covariates): *For a finite set,* $\mathcal{X} = \{x_1, \ldots, x_K\}$, $\mathbb{P}[\mathrm{x} \in \mathcal{X}] = 1$.

The effects in Theorems 2 and 3 can be written as follows:

$$\Delta^{\mathrm{L}}_{\mathrm{s}^{\mathrm{U}}} = \mathbb{E}[\Delta^{\mathrm{L}}(\mathrm{x})\omega^{\mathrm{U}}_{\mathrm{s}}(\mathrm{x})],\ \omega^{\mathrm{U}}_{\mathrm{s}}(x) = \mathbb{P}[\mathrm{s}^{\mathrm{U}} = 1 \mid \mathrm{x} = x]/\mathbb{P}[\mathrm{s}^{\mathrm{U}} = 1]$$

where $\mathrm{L} = \mathrm{U} = \mathrm{z}$ for Theorem 2 and $\mathrm{L} = \mathrm{z}$ and $\mathrm{U} = \mathrm{w}$ or vice versa for Theorem 3. More generally, superscript L indexes the group where conditional LATEs are obtained and s^{U} is an indicator for the group with the covariate distribution of interest, defined using instrument U.

Estimation is straightforward in our finite-dimensional setting. We replace expectations $\mathbb{E}$ and probabilities $\mathbb{P}$ by empirical analogs $\mathbb{E}_n$ and $\mathbb{P}_n$, where $\mathbb{E}_n[g(\mathrm{R})] = n^{-1}\sum_{i=1}^n g(\mathrm{R}_i)$ and $\mathbb{P}_n[g(\mathrm{R}) \in C] = n^{-1}\sum_{i=1}^n 1[g(\mathrm{R}_i) \in C]$ for any function g and set C. For conditional expectations and probabilities, we let $\mathbb{E}_n[\cdot \mid \mathrm{x} = x, \mathrm{U} = u]$ and $\mathbb{P}_n[\cdot \mid \mathrm{x} = x, \mathrm{U} = u]$ denote empirical analogs in the covariate cell, where $\mathrm{x} = x$ and $\mathrm{U} = u$ for $\mathrm{U} \in \{\mathrm{z}, \mathrm{w}\}$ and $u \in \{z, w\}$. This gives:

$$\hat{\Delta}^{\mathrm{L}}_{\mathrm{s}^{\mathrm{U}}} = \mathbb{E}_n[\hat{\Delta}^{\mathrm{L}}(\mathrm{x})\hat{\omega}^{\mathrm{U}}_{\mathrm{s}}(\mathrm{x})],\ \hat{\omega}^{\mathrm{U}}_{\mathrm{s}}(x) = \mathbb{P}_n[\mathrm{s}^{\mathrm{U}} = 1 \mid \mathrm{x} = x]/\mathbb{P}_n[\mathrm{s}^{\mathrm{U}} = 1] \quad (13)$$

where $\hat{\Delta}^{\mathrm{L}}(x)$ is any consistent estimator of $\Delta^{\mathrm{L}}(x)$. For example, $\hat{\Delta}^{\mathrm{L}}(x)$ can be the Wald estimator with instrument $\mathrm{L} \in \{\mathrm{z}, \mathrm{w}\}$ in cell $\mathrm{x} = x$, or 2SLS using both z and w as instruments in cell $\mathrm{x} = x$. For treated, nontreated, and the entire population, the indicator s^{U} is observed, so construction of the empirical $\hat{\omega}^{\mathrm{U}}_{\mathrm{s}}(x)$ is straightforward. For compliers, we can estimate $\omega^{\mathrm{U}}_{\mathrm{s}}(x)$ using the sample analog of Equation (6):

$$\hat{\omega}^{\mathrm{U}}_{\Delta}(x) = \frac{\mathbb{E}_n[\mathrm{D} \mid \mathrm{U} = 1, \mathrm{x} = x] - \mathbb{E}_n[\mathrm{D} \mid \mathrm{U} = 0, \mathrm{x} = x]}{\mathbb{E}_n[\mathrm{D} \mid \mathrm{U} = 1] - \mathbb{E}_n[\mathrm{D} \mid \mathrm{U} = 0]},\ \mathrm{U} \in \{\mathrm{z}, \mathrm{w}\}. \quad (14)$$

Consistency of $\hat{\Delta}^{\mathrm{L}}_{\mathrm{s}^{\mathrm{U}}}$ follows from the law of large numbers and the Slutsky theorem.

Theorem 4 (Consistency): *Let* z *and* w *be two instruments that satisfy Assumptions 2 and 3. Under Assumptions 4 and 5, and* $\mathbb{E}[|\mathrm{y}|] < \infty$*:*

$$\hat{\Delta}^{\mathrm{L}}_{\mathrm{s}^{\mathrm{u}}} = \mathbb{E}_n[\hat{\Delta}^{\mathrm{L}}(\mathrm{x})\hat{\omega}^{\mathrm{u}}_{\mathrm{s}}(\mathrm{x})] \to_p \Delta^{\mathrm{L}}_{\mathrm{s}^{\mathrm{u}}} = \mathbb{E}[\mathrm{y}_1 - \mathrm{y}_0 \mid \mathrm{s}^{\mathrm{w}} = 1];\ \mathrm{l}, \mathrm{u} \in \{\mathrm{z}, \mathrm{w}\}$$

where $\hat{\omega}^{\mathrm{u}}_{\mathrm{s}}(x)$ *and* $\hat{\Delta}^{\mathrm{L}}(x)$ *are any consistent estimators of* $\omega^{\mathrm{u}}_{\mathrm{s}}(x)$ *and* $\Delta^{\mathrm{L}}(x)$*, for all* $x \in \mathcal{X}$*.*

The estimators developed from Theorems 2 and 3 are smooth functions of GMM-type estimators and are therefore asymptotically normal under general conditions. The following result uses the delta method to characterize the relevant limiting distributions. In particular, we show that $(\hat{\Delta}^{\mathrm{z}}_{\mathrm{s}^{\mathrm{u}}}, \hat{\Delta}^{\mathrm{w}}_{\mathrm{s}^{\mathrm{u}}})$ are asymptotically jointly normal, allowing us to draw inferences about the effects of interest and to test some of the implications of CEI. Let $p_k = \mathbb{P}(\mathrm{x} = x_k)$, $\vartheta^{\mathrm{u}}_{\mathrm{s}}(x) = \mathbb{P}[\mathrm{s}^{\mathrm{u}} = 1 \mid \mathrm{x} = x]$ (i.e., the numerator of $\omega^{\mathrm{u}}_{\mathrm{s}}(x)$), and let $\hat{\vartheta}^{\mathrm{u}}_{\mathrm{s}}(x)$ be an estimator of $\vartheta^{\mathrm{u}}_{\mathrm{s}}(x)$.

Theorem 5 (Asymptotic Distribution): *Let* z *and* w *be two instruments that satisfy Assumptions 2 and 3. Assume that* $p_k > \varepsilon > 0$ *for all* $k \in \{1, \ldots, K\}$*, and for* $k \in \{1, \ldots, K\}$ *and* $\mathrm{u} \in \{\mathrm{z}, \mathrm{w}\}$*:*

$$\sqrt{n}\begin{pmatrix} \hat{\vartheta}^{\mathrm{u}}_{\mathrm{s}}(x_k) - \vartheta^{\mathrm{u}}_{\mathrm{s}}(x_k) \\ \hat{\Delta}^{\mathrm{z}}(x_k) - \Delta^{\mathrm{z}}(x_k) \\ \hat{\Delta}^{\mathrm{w}}(x_k) - \Delta^{\mathrm{w}}(x_k) \end{pmatrix}$$

$$\to_d Z^{\mathrm{u}}_k \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} V^{\mathrm{u}}_{\mathrm{s}}(x_k)/p_k & C^{\mathrm{uz}}_{\mathrm{s}\Delta}(x_k)/p_k & C^{\mathrm{uw}}_{\mathrm{s}\Delta}(x_k)/p_k \\ C^{\mathrm{uz}}_{\mathrm{s}\Delta}(x_k)/p_k & V^{\mathrm{z}}_{\Delta}(x_k)/p_k & C^{\mathrm{zw}}_{\Delta\Delta}(x_k)/p_k \\ C^{\mathrm{uw}}_{\mathrm{s}\Delta}(x_k)/p_k & C^{\mathrm{zw}}_{\Delta\Delta}(x_k)/p_k & V^{\mathrm{w}}_{\Delta}(x_k)/p_k \end{bmatrix}\right)$$

where $(Z^{\mathrm{u}}_1, \ldots, Z^{\mathrm{u}}_K)$ *are independent. Under Assumptions 9 and 10, for* $\mathrm{u} \in \{\mathrm{z}, \mathrm{w}\}$*:*

$$\sqrt{n}\begin{pmatrix} \hat{\Delta}^{\mathrm{z}}_{\mathrm{s}^{\mathrm{u}}} - \Delta^{\mathrm{z}}_{\mathrm{s}^{\mathrm{u}}} \\ \hat{\Delta}^{\mathrm{w}}_{\mathrm{s}^{\mathrm{u}}} - \Delta^{\mathrm{w}}_{\mathrm{s}^{\mathrm{u}}} \end{pmatrix} \to_d \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} V^{\mathrm{z}}_{\mathrm{s}^{\mathrm{u}}} & C^{\mathrm{zw}}_{\mathrm{s}^{\mathrm{u}}} \\ C^{\mathrm{zw}}_{\mathrm{s}^{\mathrm{u}}} & V^{\mathrm{w}}_{\mathrm{s}^{\mathrm{u}}} \end{bmatrix}\right)$$

where:

$$V^{\mathrm{L}}_{\mathrm{s}^{\mathrm{u}}} = \sum_{k=1}^{K} p_k[\omega^{\mathrm{u}}_{\mathrm{s}}(x_k)^2 + \tilde{V}^{\mathrm{u}}_{\mathrm{s}}(x_k) + 2\omega^{\mathrm{u}}_{\mathrm{s}}(x_k)\tilde{C}^{\mathrm{uL}}_{\mathrm{s}\Delta}(x_k)][\Delta^{\mathrm{L}}(x_k) - \Delta^{\mathrm{L}}_{\mathrm{s}^{\mathrm{u}}}]^2$$
$$+ \sum_{k=1}^{K} p_k \omega^{\mathrm{u}}_{\mathrm{s}}(x_k)^2 V^{\mathrm{L}}_{\Delta}(x_k),$$

for $\mathrm{L} \in \{\mathrm{z}, \mathrm{w}\}$, $\tilde{V}^{\mathrm{U}}_{\mathrm{s}}(x_k) = V^{\mathrm{U}}_{\mathrm{s}}(x_k)/[\sum_{k=1}^{K} p_k \vartheta^{\mathrm{U}}_{\mathrm{s}}(x_k)]^2$, $\tilde{C}^{\mathrm{UL}}_{\mathrm{s}\Delta}(x_k) = C^{\mathrm{UL}}_{\mathrm{s}\Delta}(x_k)/\sum_{k=1}^{K} p_k \vartheta^{\mathrm{U}}_{\mathrm{s}}(x_k)$, *and*

$$C^{\mathrm{ZW}}_{\mathrm{s}^{\mathrm{U}}} = \sum_{k=1}^{K} p_k[\omega^{\mathrm{U}}_{\mathrm{s}}(x_k)^2 + \tilde{V}^{\mathrm{U}}_{\mathrm{s}}(x_k) + \omega^{\mathrm{U}}_{\mathrm{s}}(x_k)\{\tilde{C}^{\mathrm{UZ}}_{\mathrm{s}\Delta}(x_k) + \tilde{C}^{\mathrm{UW}}_{\mathrm{s}\Delta}(x_k)\}][\Delta^{\mathrm{Z}}(x_k) - \Delta^{\mathrm{Z}}_{\mathrm{s}^{\mathrm{U}}}][\Delta^{\mathrm{W}}(x_k) - \Delta^{\mathrm{W}}_{\mathrm{s}^{\mathrm{U}}}] + \sum_{k=1}^{K} p_k \omega^{\mathrm{U}}_{\mathrm{s}}(x_k)^2 C^{\mathrm{ZW}}_{\Delta\Delta}(x_k).$$

Proof: Let $\hat{p}_k = \mathbb{P}_n(\mathrm{x} = x_k)$. By a standard Central Limit Theorem for multinomial sequences, $\sqrt{n}(\hat{p}_1 - p_1, \ldots, \hat{p}_K - p_K)$ converges in distribution to a multivariate normal with zero mean, variances $p_k(1 - p_k)$ and covariances $-p_k p_j$, for $k, j = 1, \ldots, K$, $k \neq j$. Write, for $\mathrm{L} \in \{\mathrm{z}, \mathrm{w}\}$:

$$\hat{\Delta}^{\mathrm{L}}_{\mathrm{s}^{\mathrm{U}}} = \frac{\sum_{k=1}^{K} \hat{p}_k \hat{\vartheta}^{\mathrm{U}}_{\mathrm{s}}(x_k)\hat{\Delta}^{\mathrm{L}}(x_k)}{\sum_{k=1}^{K} \hat{p}_k \hat{\vartheta}^{\mathrm{U}}_{\mathrm{s}}(x_k)} \text{ and } \Delta^{\mathrm{L}}_{\mathrm{s}^{\mathrm{U}}} = \frac{\sum_{k=1}^{K} p_k \vartheta^{\mathrm{U}}_{\mathrm{s}}(x_k)\Delta^{\mathrm{L}}(x_k)}{\sum_{k=1}^{K} p_k \vartheta^{\mathrm{U}}_{\mathrm{s}}(x_k)}.$$

Let $\vec{\pi} = (\pi_1, \ldots, \pi_K)$, $\vec{\upsilon} = (\upsilon_1, \ldots, \upsilon_K)$, and $\vec{\delta} = (\delta_1, \ldots, \delta_K)$. If $\sum_k \pi_k \upsilon_k \neq 0$, the function $f(\vec{\pi}, \vec{\upsilon}, \vec{\delta}) = \sum_k \pi_k \upsilon_k \delta_k / \sum_k \pi_k \upsilon_k$ is continuously differentiable in $(\vec{\pi}, \vec{\upsilon}, \vec{\delta})$ with partial derivatives:

$$\frac{\partial f(\vec{\pi}, \vec{\upsilon}, \vec{\delta})}{\partial \pi_k} = \tilde{\upsilon}_k \tilde{\delta}_k, \frac{\partial f(\vec{\pi}, \vec{\upsilon}, \vec{\delta})}{\partial \upsilon_k} = \frac{\pi_k \tilde{\delta}_k}{\sum_k \pi_k \upsilon_k}, \frac{\partial f(\vec{\pi}, \vec{\upsilon}, \vec{\delta})}{\partial \delta_k} = \pi_k \tilde{\upsilon}_k$$

for $\tilde{\upsilon}_k = \upsilon_k / \sum_k \pi_k \upsilon_k$ and $\tilde{\delta}_k = \delta_k - \sum_k \pi_k \tilde{\upsilon}_k \delta_k$.

Set $\pi_k = p_k$, $\upsilon_k = \vartheta^{\mathrm{U}}_{\mathrm{s}}(x_k)$, and $\delta_k = \Delta^{\mathrm{L}}(x\vartheta_k)$. The result then follows using the delta method. ■

The joint normality assumption for the components of our reweighting estimators holds under standard regularity conditions. In particular, this follows for the estimators of the weighting function by the Central Limit Theorem for binary sequences. For IV, GMM, and other moment-based estimators of the conditional LATEs, such as generalized empirical likelihood, existence of second moments (i.e., $\mathbb{E}[\mathrm{Y}^2] < \infty$) is sufficient. The first term in the expressions for $V^{\mathrm{Z}}_{\mathrm{s}^{\mathrm{U}}}$, $V^{\mathrm{W}}_{\mathrm{s}^{\mathrm{U}}}$, and $C^{\mathrm{ZW}}_{\mathrm{s}^{\mathrm{U}}}$ reflects sampling variation due to the estimation of the covariate-cell probabilities, p_k, and the weighting functions, whereas the second term arises from the estimation of conditional LATEs. The first term is zero if, for example, the conditional LATEs are constant (i.e., $\Delta^{\mathrm{L}}(x) = \Delta^{\mathrm{L}}_{\mathrm{s}^{\mathrm{U}}}$ for all $x \in \mathcal{X}$).

In practice, there are two routes to asymptotic inference. We can estimate asymptotic distributions analytically using sample analogs or approximate them numerically by resampling or simulation. We use bootstrap methods

in the empirical application. Resampling methods are convenient and save us from having to estimate complicated analytical formulas for asymptotic variances and covariances. Consistency of a bootstrap approximation to the distributions of our reweighting estimators follows from Hall and Horowitz (1996), Hahn (1996), and Brown and Newey (2002) theorems for GMM, and application of the delta method for the bootstrap (see, e.g., Theorem 23.5 in van der Vaart 1998).

There are many ways to bootstrap. We use the empirical likelihood (EL) bootstrap proposed by Brown and Newey (2002) for GMM estimators. This method resamples from the empirical likelihood distribution (ELD) that imposes the moment conditions in the sample, instead of from the empirical distribution (ED). Let $\mathbb{E}[g(\mathrm{R}, \theta)] = 0$ be the moment conditions that define the conditional LATEs, weighting functions, and effects of interest, where θ includes all unknown parameters. The ELD $(\hat{\pi}_1, \ldots, \hat{\pi}_n)$ is the solution to:

$$\max_{\pi_1,\ldots,\pi_n} \sum_{i=1}^{n} \ln(\pi_i), \text{ s.t. } \sum_{i=1}^{n} \pi_i g(\mathrm{R}_i, \hat{\theta}) = 0, \sum_{i=1}^{n} \pi_i = 1, \pi_i \geq 0$$

where $\hat{\theta}$ is the EL or another consistent estimator of θ.[8] ELD therefore is the closest to ED in terms of Kullback–Leibler distance. ELD and ED are equal in exactly identified models, but they generally differ under overidentification. In practice, the difference between the EL bootstrap and the standard nonparametric bootstrap is that the former resamples from the data with probabilities $\hat{\pi}_i$ instead of $1/n$.

Consistent estimation of CATE requires that we condition on the unobservable events $\{\Delta^{\mathrm{Z}}(x) = \Delta^{\mathrm{W}}(x)\}$. In finite samples, we never have $\hat{\Delta}^{\mathrm{Z}}(x) = \hat{\Delta}^{\mathrm{W}}(x)$. We therefore use cell-by-cell overidentification tests to find compatible values of x. Instead of discarding cells that fail the identification test for some small significance level, we reweight estimates of conditional LATE by a decreasing function of the overidentification test statistic. Letting $J(x)$ denote the overidentification test statistic for the instruments z and w in the cell $\mathrm{x} = x$, the resulting estimator of CATE is:

$$\begin{aligned} \hat{\Delta}^{\mathrm{z,w}} &= \mathbb{E}_n[\hat{\Delta}^{\mathrm{z,w}}(\mathrm{x})\hat{\omega}_{CATE}(\mathrm{x})], \\ \hat{\omega}_{CATE}(x) &= \exp\{-J(x)/a_n(x)\}/\mathbb{E}_n[\exp\{-J(\mathrm{x})/a_n(\mathrm{x})\}] \end{aligned} \tag{15}$$

where $\hat{\Delta}^{\mathrm{z,w}}(x)$ is the GMM estimate in cell $\mathrm{x} = x$ that uses z and w as instruments or any other moment estimator (e.g., 2SLS), and $a_n(x)$ is a sequence such that $a_n(x) \to \infty$ and $a_n(x) = o(n)$, for $x \in \mathcal{X}$.

[8] In the empirical application, we use the EL estimator of θ to obtain the ELD.

The sequences $a_n(x)$ guarantee the consistency of the reweighting estimator (15) for CATE. These sequences play a role similar to the penalty terms used in Andrews (1999) to obtain consistent model-selection procedures for GMM estimators. To formally establish consistency, it is convenient to introduce additional notation. Let $\mathcal{X}_0 = \{x \in \mathcal{X} : \Delta^{\mathrm{z}}(x) = \Delta^{\mathrm{w}}(x)\}$ denote the set of covariate values that satisfy the compatibility condition, with complement $\bar{\mathcal{X}}_0 = \{x \in \mathcal{X} : \Delta^{\mathrm{z}}(x) \neq \Delta^{\mathrm{w}}(x)\}$.

Theorem 6 (CATE Consistency): *Let* z *and* w *be two instruments that satisfy Assumption 2. Let* $a_n(x)$ *be sequences such that* $a_n(x) \to \infty$ *and* $a_n(x) = o(n)$ *for all* $x \in \mathcal{X}$. *Assume that* $\hat{\Delta}^{\mathrm{z,w}}(x) \to_p \Delta^{\mathrm{z}}(x)$ *and* $J(x) = O_p(1)$ *for all* $x \in \mathcal{X}_0$, *and* $\hat{\Delta}^{\mathrm{z,w}}(x) = O_p(1)$ *and* $J(x) = O_p(n)$ *for all* $x \in \bar{\mathcal{X}}_0$. *Under Assumptions 4 and 5,* $\mathbb{P}\{\mathrm{x} \in \mathcal{X}_0\} > 0$, *and* $\mathbb{E}[|\mathrm{y}|] < \infty$:

$$\hat{\Delta}^{\mathrm{z,w}} = \mathbb{E}_n[\hat{\Delta}^{\mathrm{z,w}}(\mathrm{x})\hat{\omega}_{CATE}(\mathrm{x})] \to_p \Delta^{\mathrm{z,w}} = \mathbb{E}[\Delta^{\mathrm{z}}(\mathrm{x}) \mid \mathrm{x} \in \mathcal{X}_0]$$

where $\hat{\omega}_{CATE}(x) = \exp\{-J(x)/a_n(x)\}/\mathbb{E}_n[\exp\{-J(\mathrm{x})/a_n(\mathrm{x})\}]$.

Proof: We write:

$$\hat{\Delta}^{\mathrm{z,w}} = \sum_{k=1}^{K} \hat{p}_k \hat{\Delta}^{\mathrm{z,w}}(x_k)\hat{\omega}_{CATE}(x_k) \text{ and}$$

$$\Delta^{\mathrm{z,w}} = \sum_{k=1}^{K} p_k \Delta^{\mathrm{z}}(x_k) 1\{x_k \in \mathcal{X}_0\} \Big/ \sum_{k=1}^{K} p_k 1\{x_k \in \mathcal{X}_0\}.$$

By the Law of Large Numbers, $\hat{p}_k \to_p p_k$. For $x_k \in \mathcal{X}_0$, $\hat{\Delta}^{\mathrm{z,w}}(x_k) \to_p \Delta^{\mathrm{z}}(x_k)$, $J(x_k) = O_p(1)$ and $\exp\{-J(x_k)/a_n(x_k)\} \to_p 1$. For $x_k \in \bar{\mathcal{X}}_0$, $\hat{\Delta}^{\mathrm{z,w}}(x_k) = O_p(1)$, $J(x_k) = O_p(n)$ and $\exp\{-J(x_k)/a_n(x_k)\} \to_p 0$. Hence, $\hat{\Delta}^{\mathrm{z,w}}(x_k)\exp\{-J(x_k)/a_n(x_k)\} \to_p \Delta^{\mathrm{z}}(x)1\{x_k \in \mathcal{X}_0\}$.

The result follows by the Slutsky Theorem, noting that:

$$\begin{aligned}\mathbb{E}_n[\exp\{-J(\mathrm{x})/a_n(\mathrm{x})\}] &= \sum_{k=1}^{K} \hat{p}_k \exp\{-J(x_k)/a_n(x_k)\} \\ &\to_p \sum_{k=1}^{K} p_k 1\{x_k \in \mathcal{X}_0\} = \mathbb{P}\{\mathrm{x} \in \mathcal{X}_0\} > 0.\end{aligned}$$

■

For both $x \in \mathcal{X}_0$ and $x \in \bar{\mathcal{X}}_0$, the convergence-rate assumptions for $\hat{\Delta}^{\mathrm{z,w}}(x)$ and $J(x)$ in the statement of the theorem are satisfied by GMM-type

estimators under standard regularity conditions for asymptotic normality (see, e.g., Assumption 1 in Andrews 1999).

Although consistency of our CATE estimator is relatively easy to show, inference is challenging. As with the reweighting estimators in Theorem 5, the limiting distribution of $\hat{\Delta}^{z,w}$ depends on the limiting distributions of the weighting functions. Here, however, the limiting distribution of $\hat{\omega}_{CATE}(x)$ converges at nonstandard rates, complicating the analysis. Slow convergence in this case is a by-product of the need for a term like $a_n(x)$ in the weighting function to ensure consistency and the fact that, in practice, we choose this to grow more slowly than $\sqrt{n}$.[9] Moreover, convergence to the limiting distribution cannot be uniform in the data-generating process because the CATE estimator implicitly conditions on a pretest (see, e.g., Leeb and Potscher 2008). In a related setting, Andrews and Guggenberger (2009) addressed a pretest problem using subsampling. Subsampling is imprecise in our application due to small cell sizes. We have no easy solution in this case other than to caution that convergence to the relevant limiting distribution may be slow and to conjecture the pointwise validity of bootstrap methods.

A second and less serious inference complication arises from the fact that the EL bootstrap imposes CEI at all covariate values, while the purpose of CATE is to allow deviations from CEI. In the empirical application, therefore, we supplement the standard errors obtained from the EL bootstrap with standard errors obtained by a nonparametric bootstrap that does not impose the compatibility condition on the bootstrap data generating process.

5.0 Results

Our empirical exploration of covariate-reweighting focuses on a categorical representation of second-child age and mother's schooling. As shown in Table 2, these covariates are strongly related to compliance probabilities. A younger second child in the household reduces the likelihood of a third birth, if only because less time has passed since the birth of the second. Education matters because college-educated women are less likely to choose to have a third child than less-educated women. Twins compliers therefore are relatively more likely to have a young second-born and to be highly educated. Sex-composition compliers, by contrast, are relatively unlikely

[9] As in Crump, Hotz, Imbens, and Mitnik (2009), we could simplify here by doing inference conditional on the sample. We do not take this route because we are interested in predictive population inference as opposed to sample-specific causal inference.

to be college graduates or to have a third child soon after the birth of the second. Labor supply effects also are likely to vary with second-child age and mother's schooling. The birth of a third child has little effect on the work behavior of a woman with a young second-born who is at home anyway. Likewise, a relatively educated woman should be affected less by the birth of a child than other women because for women earning higher wages, it makes sense to pay for child care in the market. Differences in twins and same-sex complier subpopulations therefore might account for the fact that twins instruments generate a smaller effect of childbearing than sex-composition instruments.

In an effort to see whether this conjecture is substantiated empirically, Table 3 reports IV estimates in each of 12 cells defined by second-child age and mother's schooling. The age categories are less than or equal to 4 years, greater than 4 and less than or equal to 8, and greater than 8. The schooling categories are high school dropout, high school graduate, some college, and college graduate. The first column of the table reports the probability mass function in the contingency table generated by these categories. The next two columns describe the distribution of covariates for the treated and untreated, relative to the entire population. Women who do and do not have a third child are clearly very different.

Cell-specific IV estimates are fairly imprecise, as shown in Columns 4 and 6 of Table 3. From these noisy cell-by-cell estimates alone, it is difficult to see how the causal effect of a third birth varies with individual characteristics. A clearer pattern emerges, however, once the cells are "weighted-up," a point we return to in Table 4.

Table 3 also reports estimates of:

$$\omega_{\mathrm{s}}^{\mathrm{z}}(x) = \frac{\mathbb{P}[\mathrm{s}^{\mathrm{z}} = 1 \mid \mathrm{x} = x]}{\mathbb{P}[\mathrm{s}^{\mathrm{z}} = 1]}$$

the weighting function for twins compliers (Column 5) and same-sex compliers (Column 7). This is the ratio of the relevant first stage in the cell to the overall first stage of Equation (6). The distribution of twins and same-sex compliers over cells clearly is different from the cell distribution in the random sample. As suggested by the summary comparisons in Table 2, the complier distributions for the two instruments also are very different from one another. Specifically, twins compliers are much more likely to have a young second-born child, while few same-sex compliers are in this group. Twins compliers also are relatively educated, while the schooling gradient in compliance probabilities for same-sex is less pronounced.

Table 3. *LATE Decompositions*

Covariate		Covariate pmf			IV Estimates and Weighting Functions						
					Twins Instrument		Same-Sex Instrument		Both Instruments		
Age	Education	$P(X)$ (1)	$P(X\|D=1)/P(X)$ (2)	$P(X\|D=0)/P(X)$ (3)	$\Delta^z(X)$ (4)	$\omega^z_\Delta(X)$ (5)	$\Delta^w(X)$ (6)	$\omega^w_\Delta(X)$ (7)	$\Delta^{z,w}(X)$ (8)	J-p value (9)	$\omega_{CATE}(X)$ (10)
					A. Weeks worked						
[0,4]	HS drop	0.06	0.65	1.24	−4.63	1.23	−1.26	0.77	−4.35	0.55	1.45
					(1.84)		(5.49)		(1.60)		
	HS grad	0.15	0.46	1.36	−4.24	1.36	−2.66	0.59	−4.17	0.76	1.99
					(1.15)		(5.00)		(1.04)		
	Some col	0.06	0.46	1.36	−3.79	1.36	−4.93	0.66	−3.86	0.87	2.14
					(1.69)		(6.86)		(1.59)		
	Col grad	0.05	0.41	1.40	−5.36	1.40	−2.21	0.54	−5.24	0.75	1.95
					(1.99)		(9.77)		(1.84)		
(4.8]	HS drop	0.07	1.29	0.81	−4.43	0.81	−9.35	1.20	−6.29	0.25	0.47
					(2.65)		(3.36)		(2.06)		
	HS grad	0.17	0.93	1.05	−3.07	1.05	−5.59	1.38	−3.94	0.33	0.79
					(1.52)		(2.09)		(1.23)		
	Some col	0.07	0.91	1.06	−1.02	1.06	−7.39	1.13	−2.59	0.16	0.21
					(2.22)		(3.91)		(1.93)		
	Col grad	0.04	0.89	1.08	−1.52	1.08	−2.88	0.94	−1.72	0.83	2.09
					(2.63)		(5.97)		(2.32)		
(8+]	HS drop	0.10	1.79	0.47	0.29	0.47	−12.04	0.80	−5.53	0.06	0.04
					(4.47)		(4.74)		(3.24)		
	HS grad	0.17	1.40	0.73	−2.41	0.73	−9.76	1.30	−6.15	0.02	0.01
					(2.22)		(2.25)		(1.60)		
	Some col	0.06	1.33	0.78	−4.40	0.78	−4.72	1.10	−4.52	0.96	2.20
					(3.55)		(4.54)		(2.82)		
	Col grad	0.02	1.15	0.90	6.78	0.90	17.90	1.12	9.48	0.28	0.44
					(4.99)		(9.01)		(4.17)		

B. Employment

[0,4]	HS drop	0.06	0.65	1.24	−0.154	1.23	−0.035	0.77	−0.143	0.43	1.15
					(0.048)		(0.143)		(0.043)		
	HS grad	0.15	0.46	1.36	−0.081	1.36	−0.123	0.59	−0.083	0.73	2.14
					(0.027)		(0.117)		(0.026)		
	Some col	0.06	0.46	1.36	−0.089	1.36	−0.035	0.66	−0.086	0.74	2.15
					(0.038)		(0.157)		(0.037)		
	Col grad	0.05	0.41	1.40	−0.130	1.40	−0.023	0.54	−0.125	0.65	1.90
					(0.047)		(0.231)		(0.046)		
(4.8]	HS drop	0.07	1.29	0.81	−0.140	0.81	−0.150	1.20	−0.144	0.92	2.42
					(0.064)		(0.082)		(0.051)		
	HS grad	0.17	0.93	1.05	−0.081	1.05	−0.156	1.38	−0.108	0.19	0.39
					(0.034)		(0.046)		(0.028)		
	Some col	0.07	0.91	1.06	−0.034	1.06	−0.161	1.13	−0.066	0.19	0.34
					(0.048)		(0.084)		(0.042)		
	Col grad	0.04	0.89	1.08	0.075	1.08	−0.202	0.94	0.028	0.06	0.03
					(0.061)		(0.134)		(0.054)		
(8+]	HS drop	0.10	1.79	0.47	0.047	0.47	−0.188	0.80	−0.064	0.11	0.13
					(0.101)		(0.106)		(0.073)		
	HS grad	0.17	1.40	0.73	−0.066	0.73	−0.167	1.30	−0.117	0.13	0.20
					(0.046)		(0.047)		(0.033)		
	Some col	0.06	1.33	0.78	−0.110	0.78	−0.023	1.10	−0.075	0.47	1.32
					(0.071)		(0.092)		(0.058)		
	Col grad	0.02	1.15	0.90	0.034	0.90	0.224	1.12	0.082	0.33	0.68
					(0.098)		(0.171)		(0.081)		

Note: Standard errors for estimates in parentheses. The p-value for the joint J-statistic for all covariate values is 0.25 for weeks and 0.29 for LFP. The sample size is 394,840.

Table 4. *Reweighting LATE*

Population Represented (effect)	Instrument Used for $\Delta(X)$	Weighting Function $\omega(X)$	Weeks Worked		Employment	
			Estimate (1)	|t| for Diff (2)	Estimate (3)	|t| for Diff (4)
Twins compliers	twins	twins	−3.15		−0.075	
(LATE)			(0.62)	1.32	(0.014)	0.93
	same-sex		−2.71		−0.068	
			(0.81)		(0.018)	
Same-sex compliers	same-sex	same-sex	−6.30		−0.131	
(LATE)			(1.15)	1.44	(0.026)	0.77
	twins		−5.08		−0.115	
			(1.58)		(0.037)	
Everyone	twins	1	−2.84		−0.067	
(ATE)			(0.76)	1.99	(0.017)	1.58
	same-sex		−5.88		−0.123	
			(1.35)		(0.031)	
Treated	twins	$P(X\|D=1)/P(X)$	−2.38		−0.056	
(TOT)			(1.07)	2.85	(0.024)	2.14
	same-sex		−7.08		−0.136	
			(1.28)		(0.029)	
Nontreated	twins	$P(X\|D=0)/P(X)$	−3.15		−0.075	
(TNT)			(0.62)	1.15	(0.014)	1.02
	same-sex		−5.08		−0.115	
			(1.58)		(0.037)	
Compatible	twins	$\exp[-12^* J(X)/n(X)]$	−3.80		−0.099	
(CATE)			(0.80)	0.18	(0.018)	0.17
			[0.64]	[0.09]	[0.015]	[0.08]
	same-sex		−3.66		−0.095	
			(1.01)		(0.023)	
			[0.96]		[0.021]	
	twins, same-sex		−4.00		−0.101	
			(0.77)		(0.017)	
			[0.62]		[0.014]	

Note: Standard errors for estimates in parentheses. *T*-statistics are for the difference between same-sex and twins estimates. Standard errors and t-statistics obtained by Brown and Newey (2002) GMM bootstrap with 1,000 repetitions. Standard errors and t-statistics reported in brackets were obtained by nonparametric bootstrap with 1,000 repetitions. The sample size is 394,840.

Table 4 reports LATE-reweighted estimates combining conditional LATEs and weighting functions for twins compliers and same-sex compliers. The first and fifth row weight conditional LATE estimates to produce an overall unconditional effect using the sample analog of Equation (4). For example, in the first row, cell-by-cell twins estimates are weighted across cells using the probability of twins compliance in each cell. This is close to the Wald estimate using twins instruments in Table 1; compare −3.15 to −3.28 and −0.075 to −0.076. Likewise, the marginalized same-sex estimates using the probability of same-sex compliance are −6.30 for weeks worked and −0.132 for employment, close to the corresponding Wald estimates. The fact that weighted conditional LATE comes out close to Wald estimates that ignore covariates suggests that covariates are unnecessary for identification of LATE, although they are helpful for extrapolation.

Reweighting conditional same-sex estimates using twins weights brings these estimates remarkably close to the marginalized twins estimates. Compare, for example, the estimate of the effect on weeks worked of −3.15 using twins in each cell to the estimate of −2.71 using same-sex instruments in each cell. Reweighting conditional twins and same-sex estimates using same-sex weights also produces a good match. In this case, the estimate of −6.30 using same-sex instruments in each cell can be compared to an estimate of −5.08 using twins instruments in each cell in the fifth and seventh rows of Table 4. In other words, differing LATE estimates generated using twins and same-sex instruments can be reconciled by reweighting covariate-specific estimates using a set of common compliance weights.

This reconciliation is an encouraging finding which suggests that external validity is an attainable goal in this context. On the other hand, Table 4 also shows that ATE and TOT using twins and same-sex instruments are not well matched. This is disappointing because, by the same argument that reconciles twins and same-sex estimates of LATE, we should be able to generate similar estimates of ATE and TOT using either instrument to construct cell-specific IV estimates. At the same time, the match for TNT is not as bad as that for ATE and TOT. The fact that some parameters can be matched more easily across instruments than others suggests that a few poorly matched cells are what drives the cross-instrument imbalance in estimates of ATE and TOT – that is, a handful of cells generate estimates that depend on the instrument.

CATE, shown at the bottom of Table 4, solves this problem, and produces a good match for the average treatment effect in compatible cells by down-weighting cells where CEI is most at odds with the data. CATE estimates

are −3.80 to −3.66 for effects on weeks worked and −0.099 to −0.095 for effects on employment. The estimates of CATE reported in Table 4 set $a_n(x) = \log n(x)/K$, where $n(x)$ is the number of individuals with $\mathrm{x} = x$ and K is the number of covariate cells. This relatively slow normalization works well in our application because the cell-level overidentification test is never very large. Still, by applying the most weight to cells in which CEI appears to be satisfied, CATE generates remarkably similar estimates of the population ATE, whether the underlying cell-level IV estimates use twins or same-sex instruments.

The match generated by CATE is for a subpopulation that need not be representative of the entire population of interest. Figure 1 describes the compatible subpopulation by plotting the weighting functions used by different estimators. CATE essentially discards the two low-education cells for women with an older second-born, with weights something like the histogram for twins compliers. Because twins instruments induce one-sided noncompliance, the group of twins compliers is the same as the nontreated population (Angrist, Lavy, and Schlosser 2010). Thus, the compatible subpopulation is similar to the population of nontreated women, although not exactly the same (Table 3 shows that reweighting to estimate effects on TNT is not very successful). Still, like the nontreated population, the subpopulation for which IV estimates of the effects of childbearing appear to have strong predictive value consists mainly of mothers who have a young second-born child and are somewhat more likely to have gone to college.

6.0 Summary and Directions for Further Work

In the LATE framework, differences in IV estimates need not signal a failure of the exclusion restriction. Rather, these differences may be attributable to differences in the type of people who are affected by the underlying experiments implicit in any IV identification strategy. At the same time, we often hope to use one set of IV estimates to predict causal effects in settings other than the one generating the estimates; the question of external validity turns on our ability to do this reliably.

Here, we begin with the idea that differences in IV estimates of LATE for the same causal relation might be driven by a combination of treatment-effect heterogeneity across covariate cells and differences in covariate distributions for instrument-specific–compliant subpopulations. Limiting heterogeneity across cells and instruments to be a function solely of observed characteristics, we can reweight one set of IV estimates to generate

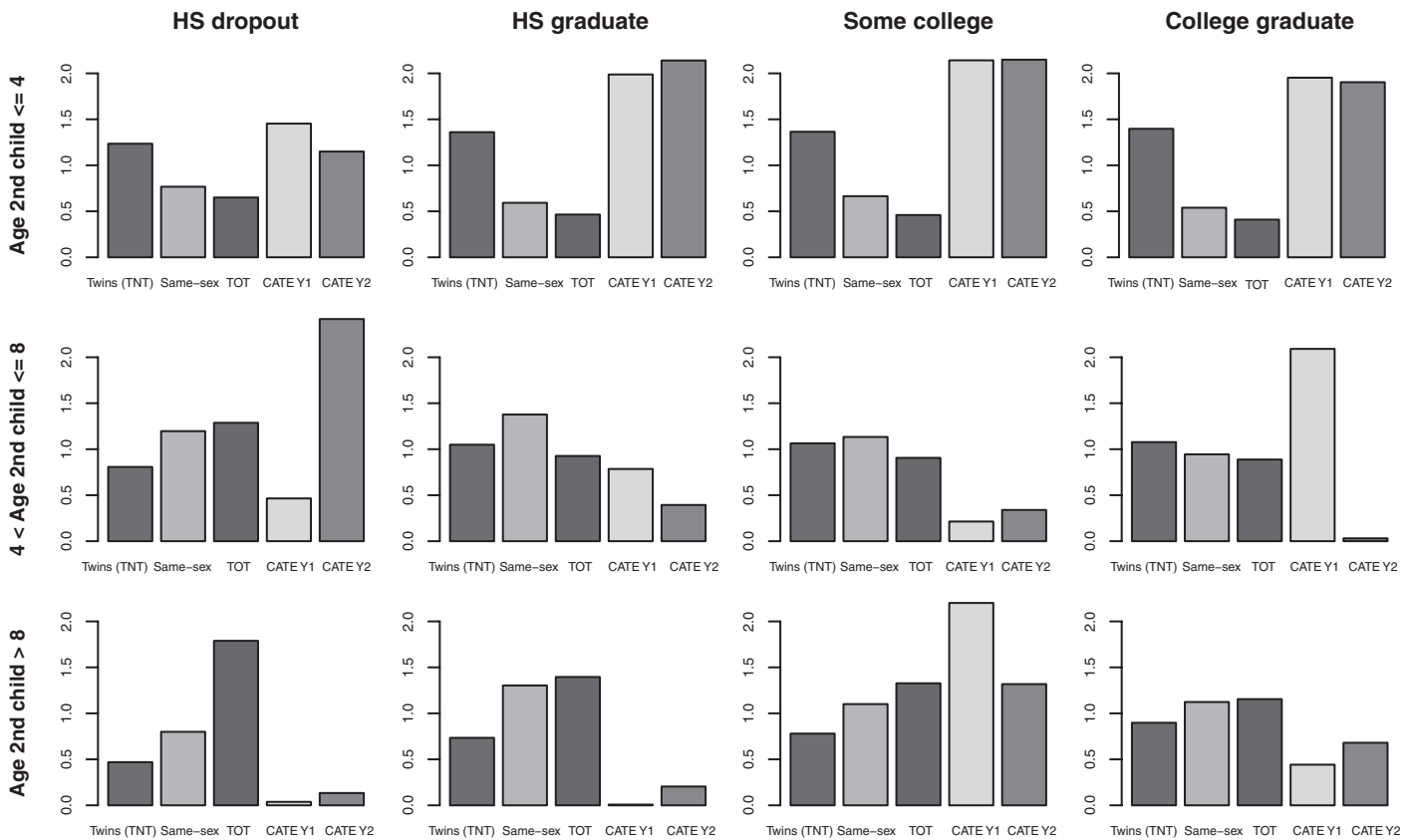

Figure 1. The distribution of compliers and related subpopulations across covariate cells: Y1 = weeks worked, Y2 = employment.

effects for compliant subpopulations other than the one defined by the instrument at hand. This approach turns out to do a good job of explaining why twins instruments produce smaller estimates of the labor supply consequences of childbearing than sex-composition instruments in the Angrist and Evans (1998) dataset.

The CEI assumption that lies at the heart of our approach rules out Roy (1951)–type selection into treatment on the basis of outcome gains. "No Roy selection" is unlikely to be compelling in many settings; gain-driven selection motivates a wide range of theoretical discussions of causal effects in labor economics and other applied fields (see, e.g., Rosen and Willis 1979 for a Roy model of schooling). At the same time, it seems hard to argue with the idea that any analysis of treatment-effect heterogeneity should *begin* with effect variation that is associated with the characteristics we observe.

In an effort to bridge the gap between heterogeneity associated with observed characteristics and latent gains, we also have explored an approach that allows some covariate values to satisfy our CEI assumption, while others – perhaps only a few – do not. This idea seems to work well in our application, generating, for example, remarkably similar estimates of population ATE using twins and same-sex instruments when the sample is reweighted toward cells that appear to satisfy CEI. At the same time, we acknowledge that brazen pretesting induces a complicated limiting distribution that we have not yet succeeded in characterizing and that may not always be useful for applied work. The development of robust and convenient inference procedures for CATE-type estimators seems a natural direction for further work on the external validity of IV estimates.

APPENDIX A: CEI IN AN ILLUSTRATIVE LDV MODEL

Because the text illustration of CEI uses a linear additive structure for potential outcomes, this example shows how we might have CEI in a nonadditive model with nonrandom selection into treatment and a Bernoulli outcome. To simplify, we drop covariates.

Suppose that treatment is determined by:

$$\mathrm{D} = 1[h_0(1-\mathrm{Z}) + h_1\mathrm{Z} > \eta],\ \eta \mid \mathrm{Z} \sim U(0,1/2),\ 0 \le h_0 < h_1 \le 1/2$$

so that potential assignments are:

$$\mathrm{D}_0^{\mathrm{Z}} = 1[h_0 > \eta] \text{ and } \mathrm{D}_1^{\mathrm{Z}} = 1[h_1 > \eta].$$

The associated model for potential outcomes also has the latent-index representation:

$$\mathrm{Y}_j = 1[\eta + g_j > \mathrm{U}_j],\ \mathrm{U}_j \mid \eta, \mathrm{Z} \sim U(0,1),\ 0 \leq g_j \leq 1/2,\ \ j \in \{0,1\}.$$

Iterating expectations, we can show that $\mathbb{E}[\mathrm{Y}_0] = g_0 + 1/4$ and

$$\mathbb{E}[\mathrm{Y}_0 \mid \mathrm{D} = 1] = g_0 + \frac{1}{2}\frac{h_0^2 + (h_1^2 - h_0^2)\mathbb{P}[\mathrm{Z} = 1]}{h_0 + (h_1 - h_0)\mathbb{P}[\mathrm{Z} = 1]}.$$

Selection bias arises because the conditional mean of Y_0 is a function of D. However, CEI holds because:

$$\Delta^{\mathrm{Z}} = \mathbb{E}[\mathrm{Y}_1 - \mathrm{Y}_0 \mid \mathrm{D}_1^{\mathrm{Z}}, \mathrm{D}_0^{\mathrm{Z}}] = \mathbb{E}[\mathbb{E}[\mathrm{Y}_1 - \mathrm{Y}_0 \mid \mathrm{D}_1^{\mathrm{Z}}, \mathrm{D}_0^{\mathrm{Z}}, \eta, \mathrm{Z}] \mid \mathrm{D}_1^{\mathrm{Z}}, \mathrm{D}_0^{\mathrm{Z}}] = g_1 - g_0.$$

References

Abadie, A. (2003), "Semiparametric Instrumental Variable Estimation of Treatment Response Models." *Journal of Econometrics*, 113(2), 231–63.

Andrews, D. W. K. (1999), "Estimation When a Parameter Is on a Boundary." *Econometrica*, 67(6), 1341–83.

Andrews, D. W. K., and P. Guggenberger (2009), "Hybrid and Size-Corrected Subsampling Methods," *Econometrica*, 77(3), 721–62.

Angrist, J. D. (2004), "Treatment Effect Heterogeneity in Theory and Practice," *Economic Journal*, 114(494), C52–C83.

Angrist, J. D., and W. N. Evans (1998), "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *American Economic Review*, 88(3), 450–77.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–72.

Angrist, J. D., V. Lavy, and A. Schlosser (2010), "Multiple Experiments for the Causal Link between the Quantity and Quality of Children," *Journal of Labor Economics*, 28, 773–824.

Aronow, P., and A. J. Sovey (2010), "Beyond LATE: A Simple Method for Recovering Sample Average Treatment Effects," Yale University, discussion paper.

Brown. B. W., and W. K. Newey (2002), "Generalized Method of Moments, Efficient Bootstrapping, and Improved Inference," *Journal of Business & Economic Statistics*, 20(4), 507–17.

Chamberlain, G. (2010), "Bayesian Aspects of Treatment Choice," Harvard University, discussion paper.

Cruces, G. (2005), "Poverty, Income Fluctuations and Work: Argentina 1991–2002," London School of Economics and Political Science, Ph.D. dissertation.

Cruces, G., and S. Galiani (2007), "Fertility and Female Labor Supply in Latin America: New Causal Evidence," *Labour Economics*, 14, 565–73.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009), "Dealing with Limited Overlap in Estimation of Average Treatment Effects," *Biometrika*, 96(1), 187–99.

Ebenstein, A. (2009), "When Is the Local Average Treatment Close to the Average? Evidence from Fertility and Labor Supply," *Journal of Human Resources*, 44(4), 955–75.

Frangakis, C. E., and D. B. Rubin (2002), "Principal Stratification in Causal Inference," *Biometrics*, 58(1), 21–9.

Frölich, M. (2007), "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates," *Journal of Econometrics*, 139(1), 35–75.

Gelbach, J. B. (2002), "Public Schooling for Young Children and Maternal Labor Supply," *American Economic Review*, 92(1), 307–22.

Hahn, J. (1996), "A Note on Bootstrapping Generalized Method of Moments Estimators," *Econometric Theory*, 12, 187–97.

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66(2), 315–32.

Hall, P., and J. L. Horowitz (1996), "Bootstrap Critical Values for Tests Based on Generalized-Method-of-Moments Estimators," *Econometrica*, 64(4), 891–916.

Heckman, J. J. (2010), "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy," *Journal of Economic Literature*, 48(2), 356–98.

Heckman, J. J., J. L. Tobias, and E. Vytlacil (2001), "Four Parameters of Interest in the Evaluation of Social Programs," *Southern Economic Journal*, 68(2), 210–23.

Heckman, J. J., J. L. Tobias, and E. Vytlacil (2003), "Simple Estimators for Treatment Parameters in a Latent-Variable Framework," *Review of Economics and Statistics*, 85(3), 748–55.

Ichino, A., and R. Winter-Ebmer (1999), "Lower and Upper Bounds of Returns to Schooling: An Exercise in IV Estimation with Different Instruments," *European Economic Review*, 43, 889–901.

Imbens, G. W., and J. D. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–75.

Imbens, G. W., and W. K. Newey (2009), "Identification and Estimation of Triangular Simultaneous Equations Models without Additivity," *Econometrica*, 77(5), 1481–512.

Leeb, H., and M. M. Potscher (2008), "Model Selection," in T. G. Andersen, R. A. Davis, J.-P. Kreiss, and T. Mikosch (eds.), *Handbook of Financial Time Series*, Berlin Heidlberg, Springer-Verlag.

Olsen, R. J. (1980), "A Least Squares Correction for Selectivity Bias," *Econometrica*, 48(7), 1815–20.

Oreopoulos, P. (2006), "Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter," *American Economic Review*, 96(1), 152–75.

Rosenzweig, M. R., and K. I. Wolpin (1980), "Testing the Quantity–Quality Fertility Model: The Use of Twins as a Natural Experiment," *Econometrica*, 48(1), 227–40.

Roy, A. D. (1951), "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 3(2), 135–46.

Rubin, D. B. (1977), "Assignment to a Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics*, 2(1), 1–26.

Van der Vaart, A. W. (1998), *Asymptotic Statistics*, New York: Cambridge University Press.

Vella, F., and M. Verbeek (1999), "Estimating and Interpreting Models with Endogenous Treatment Effects," *Journal of Business & Economic Statistics*, 17(4), 473–78.
Vytlacil, E. (2002), "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70(1), 331–41.
Wald, A. (1940), "The Fitting of Straight Lines if Both Variables Are Subject to Error," *Annals of Mathematical Statistics*, 11(3), 284–300.
Willis, R. J., and S. Rosen (1979), "Education and Self-Selection," *The Journal of Political Economy*, 87(5), S7–S36.