NBER WORKING PAPER SERIES

EXTRAPOLATE-ING:
EXTERNAL VALIDITY AND OVERIDENTIFICATION IN THE LATE FRAMEWORK

Joshua Angrist
Ivan Fernandez-Val

Working Paper 16566
http://www.nber.org/papers/w16566

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2010

Our thanks to Alberto Abadie, Victor Chernozhukov, Gary Chamberlain, Guido Imbens, Frank Vella and participants in the 2010 Econometric Society World Congress, Boston University, Carnegie Mellon, Georgetown and Harvard-MIT Econometrics Workshops for helpful discussions and comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework
Joshua Angrist and Ivan Fernandez-Val
NBER Working Paper No. 16566
December 2010
JEL No. C01,C13,C31,C53

ABSTRACT

This paper develops a covariate-based approach to the external validity of instrumental variables (IV) estimates. Assuming that differences in observed complier characteristics are what make IV estimates differ from one another and from parameters like the effect of treatment on the treated, we show how to construct estimates for new subpopulations from a given set of covariate-specific LATEs. We also develop a reweighting procedure that uses the traditional overidentification test statistic to define a population for which a given pair of IV estimates has external validity. These ideas are illustrated through a comparison of twins and sex-composition IV estimates of the effects childbearing on labor supply.

Joshua Angrist
Department of Economics
MIT, E52-353
50 Memorial Drive
Cambridge, MA 02142-1347
and NBER
angrist@mit.edu

Ivan Fernandez-Val
Department of Economics
Boston University
270 Bay State Rd
Boston, MA 02215
ivanf@bu.edu

1 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 IV estimates as instrument-specific should not be surprising or troubling - when this point is cast in terms of specific examples, one wonders 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 the causal effects revealed by any instrument raises questions about the external validity of this estimate. Can we use a given instrumental variables (IV) estimate to identify the effects induced by another source of a variation? How about an unconditional average effect? Can we go from average effects on compliers to average effects on the entire population?

The usual answer to these questions is "no". Except in special cases, we can't go further, at least, not without additional assumptions. As described by Angrist, Imbens, and Rubin (1996), the treated population includes two groups: compliers whose behavior is affected by the instrument at hand and always-takers who are treated irrespective of whether a Bernoulli instrument is switched off or on. The non-treated are likewise composed of compliers and never-takers; the latter group that avoids treatment 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 road to go down in the search for external validity is structural. A latent-index choice framework sometimes allows us to fill in gaps in the data. For example, Heckman, Tobias, and Vytlacil (2001, 2003) and Angrist (2004) use parametric latent-index models to identify and compare different causal effects. Chamberlain (2010) develops a Bayesian semi-parametric 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) summarizes a literature on IV that establishes theoretical links between parameters like LATE and effects on the treated. There's no free lunch, however; these links can be used for *nonparametric* identification of effects other than LATE only with instrumental variables 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 are discrete with finite support, 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 Oreopoulos'

(2006) study of the economic returns to schooling. Oreopoulos compares IV estimates of the returns to schooling across instruments of different strengths. Some of Oreopoulos' instruments are derived from compulsory attendance policies that had modest effects on schooling. But two of the policy experiments in the Oreopoulos study generate instruments with dramatic first-stage effects, something 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 non-treated. As it turns out, Oreopoulos' 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) make 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, where the first stage is larger by an order of magnitude and there are 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. In this paper, 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 other 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.

A second purpose of the investigation described here is to use 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, different (internally valid) instruments capture different causal effects. On the other hand, covariate-specific overidentification tests and summary conditional tests weighted across covariate cells tell us whether differences in the observed characteristics of compliant subpopulations are enough to explain differences in unconditional effects. If so, it seems fair to say that the IV estimates at hand

[1] See also Ebenstein (2009) who compares LATEs generating by first stages of varying strength for the effect of fertility on labor supply in the US and Taiwan.

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 yes or no 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 values as defining a population for which heterogeneity in treatment effects is solely a function of observed characteristics. For this population, the predictive value of IV estimates is likely to be especially high.

The ideas in the paper 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 populations can account for most of the difference between the two sets of IV estimates.

2 Framework

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

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

A random-coefficients notation for this is

$$\text{Y} = \alpha + r\text{D} + \eta,$$

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

We also define potential treatment status indexed against the instrument, Z. Potential treatment status is D_1^{Z} when the instrument is switched on and D_0^{Z} when the instrument is switched off. The variables D_1^{Z} and D_0^{Z} are superscripted to signal the fact that they are tied to Z. Observed treatment status is

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

or in random-coefficients notation

$$\text{D} = \gamma + p\text{Z} + \upsilon,$$

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

IV using Z as an instrument for the effect of D on Y with no covariates is the (1940) Wald 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 only make treatment move in one direction (say, make treatment more likely), those whose treatment status is changed by Z have $\text{D}_1^{\text{Z}} = 1$ and $\text{D}_0^{\text{Z}} = 0$. The causal effect on this group is called the local average treatment effect (LATE; Imbens and Angrist, 1994). Formally, we have:

Assumption 1 (LATE) *(a) Independence and Exclusion:* $(\text{Y}_1, \text{Y}_0, \text{D}_1^{\text{Z}}, \text{D}_0^{\text{Z}}) \amalg \text{Z}$.

(b) First-stage: $\mathbb{E}[\text{D}_1^{\text{Z}} - \text{D}_0^{\text{Z}}] \neq 0$ *and* $0 < \mathbb{P}[\text{Z} = 1] < 1$.

(c) Monotonicity: $\text{D}_1^{\text{Z}} \geq \text{D}_0^{\text{Z}}$ *a.s., or vice versa.*

Theorem 1 (LATE) *Under Assumption 1*

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

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 a set of three instrument-dependent subgroups. These subgroups are defined by the way people react to the instrument:

Definition 1 (subgroups defined by instrument z) *(a)* Z*-Compliers. The subpopulation with* $\text{D}_1^{\text{Z}} = 1$ *and* $\text{D}_0^{\text{Z}} = 0$.

(b) Z*-Always-takers. The subpopulation with* $\text{D}_1^{\text{Z}} = \text{D}_0^{\text{Z}} = 1$.

(c) Z*-Never-takers. The subpopulation with* $\text{D}_1^{\text{Z}} = \text{D}_0^{\text{Z}} = 0$.

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

3 Covariate Heterogeneity

3.1 Two instruments for one effect

The relationship between fertility and labor supply has long been of interest to labor economists, while the case for omitted variables bias in this context 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 hard to interpret since mothers who have big families may have worked less anyway. Angrist and Evans (1998) solve this omitted-variables problem using two instrumental variables, 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 for a multiple second birth in a sample of mothers with at least two children. The twins first-stage is about .6, an estimate reported here 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 .38). Internal validity of the same-sex instrument rests on the claim that sibling sex composition is essentially random and affects labor supply solely by increasing fertility.

Twins and sex-composition IV estimates both suggest that the birth of a third child has a substantial effect on weeks worked and employment. Wald estimates using twins instruments show a precisely-estimated reduction in weeks worked of a little over 3 weeks, with an employment reduction of about .08. These results, which appear in column 3 of Table 1, are smaller in absolute value than the corresponding OLS estimates reported in column 1 (the latter include a set of controls listed in the table). This suggests the OLS estimates are exaggerated by selection bias. Interestingly, however, 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 .02 and .06, implying that the twins and sex-composition IV estimates are at least marginally significantly different from one another.

Twins and samesex IV estimates reflect behavior in different compliant subpopulations. To see this, let x be a Bernoulli-distributed characteristic, say 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 calculation:

$$\frac{\mathbb{P}[\text{X} = 1 \mid \text{D}_1^{\text{Z}} > \text{D}_0^{\text{Z}}]}{\mathbb{P}[\text{X} = 1]} = \frac{\mathbb{P}[\text{D}_1^{\text{Z}} > \text{D}_0^{\text{Z}} \mid \text{X} = 1]}{\mathbb{P}[\text{D}_1^{\text{Z}} > \text{D}_0^{\text{Z}}]} = \frac{\mathbb{E}[\text{D} \mid \text{Z} = 1, \text{X} = 1] - \mathbb{E}[\text{D} \mid \text{Z} = 0, \text{X} = 1]}{\mathbb{E}[\text{D} \mid \text{Z} = 1] - \mathbb{E}[\text{D} \mid \text{Z} = 0]},$$

where the second equality follows by Bayes rule. In other words, the relative likelihood 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 mothers schooling as described by dummies for high school graduates, mothers with some college, and college graduates. Twins compliers have younger second-born children, reflecting the fact that few women who had their second child recently 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. This is important because the birth of a third child may matter less if the second child is also young, helping to explain the finding that twins-IV estimates are smaller than samesex estimates (Gelbach 2002 suggests the presence of a child younger than age 5 in the household is a key labor supply mediator).

Twins compliers are also more likely to be college graduates than the average mother, while sex-composition compliers are less educated. This fact also helps to explain the smaller 2SLS estimates generated by twins instruments, since Angrist and Evans (1998) show 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

$$\mathbb{E}[\text{X} \mid \text{D}_1 > \text{D}_0] = \frac{\mathbb{E}[\kappa^{\text{Z}}(\text{X})\ \text{X}]}{\mathbb{E}[\kappa^{\text{Z}}(\text{X})]}, \tag{2}$$

where

$$\kappa^{\text{Z}}(x) = 1 - \frac{\text{D}(1 - \text{Z})}{1 - \mathbb{P}[\text{Z} = 1 \mid \text{X} = x]} - \frac{(1 - \text{D})\text{Z}}{\mathbb{P}[\text{Z} = 1 \mid \text{X} = x]}.$$

Intuitively, this works because, as Abadie shows, the weighting function, $\kappa^{\text{Z}}(x)$, "finds compliers," even though it is not a simple indicator for compliers. Estimates of $\mathbb{E}[\text{X} \mid \text{D}_1 > \text{D}_0]$ for age of second child and mothers 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 small difference in schooling. The latter result is due to the fact that the main difference between the schooling of twins and samesex compliers is in the proportion of college graduates. Relatively few of the women in the high-fertility 1980 census sample used to construct these estimates were 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 like to control for race and maternal age when using twins instruments since the probability of multiple births varies by race and increases with maternal age. Second, we propose to use covariates for extrapolation. Specifically, we argue that in some cases, including the twins and samesex comparison, 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 express the idea that we think of the instrumental variables as being "as good as randomly assigned," conditional on covariates, $\textsc{x}$. This generalizes Assumption 1:

Assumption 2 (Conditional LATE) *(a) Independence and Exclusion:* $(\textsc{y}_1, \textsc{y}_0, \textsc{d}_1^{\textsc{z}}, \textsc{d}_0^{\textsc{z}}) \amalg \textsc{z} \mid \textsc{x}$ *a.s.;*

(b) First-stage: $\mathbb{E}[\textsc{d}_1^{\textsc{z}} - \textsc{d}_0^{\textsc{z}} \mid \textsc{x}] \neq 0$ *and* $0 < \mathbb{P}[\textsc{z} = 1 \mid \textsc{x}] < 1$ *a.s.*

(c) Monotonicity: $\mathbb{P}[\textsc{d}_1^{\textsc{z}} \geq \textsc{d}_0^{\textsc{z}} \mid \textsc{x}] = 1$ *a.s., or* $\mathbb{P}[\textsc{d}_1^{\textsc{z}} \leq \textsc{d}_0^{\textsc{z}} \mid \textsc{x}] = 1$ *a.s.*

For each value of $\textsc{x}$, we define covariate-specific LATE constructed using $\textsc{z}$, as

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

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

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

where $F_{\textsc{x}}(\cdot \mid \textsc{d}_1^{\textsc{z}} > \textsc{d}_0^{\textsc{z}})$ is the distribution of $\textsc{x}$ for $\textsc{z}$-compliers and $F_{\textsc{x}}(\cdot)$ is the distribution of $\textsc{x}$ in the population.

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

Assumption 3 (CEI) *Conditional Effect Ignorability for an instrument* $\textsc{z}$*:*

$$\mathbb{E}[\textsc{y}_1 - \textsc{y}_0 \mid \textsc{d}_1^{\textsc{z}}, \textsc{d}_0^{\textsc{z}}, \textsc{x}] = \mathbb{E}[\textsc{y}_1 - \textsc{y}_0 | \textsc{x}] \quad a.s.$$

A sufficient condition for CEI is

$$\textsc{y}_1 = \textsc{y}_0 + g(\textsc{x}) + \nu,$$

where $g(\text{X})$ is any function and ν is mean-independent of $(\text{D}_1^{\text{Z}}, \text{D}_0^{\text{Z}})$ conditional on X. In other words, heterogeneity in average causal effects is solely due to observed covariates.[2]

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

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

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:

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

The associated model for potential outcomes is

$$\begin{aligned} \text{Y}_0 &= g_0(\text{X}) + \epsilon_0, \\ \text{Y}_1 &= g_1(\text{X}) + \epsilon_1. \end{aligned}$$

The errors are mean independent of the covariates and instrument. Assuming $h^{\text{Z}}(\text{X}, 1) \geq h^{\text{Z}}(\text{X}, 0)$ a.s., conditional LATE can be written

$$\Delta^{\text{Z}}(\text{X}) = \mathbb{E}[\text{Y}_1 - \text{Y}_0 \mid \text{X}, h^{\text{Z}}(\text{X}, 1) > \eta > h^{\text{Z}}(\text{X}, 0)] = g_1(\text{X}) - g_0(\text{X}) + \mathbb{E}[\epsilon_1 - \epsilon_0 \mid \text{X}, \text{D}_1^{\text{Z}} > \text{D}_0^{\text{Z}}]. \quad (4)$$

The CEI assumption says that $\epsilon_1 - \epsilon_0$ is mean independent of $(\text{D}_1^{\text{Z}}, \text{D}_0^{\text{Z}})$ conditional on X, so that (4) simplifies to

$$\mathbb{E}[\text{Y}_1 - \text{Y}_0 \mid \text{X}] = g_1(\text{X}) - g_0(\text{X}).$$

In the language of Rubin (1977), CEI is a kind of ignorability assumption for treatment effects. Given ignorability, you might wonder why we need be concerned about selection bias in the first place. Under CEI, selection bias arises due to correlation between η and ϵ_0. Alternatively, a latent-index specification compatible with the CEI sets $\epsilon_j = \theta + \xi_j$ where θ is correlated with η but ξ_j is not. In a latent index formulation of the assignment mechanism, 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.[3] 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.

[2] 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 where selection bias is likely to minimal.

[3] A fact noted by many authors working with models of this type; see, e.g., Vella and Verbeek,1999

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

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

Conditional LATE then becomes

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

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

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

where $\lambda(x)$ is the weight given to outcome gaps, $c(x,z)$ is the expected cost of having a third child, the instrument, $\textsc{z}$, is a cost shifter independent of potential outcomes, and η is private information about $\textsc{y}_1 - \textsc{y}_0$, orthogonal to $\textsc{x}$. Then $\rho(x)$ is small (close to zero) when either $\lambda(x)$ is small, i.e., labor supply consequences are of little import; or η matters little given $\textsc{x}$, e.g., for women with a young second-born that are already at home or for relatively educated women who can more easily afford to pay for child care. This offers a possible explanation for why CEI might be satisfied for women with some characteristics but not for others. We consider econometric models where CEI is partially satisfied in Section 3.4.

3.3 Reweighting LATE

Our covariate-based strategy reweights treatment effects across covariates cells. This is similar to matching estimators designed to control for selection bias; such estimators reweight the conditional mean function for outcome variables to identify 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* $\textsc{z}$ *be an instrument that satisfy Assumption 3 and let* $\textsc{s}^{\textsc{z}} = \textsc{s}(\textsc{d}_0^{\textsc{z}}, \textsc{d}_1^{\textsc{z}}, \textsc{z})$ *be an indicator for any subpopulation defined by* $\textsc{z}$*. For example, for* $\textsc{z}$*-compliers, we have* $\textsc{s}^{\textsc{z}} = \textsc{d}_1^{\textsc{z}} - \textsc{d}_0^{\textsc{z}}$*, for the treated* $\textsc{s}^{\textsc{z}} = (1-\textsc{z})\textsc{d}_0^{\textsc{z}} + \textsc{z}\textsc{d}_1^{\textsc{z}} = \textsc{d}$*, for the non-treated* $\textsc{s}^{\textsc{z}} = (1-\textsc{z})(1-\textsc{d}_0^{\textsc{z}}) + \textsc{z}(1-\textsc{d}_1^{\textsc{z}}) = 1 - \textsc{d}$*, and for the entire population* $\textsc{s}^{\textsc{z}} = 1$*. Under Assumption 2* *and* $\mathbb{E}[|\textsc{y}|] < \infty$,

$$\mathbb{E}[\textsc{y}_1 - \textsc{y}_0 \mid \textsc{s}^{\textsc{z}} = 1] = \mathbb{E}[\Delta^{\textsc{z}}(\textsc{x}) \mid \textsc{s}^{\textsc{z}} = 1] = \int \Delta^{\textsc{z}}(x) dF_{\textsc{x}}(x \mid \textsc{s}^{\textsc{z}} = 1).$$

This is

$$\int \Delta^{\textsc{z}}(x)\omega_{\textsc{s}}^{\textsc{z}}(x) dF_{\textsc{x}}(x),$$

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

Proof. By the law of iterated expectations

$$\mathbb{E}[\textsc{y}_1 - \textsc{y}_0 \mid \textsc{s}^{\textsc{z}} = 1] = \mathbb{E}[\mathbb{E}[\textsc{y}_1 - \textsc{y}_0 \mid \textsc{s}^{\textsc{z}} = 1, \textsc{x}] \mid \textsc{s}^{\textsc{z}} = 1] = \mathbb{E}[\mathbb{E}[\textsc{y}_1 - \textsc{y}_0 \mid \textsc{x}] \mid \textsc{s}^{\textsc{z}} = 1] = \mathbb{E}[\Delta^{\textsc{z}}(\textsc{x}) \mid \textsc{s}^{\textsc{z}} = 1],$$

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

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

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

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

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 non-treated (TNT). The relevant weighting functions, $\omega_{\textsc{s}}^{\textsc{z}}(x)$, can be written in terms of observed variables as:

$$\omega_{\Delta}^{\textsc{z}}(x) = \frac{\mathbb{E}[\textsc{d} \mid \textsc{z} = 1, \textsc{x} = x] - \mathbb{E}[\textsc{d} \mid \textsc{z} = 0, \textsc{x} = x]}{\mathbb{E}[\textsc{d} \mid \textsc{z} = 1] - \mathbb{E}[\textsc{d} \mid \textsc{z} = 0]} \tag{5}$$

for effects on z-compliers;

$$\omega_{TOT}^{\textsc{z}}(x) = \mathbb{E}[\textsc{d} \mid \textsc{x} = x]/\mathbb{E}[\textsc{d}] \tag{6}$$

for effects on the treated;

$$\omega_{TNT}^{\textsc{z}}(x) = \mathbb{E}[1 - \textsc{d} \mid \textsc{x} = x]/\mathbb{E}[1 - \textsc{d}] \tag{7}$$

for effects on the non-treated; and

$$\omega_{ATE}^{\textsc{z}}(x) = 1 \tag{8}$$

for the population.

3.4 Overidentification

Differences in the observable characteristics of twins and samesex 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. To see this, assume that two instruments Z and W are available. The difference between the LATE generated by each can be decomposed as

$$\Delta^{\mathrm{Z}} - \Delta^{\mathrm{W}} = \int [\Delta^{\mathrm{Z}}(x) - \Delta^{\mathrm{W}}(x)] \omega_{\Delta}^{\mathrm{Z}}(x) dF_{\mathrm{X}}(x) + \int \Delta^{\mathrm{W}}(x) [\omega_{\Delta}^{\mathrm{Z}}(x) - \omega_{\Delta}^{\mathrm{W}}(x)] dF_{\mathrm{X}}(x). \tag{9}$$

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

$$\Delta^{\mathrm{W}}(\mathrm{X}) = \Delta^{\mathrm{Z}}(\mathrm{X}) \text{ a.s.} \tag{10}$$

and the first term of the decomposition (9) is zero. In other words, joint CEI implies that the difference between LATEs is driven solely by differences in the observed characteristics of compliers (the covariate vector X is assumed to be observed). The following theorem shows that under joint CEI we can use the distribution of compliers for a hypothetical instrument to construct the treatment effects that might be generated by instruments other than the ones we've got.

Theorem 3 (LATE-Overid) *Let Z and W be two instruments that satisfy Assumptions 2 and 3. Let $\Delta^{\mathrm{W}}(\mathrm{X})$ be defined as in (3) using the instrument W. Let $\mathrm{S}^{\mathrm{W}} = \mathrm{S}(\mathrm{D}_0^{\mathrm{W}}, \mathrm{D}_1^{\mathrm{W}}, \mathrm{W})$ be an indicator for any subpopulation defined by instrument W with corresponding treatment assignments $(\mathrm{D}_0^{\mathrm{W}}, \mathrm{D}_1^{\mathrm{W}})$. If $\mathbb{E}[|\mathrm{Y}|] < \infty$,*

$$\mathbb{E}[\mathrm{Y}_1 - \mathrm{Y}_0 \mid \mathrm{S}^{\mathrm{W}} = 1] = \int \Delta^{\mathrm{W}}(x) \omega_{\mathrm{S}}^{\mathrm{W}}(x) dF_{\mathrm{X}}(x) = \int \Delta^{\mathrm{Z}}(x) \omega_{\mathrm{S}}^{\mathrm{W}}(x) dF_{\mathrm{X}}(x). \tag{11}$$

Proof. The first equality in equation (11) follows from Theorem 2 applied to W. For the second equality, the law of iterated expectations and CEI for Z give

$$\mathbb{E}[\mathrm{Y}_1 - \mathrm{Y}_0 \mid \mathrm{S}^{\mathrm{W}} = 1] = \mathbb{E}[\mathbb{E}[\mathrm{Y}_1 - \mathrm{Y}_0 \mid \mathrm{X}] \mid \mathrm{S}^{\mathrm{W}} = 1] = \mathbb{E}[\Delta^{\mathrm{Z}}(\mathrm{X}) \mid \mathrm{S}^{\mathrm{W}} = 1].$$

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 populations with these covariate values and therefore some claim to external validity.

Joint CEI also implies that the conditional LATEs $\Delta^{\text{Z}}(x)$ are overidentified if we observe both Z and W. We can therefore construct more precise estimators for conditional LATEs with 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 (10). Comparisons of estimates of the two expressions on the RHS of (11) can 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 was given 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 population for which heterogeneous treatment effects can be understood to be solely a function of observable characteristics. For this population, 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*

$$\Delta^{\text{Z,W}} = \int \Delta^{\text{Z}}(x) dF_{\text{X}}(x \mid \Delta^{\text{Z}}(x) = \Delta^{\text{W}}(x)).$$

If the compatibility condition $\Delta^{\text{Z}}(x) = \Delta^{\text{W}}(x)$ holds for all values of X, CATE is ATE. But if compatibility holds only for some values of X, CATE is ATE for the population defined by these values. Below, we develop an estimator for CATE when the compatibility condition is only partially satisfied.

4 Estimation and Inference

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

Assumption 4 (Sampling) $\{\text{R}_i = (\text{Y}_i, \text{D}_i, \text{X}_i, \text{Z}_i, \text{W}_i), i = 1, \ldots, n\}$ *are i.i.d. observations from* $\text{R} = (\text{Y}, \text{D}, \text{X}, \text{Z}, \text{W})$.

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'd like to 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* $\mathcal{X} = \{x_1, ..., x_K\}$, $\mathbb{P}[\text{X} \in \mathcal{X}] = 1$.

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

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

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, the superscript L indexes the population where conditional LATEs are obtained and $\mathrm{S^U}$ is an indicator for the population with the distribution of covariate characteristics 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 $\mathbb{P}_n$. For conditional expectations and probabilities, 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 $\mathrm{U} = u$ (or for everyone in the covariate cell if we don't condition on U), for $\mathrm{U} \in \{\mathrm{Z}, \mathrm{W}\}$ and $u \in \{z, w\}$. This gives

$$\hat{\Delta}^{\mathrm{L}}_{\mathrm{S^U}} = \mathbb{E}_n[\hat{\Delta}^{\mathrm{L}}(\mathrm{X}_i)\hat{\omega}^{\mathrm{U}}_{\mathrm{S}}(\mathrm{X}_i)],\ \hat{\omega}^{\mathrm{U}}_{\mathrm{S}}(x) = \mathbb{P}_n[\mathrm{S}^{\mathrm{U}}_i = 1 \mid \mathrm{X} = x]/\mathbb{P}_n[\mathrm{S}^{\mathrm{U}}_i = 1], \tag{12}$$

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 the GMM estimator using both Z and W as instruments in cell $\mathrm{X} = x$. For treated, non-treated, and the entire population, the indicator $\mathrm{S}^{\mathrm{U}}_i$ is observed and 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 (5):

$$\hat{\omega}^{\mathrm{U}}_{\Delta}(x) = \frac{\mathbb{E}_n[\mathrm{D}_i \mid \mathrm{U} = 1, \mathrm{X} = x] - \mathbb{E}_n[\mathrm{D}_i \mid \mathrm{U} = 0, \mathrm{X} = x]}{\mathbb{E}_n[\mathrm{D}_i \mid \mathrm{U} = 1] - \mathbb{E}_n[\mathrm{D}_i \mid \mathrm{U} = 0]},\ \mathrm{U} \in \{\mathrm{Z}, \mathrm{W}\}. \tag{13}$$

Consistency of $\hat{\Delta}^{\mathrm{L}}_{\mathrm{S^U}}$ 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* $\mathbb{E}[|\mathrm{Y}|] < \infty$

$$\hat{\Delta}^{\mathrm{L}}_{\mathrm{S^U}} = \mathbb{E}_n[\hat{\Delta}^{\mathrm{L}}(\mathrm{X}_i)\hat{\omega}^{\mathrm{U}}_{\mathrm{S}}(\mathrm{X}_i)] \to_p \Delta^{\mathrm{L}}_{\mathrm{S^U}} = \mathbb{E}[\mathrm{Y}_1 - \mathrm{Y}_0 \mid \mathrm{S^W} = 1], \quad \mathrm{L}, \mathrm{U} \in \{\mathrm{Z}, \mathrm{W}\},$$

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 here are smooth functions of GMM-type estimators and are therefore asymptotically normal under standard regularity conditions. The following result uses the delta method to characterize the relevant limiting distributions. In particular, we show that $(\hat{\Delta}^{\mathrm{Z}}_{\mathrm{S^U}}, \hat{\Delta}^{\mathrm{W}}_{\mathrm{S^U}})$ are asymptotically jointly normal, what allows us to draw inferences about the effects of interest and to test some of the implications of joint CEI. Let $p_k = \mathbb{P}(\mathrm{X} = x_k)$, $\vartheta^{\mathrm{U}}_{\mathrm{S}}(x) = \mathbb{P}[\mathrm{S^U} = 1 \mid \mathrm{X} = x]$, i.e., the numerator of $\omega^{\mathrm{U}}_{\mathrm{S}}(x)$, and $\hat{\vartheta}^{\mathrm{U}}_{\mathrm{S}}(x)$ 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* $p_k > 0$ *for all* $k \in \{1, ..., K\}$ *and, for* $k \in \{1, ..., K\}$ *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 $(Z^{\mathrm{U}}_1, ..., Z^{\mathrm{U}}_K)$ *are independent. Under Assumptions 4 and 5, 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 CLT for multinomial sequences, $\sqrt{n}(\hat{p}_1 - p_1, ..., \hat{p}_K - p_K)$ converges in distribution to a multivariate normal with zero mean, variances $p_k(1 - p_k)$ and covariances $-p_k p_j$, for $k, j = 1, ..., K$, $k \neq j$. 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 $\vec{\pi} = (\pi_1, ..., \pi_K)$, $\vec{\upsilon} = (\upsilon_1, ..., \upsilon_K)$, and $\vec{\delta} = (\delta_1, ..., \delta_K)$. 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:

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

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 $\pi_k = \hat{p}_k$, $\upsilon_k = \hat{\vartheta}^{\mathrm{U}}_{\mathrm{S}}(x_k)$, and $\delta_k = \Delta^{\mathrm{L}}(x_k)$. The result then follows using the delta method.
■

It's worth noting that 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 CLT for binary sequences. For IV, GMM, and other moment-based estimators of the conditional LATEs, such as generalized empirical likelihood (GEL), the existence of second moments, i.e., $\mathbb{E}[\text{Y}^2] < \infty$, is sufficient.

Note also that the first term in the expressions for $V^{\text{Z}}_{\text{S}^{\text{U}}}$, $V^{\text{W}}_{\text{S}^{\text{U}}}$, and $C^{\text{ZW}}_{\text{S}^{\text{U}}}$ 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 the conditional LATEs. The first term is zero if, for example, the conditional LATEs are constant, i.e., $\Delta^{\text{L}}(x) = \Delta^{\text{L}}_{\text{S}^{\text{U}}}$ for all $x \in \mathcal{X}$.

In practice, there are two routes to asymptotic inference. We can estimate the asymptotic distributions analytically using sample analogs or approximate them numerically using resampling or simulation methods. We use bootstrap methods in the empirical application. Resampling methods are convenient and save us from having to estimate complicated analytical formulas for the asymptotic variances and covariances. Consistency of the 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 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 $\mathbb{E}[g(\text{R}, \theta)] = 0$ be the moment conditions that define the conditional LATEs, weighting functions and effects of interest, where θ includes all the unknown parameters. The ELD $(\hat{\pi}_1, ..., \hat{\pi}_n)$ is the solution to

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

where $\hat{\theta}$ is the EL or another consistent estimator of θ.[4] ELD is therefore 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 EL bootstrap 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 $\{\Delta^{\text{Z}}(x) = \Delta^{\text{W}}(x)\}$. In finite samples, we never have $\hat{\Delta}^{\text{Z}}(x) = \hat{\Delta}^{\text{W}}(x)$. 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 $\text{X} = x$, the resulting estimator of

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

CATE is

$$\hat{\Delta}^{\mathrm{z,w}} = \mathbb{E}_n[\hat{\Delta}^{\mathrm{z,w}}(\mathrm{x}_i)\hat{\omega}_{CATE}(\mathrm{x}_i)],\ \hat{\omega}_{CATE}(x) = \exp\{-J(x)/a_n(x)\}/\mathbb{E}_n[\exp\{-J(\mathrm{x}_i)/a_n(\mathrm{x}_i)\}], \tag{14}$$

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 (such as 2SLS), and $a_n(x)$ is a sequence such that $a_n(x) \to \infty$ and $a_n(x) = o(n)$, for $x \in \mathcal{X}$.

The sequences $a_n(x)$ guarantee the consistency of the reweighting estimator (14) 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,* $\Pr\{\mathrm{x} \in \mathcal{X}_0\} > 0$*, and* $\mathbb{E}[|\mathrm{y}|] < \infty$,

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

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

Proof. Write

$$\hat{\Delta}^{\mathrm{z,w}} = \sum_{k=1}^{K} \hat{p}_k \hat{\Delta}^{\mathrm{z,w}}(x_k)\hat{\omega}_{CATE}(x_k) \text{ and } \Delta^{\mathrm{z,w}} = \sum_{k=1}^{K} p_k \Delta^{\mathrm{z}}(x_k) 1\{x_k \in \mathcal{X}_0\} / \sum_{k=1}^{K} p_k 1\{x_k \in \mathcal{X}_0\}.$$

By the LLN, $\hat{p}_k \to_p p_k$. 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 Slutsky Theorem, noting that

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

■

Note that 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 $\hat{\Delta}^{\text{z,w}}$ depends on the limiting distributions of the weighting functions. But here, 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 be slower than $\sqrt{n}$. [5] 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) deal with a pretest problem using subsampling, but 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 conditions on the bootstrap DGP.

5 Results

Our empirical exploration of covariate-reweighting focuses on a categorical representation of second child age and mother's schooling. As we've seen, these covariates are strongly related to compliance probabilities. On one hand, 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. Multiple pregnancies are therefore more important for women with a young second-born and for women with some college and college degrees. Sex-composition compliers, by contrast, are relatively unlikely to be college graduates or to have a third child soon after the birth of the second. Labor supply effects are also 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 already 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.

[5] 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're interested in predictive population inference as opposed to sample-specific causal inference.

Differences in twins and samesex complier populations might therefore account for the fact that twins instruments generate a smaller effect of childbearing than sex-composition instruments.

In 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, older 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 don't have a third child are clearly very different.

Not surprisingly, IV estimates are fairly noisy across cells, as can be seen in columns 4 and 6 of Table 3. From these noisy cell-by-cell estimates alone, it's hard to tell 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 samesex compliers (column 7). This is just the ratio of the relevant first stage in the cell to the overall first stage. The distribution of twins and samesex compliers over cells is clearly different from the cell distribution in the random sample. Not surprisingly, given the summary comparisons in Table 2, the complier distributions for the two instruments are also very different from each other. Specifically, twins compliers are much more likely to have a young second-born child, while few samesex compliers are in this group. Twins compliers are also relatively educated, while the schooling gradient in compliance probabilities for samesex is less pronounced.

Reweighting covariate-specific samesex IV estimates using twin-complier weights brings these estimates much closer to the IV estimate using twins instruments. This can be seen in the first two rows of Table 4, which report

$$\Delta_{\mathrm{S}^{\mathrm{U}}}^{\mathrm{L}} = \mathbb{E}[\Delta^{\mathrm{L}}(\mathrm{X})\omega_{\mathrm{S}}^{\mathrm{U}}(\mathrm{X})],\ \omega_{\mathrm{S}}^{\mathrm{U}}(x) = \mathbb{P}[\mathrm{S}^{\mathrm{U}} = 1 \mid \mathrm{X} = x]/\mathbb{P}[\mathrm{S}^{\mathrm{U}} = 1],$$

for $\mathrm{L} = \mathrm{U} = \mathrm{TWINS}$ and $\mathrm{L} = \mathrm{SAMESEX}$ and $\mathrm{U} = \mathrm{TWINS}$. Compare, for example, the estimates of effects on weeks worked of -3.15 using twins instruments and -2.71 using samesex instruments. Reweighting covariate-specific twins and samesex estimates using samesex weights also produces a good match. Compare, in this case, the estimate of -6.3 using samesex instruments to an estimate of -5.08 using twins instruments. In other words, the IV estimates generated using twins and samesex instruments can indeed be effectively reconciled by reweighting covariate-specific effects using a set of common weights.

This 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 samesex instruments are not well-matched. This is disappointing because, by the same argument that appears to reconcile twins and samesex estimates of LATE, we should be able to generate similar estimates of ATE and TOT using either instrument to construct cell-specific IV estimates (the estimators for ATE and TOT apply a common set of weights to the underlying set of IV estimates). The match for TNT is not bad as that for ATE and TOT. The fact that some parameters can be more easily matched across instruments than others suggests that a few poorly matched cells are what drives the cross-instrument imbalance in estimates of ATE and TOT, i.e., that a handful of cells generate different estimates depending on the instrument.

CATE, at the bottom of Table 4, solves this problem, and produces a good match for the average treatment effect in compatible cells by downweighting cells where CEI is most at odds with the data. Compare CATE estimates of -3.80 to -3.66 for effects on weeks worked and -.099 to -.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 over-identification test is never very large. Still, by applying the most weight to cells where CEI appears to be satisfied, CATE generates remarkably similar estimates of the population ATE, whether the underlying cell-level IV estimates use twins or samesex instruments.

The match generated by CATE weighting is for a subpopulation and not a random sample. Figure 1 describes the compatible subpopulation by plotting the weighting functions used by different estimators. CATE weights essentially discard the two low-education cell for women with an older second-born, and look a lot like the histogram for twins compliers. Because twins instruments induce one-sided non-compliance, the population of twins compliers is the same as the nontreated population (Angrist, Lavy, and Schlosser, 2010). The population for which IV estimates of the effects of childbearing have predictive value therefore consists mainly of mothers who have a young second-born child and are somewhat more likely to have gone to college.

6 Summary and Directions for Further Work

In the LATE framework, differences in IV estimates no longer signal a failure of the exclusion restriction. Rather, these differences may be attributable to differences in the sort of people who are affected by the underlying experiments implicit in any IV identification strategy. At the same time, we'd very much like 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 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 do sex composition instruments in the Angrist-Evans (1998) data set.

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 micro 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 ought to at least 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 have also explored an approach that allows for 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 average treatment effects using twins and samesex instruments when the sample is reweighted towards cells that appear to satisfy CEI. At the same time, we acknowledge that such 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.

References

[1] Abadie, Alberto. 2003. "Semiparametric instrumental variable estimation of treatment response models." *Journal of Econometrics* 113(2): 231-63.

[2] Andrews, Donald W. K. 1999. "Estimation when a parameter is on a boundary." *Econometrica* 67(6): 1341-83.

[3] Andrews, Donald W. K., and Patrik Guggenberger. 2009. "Hybrid and size-corrected subsampling methods." *Econometrica* 77(3): 721-62.

[4] Angrist, Joshua D. 2004. "Treatment effect heterogeneity in theory and practice." *Economic Journal* 114(494): C52-C83.

[5] Angrist, Joshua D., and William N. Evans. 1998. "Children and their parents' labor supply: Evidence from exogenous variation in family size." *American Economic Review* 88(3): 450-77.

[6] Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of causal effects using instrumental variables." *Journal of the American Statistical Association* 91: 444-72.

[7] Angrist, Joshua D., Victor Lavy, and Analia Schlosser. 2010. "Multiple experiments for the causal link between the quantity and quality of children." *Journal of Labor Economics* 28, 773-824.

[8] Brown, Bryan W., and Whitney K. Newey. 2002. "Generalized method of moments, efficient bootstrapping, and improved inference." *Journal of Business & Economic Statistics* 20(4): 507- 17.

[9] Chamberlain, Gary. 2010. "Bayesian aspects of treatment choice." Discussion paper, Harvard University.

[10] Crump, Richard K., V. J. Hotz, Guido W. Imbens, and Oscar A. Mitnik. 2009. "Dealing with limited overlap in estimation of average treatment effects." *Biometrika* 96(1): 187-99.

[11] Ebenstein, Avraham. 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.

[12] Frangakis, Constantine E., and Donald B. Rubin. 2002. "Principal stratification in causal inference." *Biometrics* 58(1): 21-29

[13] Frölich, Markus. 2007. "Nonparametric IV estimation of local average treatment effects with covariates." *Journal of Econometrics* 139(1): 35-75.

[14] Gelbach, Jonah. B. 2002. "Public schooling for young children and maternal labor supply." *American Economic Review*, 92(1): 307-22.

[15] Hahn, Jinyong. 1996. "A note on bootstrapping generalized method of moments estimators." *Econometric Theory* 12: 187-97.

[16] Hahn, Jinyong. 1998. “On the role of the propensity score in efficient semiparametric estimation of average treatment effects.” *Econometrica* 66(2): 315-32.

[17] Hall, Peter, and Joel L. Horowitz. 1996. “Bootstrap critical values for tests based on generalized-Method-of-moments estimators.” *Econometrica* 64(4): 891-916.

[18] Heckman, James J. 2010. “Building bridges between structural and program evaluation approaches to evaluating policy.” *Journal of Economic Literature* 48(2): 356- 98.

[19] Heckman, James J., Justin L. Tobias, and Edward Vytlacil. 2001. “Four parameters of interest in the evaluation of social programs.” *Southern Economic Journal* 68(2): 210-23.

[20] ______. 2003. “Simple estimators for treatment parameters in a latent-variable framework.” *Review of Economics and Statistics* 85 (3): 748-55.

[21] Imbens, Guido W., and Whitney K. Newey. 2009. “Identification and estimation of triangular simultaneous equations models without additivity.” *Econometrica* 77(5): 1481-1512.

[22] Imbens, Guido W., and Joshua D. Angrist. 1994. “Identification and estimation of local average treatment effects.” *Econometrica* 62(2): 467-75.

[23] Leeb, Hannes, and Meredikt M. Potscher 2008, “Model selection,” in T.G. Andersen, R.A. Davis, J.-P. Kreiss and T. Mikosch (eds.), *Handbook of Financial Time Series.* Springer-Verlag.

[24] Olsen, Randall, J. 1980. “A least squares correction for selectivity bias.” *Econometrica* 48(7): 1815-20

[25] Oreopoulos, Philip. 2006. “Estimating average and local average treatment effects of education when compulsory schooling laws really matter.” *American Economic Review* 96(1): 152-75.

[26] Rosenzweig, Mark R., and Kenneth I. Wolpin. 1980. “Testing the quantity-quality fertility model: The use of twins as a natural experiment.” *Econometrica* 48(1): 227-40.

[27] Roy, A. D. 1951. “Some thoughts on the distribution of earnings.” *Oxford Economic Papers* 3(2): 135-46.

[28] Rubin, Donald B. 1977. “Assignment to a treatment group on the basis of a covariate.” *Journal of Educational Statistics* 2(1): 1-26.

[29] van der Vaart, Aad W. 1998. *Asymptotic statistics*. New York: Cambridge University Press.

[30] Vella, Francis, and Marno Verbeek. 1999. "Estimating and Interpreting Models with Endogenous Treatment Effects." *Journal of Business & Economic Statistics* 17(4): 473-78.

[31] Vytlacil, Edward. 2002. "Independence, monotonicity, and latent index models: An equivalence result." *Econometrica* 70(1): 331-41.

[32] Wald, Abraham. 1940. "The fitting of straight lines if both variables are subject to error." *Annals of Mathematical Statistics* 11(3): 284-300.

[33] Willis Robert J., and Sherwin Rosen. 1979. "Education and self-selection." *The Journal of Political Economy* 87(5): S7-S36.

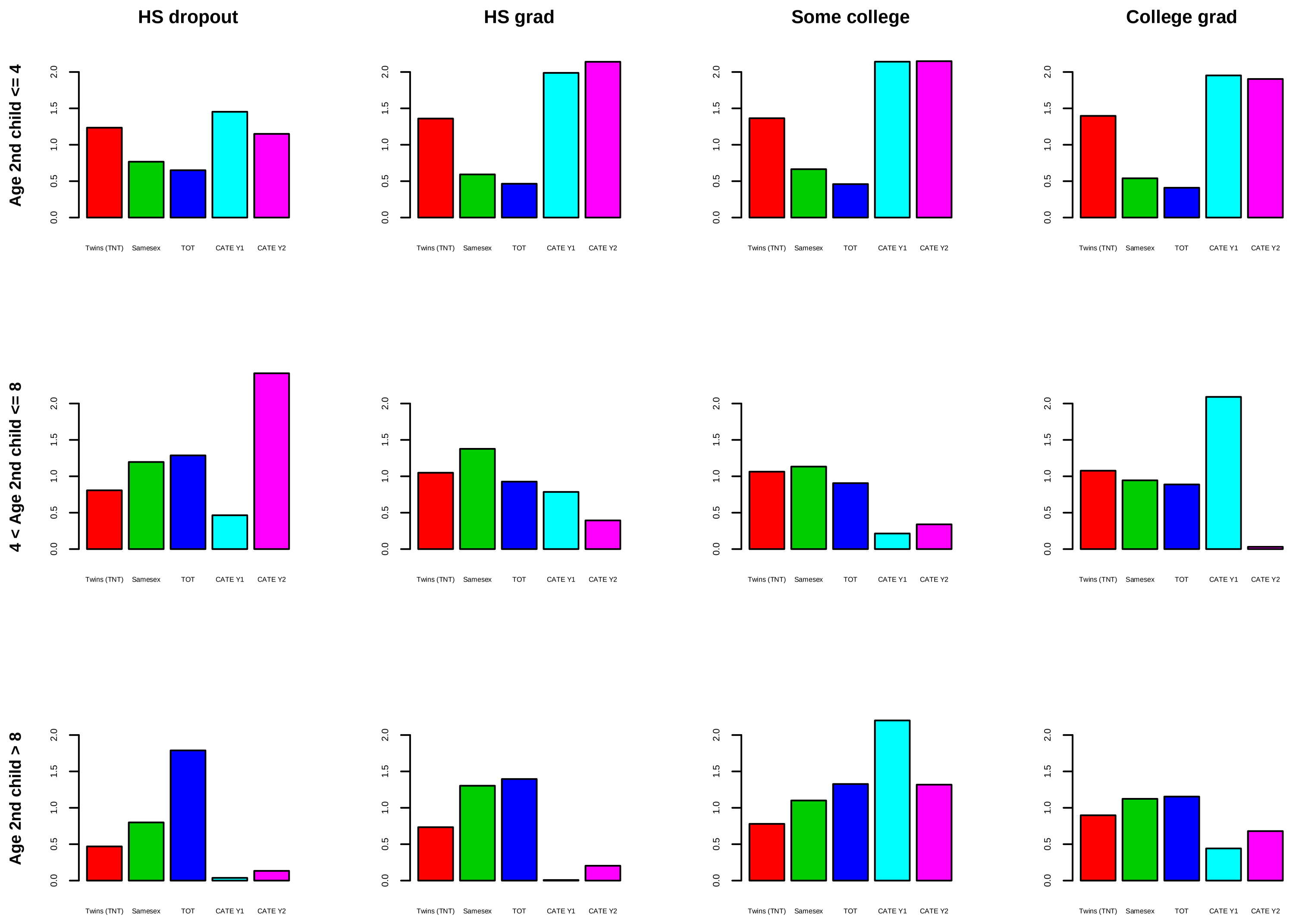

Figure 1: The distribution of compliers and related subpopulations across covariate cells: weighting functions $\hat{\omega}_{\mathrm{S}}^{\mathrm{U}}(x)$. Y1 = weeks worked, Y2= employment.

Table 1: Wald estimates of the effects of family size on labor supply

			Twins instrument		Samesex instrument		Both
Dependent Variable	Mean	OLS	First Stage	Wald Estimates	First Stage	Wald Estimates	2SLS Estimates
		(1)	(2)	(3)	(4)	(5)	(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(.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(.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 including women aged 21-35 with at least two children in the 1980 census. OLS models include controls for mother's age, age at first birth, ages of the first two children, and dummies for race. The first stage is the same for all dependent variables. The sample size is 394,840.

Table 2: Complier characteristics ratios for twins and sex composition instruments

	Population mean	Mean for twins-compliers		Mean for samesex-compliers	
	$E[x_{1i}]$	$E[x_{1i} \mid D_{1i} > D_{0i}]$	$E[x_{1i} \mid D_{1i} > D_{0i}]/$ $E[x_{1i}]$	$E[x_{1i} \mid D_{1i} > D_{0i}]$	$E[x_{1i} \mid D_{1i} > D_{0i}]/$ $E[x_{1i}]$
Variable	(1)	(2)	(3)	(4)	(5)
		A. Bernoulli			
Age of second child 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			
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	0.997

Notes: The table reports an analysis of complier characteristics for twins and sex composition instruments. The ratios in columns 3 and 5 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.

Table 3: LATE decompositions

					IV estimates and weighting functions						
Covariate		Covariate pmf			Twins instrument		Samesex instrument		Both instruments		
Age	Education	$P(X)$	$P(X \mid D=1)/P(X)$	$P(X \mid D=0)/P(X)$	$\Delta^z(X)$	$\omega_{\Delta}^{z}(X)$	$\Delta^w(X)$	$\omega_{\Delta}^{w}(X)$	$\Delta^{z,w}(X)$	J-pvalue	$\omega_c(X)$
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(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)		

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

Table 4: Reweighting LATE

			Weeks worked		Employment	
Population	Conditional LATE	Weighting function	Estimate	|t| for diff	Estimate	|t| for diff
(effect)	Δ(X)	ω(X)	(1)	(2)	(3)	(4)
Twins-compliers	twins	twins	-3.15		-0.075	
(LATE)			(0.62)	1.32	(0.014)	0.93
	samesex		-2.71		-0.068	
			(0.81)		(0.018)	
	twins, samesex		-4.19		-0.093	
			(0.53)		(0.012)	
Samesex-compliers	samesex	samesex	-6.30		-0.131	
(LATE)			(1.15)	1.44	(0.026)	0.77
	twins		-5.08		-0.115	
			(1.58)		(0.037)	
	twins, samesex		-4.40		-0.094	
			(0.62)		(0.014)	
Everyone	twins	1	-2.84		-0.067	
(ATE)			(0.76)	1.99	(0.017)	1.58
	samesex		-5.88		-0.123	
			(1.35)		(0.031)	
	twins, samesex		-4.36		-0.092	
			(0.60)		(0.014)	
Treated	twins	P(X|D=1)/P(X)	-2.38		-0.056	
(TOT)			(1.07)	2.85	(0.024)	2.14
	samesex		-7.08		-0.136	
			(1.28)		(0.029)	
	twins, samesex		-4.64		-0.092	
			(0.79)		(0.018)	
Nontreated	twins	P(X|D=0)/P(X)	-3.15		-0.075	
(TNT)			(0.62)	1.15	(0.014)	1.02
	samesex		-5.08		-0.115	
			(1.58)		(0.037)	
	twins, samesex		-4.18		-0.092	
			(0.53)		(0.012)	
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]
	samesex		-3.66		-0.095	
			(1.01)		(0.023)	
			[0.96]		[0.021]	
	twins, samesex		-4.00		-0.101	
			(0.77)		(0.017)	
			[0.62]		[0.014]	

Notes: Standard errors for estimates in parentheses. T-statistics are for the difference between samesex and twins estimates. Standard errors and t-statistics obtained by Brown and Newey (2002) GMM bootstrap with 1,000 repetions. In brackets, we report standard errors and t-statistics obtained by nonparametric bootstrap with 1,000 repetitions. The sample size is 394,840.