Using Instrumental Variables for Inference about Policy Relevant Treatment Parameters*

Magne Mogstad[†] Andres Santos[‡] Alexander Torgovitsky[§]

July 22, 2016

Abstract

We propose a method for using instrumental variables (IV) to draw inference about causal effects for individuals other than those affected by the instrument at hand. The question of policy relevance and external validity turns on our ability to do this reliably. Our method exploits the insight that both the IV estimand and many treatment parameters can be expressed as weighted averages of the same underlying marginal treatment effects. Since the weights are known or identified, knowledge of the IV estimand generally places some restrictions on the unknown marginal treatment effects, and hence on the logically permissible values of the treatment parameters of interest. We show how to extract the information about the average effect of interest from the IV estimand, and more generally, from a class of IV-like estimands which includes the TSLS and OLS estimands, among many others. Our method has several applications. First, it can be used to construct nonparametric bounds on the average causal effects of an actual or hypothetical policy change. Second, our method allows the researcher to flexibly incorporate shape restrictions and parametric assumptions, thereby enabling extrapolation of the average effects for compliers to the average effects for different or larger populations. Third, our method provides specification tests. In addition to testing the null of correctly specified model, we can use our method to test null hypotheses of no selection bias, no selection on gains and instrument validity. Importantly, specification tests using our method do not require the treatment effect to be constant over individuals with the same observables. To illustrate the applicability of our method, we use Norwegian administrative data to draw inference about the causal effects of family size on children's outcomes.

[*] We thank the audience at the 2016 Cowles Foundation Summer Econometrics Conference. Bradley Setzler provided excellent research assistant.

[†] Department of Economics, University of Chicago; Statistics Norway; NBER.

[‡] Department of Economics, University of California at San Diego. Research supported in part by National Science Foundation grant SES-1426882.

[§] Department of Economics, Northwestern University. Research supported in part by National Science Foundation grant SES-1530538. The hospitality of the Department of Economics at the University of Chicago is gratefully acknowledged.

1 Introduction

Influential work by Imbens and Angrist (1994) has provided conditions under which the instrumental variables (IV) estimand can be interpreted as the average causal effect for the subpopulation of compliers, i.e. for those whose treatment status would be affected by an exogenous manipulation of the instrument. In some cases, this local average treatment effect (LATE) is of intrinsic interest, such as when the instrument represents the intervention or policy change of interest.[1] In other settings, it may be reasonable to assume that the causal effect of treatment does not vary systematically with unobservable factors that determine treatment status. Under such an assumption, the causal effect for compliers will be representative of that for the entire population. On the other hand, in many situations, the causal effect for individuals induced to treatment by the instrument at hand might not be representative of the causal effect for those who would be induced to treatment by a given policy change of interest to the researcher. In these cases, the LATE is not the relevant parameter for evaluating the policy change.

In this paper, we show that IV estimators can be informative about the causal effect of a given policy change, even when the LATE is not the parameter of interest. Our setting is the canonical program evaluation problem with a binary treatment $D \in \{0, 1\}$ and a scalar, real-valued outcome Y.[2] Corresponding to the two treatment arms are unobservable potential outcomes, Y_0 and Y_1, that represent the realization of Y that would have been experienced by an individual had their treatment status been exogenously set to 0 or 1. The relationship between Y and D is given by the switching regression

$$Y = DY_1 + (1 - D)Y_0. \tag{1}$$

Following Heckman and Vytlacil (1999, 2005), we assume that treatment is determined by the weakly separable selection or choice equation

$$D = \mathbb{1}[\nu(Z) - U \geq 0] \tag{2}$$

[1] For example, Angrist and Krueger (1991) present IV estimates of the returns to schooling that can be interpreted as LATEs for compliers whose education levels would be affected by compulsory schooling laws.

[2] For discussions of heterogeneous effects IV models with multiple discrete treatments, we refer to Angrist and Imbens (1995), Heckman, Urzua, and Vytlacil (2006), Heckman and Vytlacil (2007b), Heckman and Urzua (2010), Kirkeboen, Leuven, and Mogstad (2015), and Lee and Salanié (2016), among others. Heterogeneous effects IV models with continuous treatments have been considered by Angrist, Graddy, and Imbens (2000), Chesher (2003), Florens, Heckman, Meghir, and Vytlacil (2008), Imbens and Newey (2009), Torgovitsky (2015), Masten (2015), and Masten and Torgovitsky (2016), among others.

where ν is an unknown function, U is a continuously distributed random variable, and Z is a vector of observable regressors. Suppose that Z is stochastically independent of (Y_0, Y_1, U), perhaps conditional on some subvector X of Z that can be thought of as observable control variables, or covariates. Under this assumption, the IV model given by (1)–(2) is equivalent to the IV model used by Imbens and Angrist (1994) and many subsequent authors (Vytlacil, 2002). In particular, the instrument monotonicity condition of Imbens and Angrist (1994) is embedded in the separability of U and Z in the latent index $\nu(Z) - U$. Importantly, the IV model allows the treatment effects $Y_1 - Y_0$ to vary across individuals with the same observable characteristics because U can be statistically dependent with Y_0 and Y_1 conditional on X.

Our goal is to develop a method that uses a random sample of (Y, D, Z) together with the structure of the IV model to draw inference about some parameter of interest, $\beta^\star$, which a researcher has decided is relevant for evaluating an actual or hypothetical intervention or policy change. Our proposed method builds on the work of Heckman and Vytlacil (2005).[3] They argued that various treatment parameters can be unified and interpreted by consideration of the marginal treatment effect (MTE) function, defined as

$$m(u, x) = \mathbb{E}\left[Y_1 - Y_0 | U = u, X = x\right]. \tag{3}$$

The MTE can be interpreted as the average treatment effect indexed as a function of an individual's latent propensity to receive treatment, U, and conditional on other covariates, X. Heckman and Vytlacil (2005) show that common parameters of interest in the program evaluation literature can be expressed as weighted averages of m, with weights that are either known or identified. They show that the same is also true of the IV estimand.

These insights suggest that even in situations where the IV estimand is not of direct interest, it still carries information about the underlying MTE function, and hence about the parameter of interest. In particular, since the weights for both the IV estimand and the parameter of interest are identified, knowledge of the IV estimand generally places some restrictions on the unknown MTE function, and hence on the

[3] See also Heckman and Vytlacil (1999, 2001a,b,c, 2007a,b).

range of values of $\beta^\star$ that are consistent with the data. This can be seen be writing:

$$\begin{aligned}
\underbrace{\beta_{\text{IV}}}_{\text{known IV estimand}} &\equiv \int \underbrace{m(u)}_{\text{unknown MTE}} \times \underbrace{\omega_{\text{IV}}(u)}_{\text{identified IV weights}} \, du \\
\underbrace{\beta^\star}_{\text{unknown target parameter}} &\equiv \int \underbrace{m(u)}_{\text{unknown MTE}} \times \underbrace{\omega^\star(u)}_{\text{identified target weights}} \, du,
\end{aligned} \tag{4}$$

where we are assuming for the moment that there are no covariates X, just for simplicity. Equation (4) suggests that we can extract information about the parameter of interest, $\beta^\star$, from the IV estimand, β_{IV}, by solving an optimization problem. In particular, $\beta^\star$ must be smaller than

$$\max_{m \in \mathcal{M}} \int m(u)\omega^\star(u)\, du \quad \text{such that} \quad \int m(u)\omega_{\text{IV}}(u)\, du = \beta_{\text{IV}}, \tag{5}$$

where the maximum is taken over a set $\mathcal{M}$ of potential MTE functions that incorporates any additional a priori assumptions that the researcher chooses to maintain. Similarly, $\beta^\star$ must be larger than the solution to the analogous minimum problem.

An important practical feature of the abstract optimization problem (5) is that both the objective and constraint are linear functions of the variable of optimization, m. Since m is a function, and therefore an infinite dimensional object, we assume that any $m \in \mathcal{M}$ can be approximated by using a linear basis as

$$m(u) \approx \sum_{k=0}^{K} \theta_k b_k(u), \tag{6}$$

where K is a known positive integer, b_k are known basis functions, and θ_k are unknown parameters. Substituting (6) into (5) renders (5) a finite linear program, at least in the absence of other restrictions on θ. Finite linear programs are convex problems that can be solved quickly and reliably with modern solvers such as CPLEX (IBM, 2010) and Gurobi (Gurobi Optimization, 2015). As a result, the computational burden of our procedure is relatively light, and, importantly, it will always converge.[4]

Equation (6) can be interpreted either as imposing an exact parametric functional form on m, or as approximating m through a finite dimensional sieve as discussed by Chen (2007). We allow for both interpretations. Under the second interpretation, (6) is not restrictive, since wide classes of functions can be approximated arbitrarily well

[4] With the assistance of Bradley Setzler, we are developing a 'push-button' package for R, allowing researchers to apply our method to identify and extrapolate treatment effects as well as to perform specification tests.

as $K \to \infty$ by appropriately chosen linear bases. Under the first interpretation, (6) can be either quite restrictive or quite permissive, depending on the application and subjective opinion. For example, (6) allows a researcher to assume that $K = 0$ and $b_0(u) = 1$, in which case $m(u) = \theta_0$ must be constant, implying no heterogeneity in treatment effects. Similarly, a researcher could assume that the marginal treatment effect is linear, i.e. $m(u) = \theta_0 + \theta_1 u$, as studied by Brinch, Mogstad, and Wiswall (2015). These specifications might be viewed as quite restrictive in many applications. However, they are not inherent to (6), which just as easily allows a researcher to assume (for example) that m is a 10^{th} order polynomial by taking $K = 10$ and $b_k(u) = u^{k-1}$ for $k = 0, 1, \ldots, 10$. Our method allows a researcher to choose K as large as they like, thereby enabling a trade-off between the strength of conclusions they draw and the credibility of the assumptions they maintain.

The optimization problem (5) has only one constraint, corresponding to β_{IV}. Using the same logic, one could also include similar constraints for other IV estimands that correspond to different functions of Z. Upon doing so, the bounds on $\beta^{\star}$ will necessarily become smaller, since each new IV estimand reduces the feasible set in (5). In Section 2.3 we show that, more generally, any cross moment between Y and a known function of D and Z can also be written as a weighted average of MTEs, as in (4). We refer to this class of cross moments as "IV–like" estimands. This class is general enough to contain the estimands corresponding to any weighted linear IV estimator, which includes as special cases the TSLS, optimal generalized method of moments (GMM) and ordinary least squares (OLS) estimands. Each member of this class provides a different weighted average of the same underlying MTE function, and therefore carries some distinct information about the possible values of the parameter of interest, $\beta^{\star}$. In Section 2.4.3 we develop results on how these IV–like estimands can be chosen systematically so as to exhaust the informational content of the model, i.e. to provide sharp bounds on $\beta^{\star}$. An interesting implication of our results is that one can generally obtain stronger conclusions by combining the IV and OLS estimands than could be obtained by using the IV estimand alone. Similarly, two (or more) IV estimates based on different instruments will generally carry independent information about the MTEs, and hence about the parameter of interest.

Our method has several applications, which we discuss in more detail in Section 3. First, it can be used to compute nonparametric bounds on a wide variety of treatment parameters used in the evaluation literature, thereby allowing us to draw inference about the causal effects of a hypothetical or actual policy change. Analytic expressions for sharp bounds have been derived for commonly studied parameters, such as the average treatment effect, by Manski (1989, 1990, 1994, 1997, 2003) and Heckman and Vytlacil (2001b). However, these results provide little guidance to a researcher interested in more complicated parameters, such as the policy relevant treatment effects studied by Heckman and Vytlacil (2001a, 2005) and Carneiro, Heckman, and Vytlacil (2010, 2011). Our general methodology can be used to compute any of these parameters. In addition, our method provides a unified framework for imposing shape restrictions such as monotonicity, concavity, monotone treatment selection (Manski and Pepper, 2000) and separability between observed and unobserved factors in the MTE function (Brinch et al., 2015). Bounds that incorporate these types of assumptions in flexible combinations would be difficult to derive analytically.

Second, our method allows researchers to extrapolate the average effects for compliers to the average effects for different or larger populations, through shape restrictions or parametric assumptions. This aspect of our method generalizes recent work by Brinch et al. (2015). Unlike those authors, we do not presume that we have enough exogenous variation in the data to secure point identification of a parametrically specified MTE function, although we allow this as a special case. As a result, target parameters other than LATEs are typically partially identified. We show through empirical examples that despite this lack of point identification, bounds on parameters other than the LATE can still be tight and informative in interesting cases. An attractive feature of our method is that the constraints in (5) require any feasible MTE function m to yield the same LATE that is already nonparametrically point identified. Hence, our method allows for extrapolation to other parameters of interest without sacrificing the internal validity of the LATE.

Third, our method provides a general framework for testing hypotheses about the IV model. To see this, suppose that the program in (5) is infeasible, so that there does not exist an $m \in \mathcal{M}$ that could have lead to the observed IV estimand. Then the model is misspecified: Either $\mathcal{M}$ is too restrictive, or Z is not exogenous, or the selection equation (2) is rejected by the data, or some combination of the three. That the IV model has testable implications has been observed and studied previously in the case of a binary instrument Z by Balke and Pearl (1997), Imbens and Rubin (1997), Huber and Mellace (2014), Kitagawa (2015), Brinch et al. (2015) and Mourifié and Wan (2015). Our method builds on the setting studied by these authors by allowing Z to be arbitrary and allowing (but not requiring) the researcher to maintain additional assumptions. These additional assumptions could include the parametric and/or shape restrictions previously described. In addition to testing whether the model is misspecified, our method can be used to test the null hypotheses such as no selection bias, no selection on gains, or instrument validity.[5]

Our method takes partial identification of the MTE and target parameter as the typical case and nests point identification as special cases. Heckman and Vytlacil (2005) observe that if Z is continuously distributed and has a sufficiently large impact on treatment choices D so that the propensity score $\mathbb{P}[D = 1|Z = z]$ varies over the entire $[0, 1]$ interval, then the MTE function is point identified and hence any target parameter $\beta^\star$ is also point identified. In practice, however, instruments are often discrete, and many are binary. In such situations, our method helps researchers study the spectrum of information that they can harness about treatment parameters of interest while incorporating a menu of a priori assumptions. In particular, it allows researchers to examine the informational content of the assumptions they use for identification or extrapolation, and the impact of these assumptions on the resulting inferences drawn.

The main technical difficulty raised by allowing for partial identification concerns conducting statistical inference that is asymptotically valid and yet appropriately robust to a sufficiently large class of data generating processes. In Section 5, we provide a method for constructing asymptotically uniformly valid confidence regions for the target parameter. This method is based on naive projection and so, while it does not require any tuning parameters, it tends to be quite conservative. We are in the process of developing less conservative inference methods that more fully exploit the structure of our stochastic linear programming problem. These methods will be included in a future draft of this paper.

In Section 6, we illustrate some of the applications of our method by conducting an empirical assessment of the effects of family size on children's outcomes. To address the problem of selection bias in family size, existing research often use twin births or same-sex sibship as instruments for the number of children. The estimated LATEs suggest that family size has little effect on children's outcomes. In interpreting these results, however, a natural question is whether we are interested in the average causal effects for children whose sibship size is affected by these instruments. An obvious concern is that families induced to have another child because of twin birth or same-sex sibship may differ from families induced to have additional children by a given policy change. In particular, tax and transfer policies would typically affect the households budget constraint, and households would optimally choose family size considering any number of factors. By comparison, everyone who has twins will have another child, whereas the same-sex instrument isolates shifts in family size due to parental preferences for variety in the sex composition.

[5] Tests of no selection on gains have been previously considered in more restrictive settings by Heckman and Schmierer (2010), Heckman, Schmierer, and Urzua (2010) and Angrist and Fernandez-Val (2013).

Instead of assuming that the estimated LATEs of family size are of intrinsic interest, we use the IV estimates to draw inference about the causal effects for different and larger populations. The question of policy relevance and external validity turns on our ability to do this reliably. Applying our method to administrative data from Norway, we find that bounds on treatment parameters other than the LATEs are informative. Additionally, we show that only weak auxiliary assumptions are necessary to extrapolate the average causal effects for a small complier group to the average causal effect for a much larger population.

2 A General Framework for Inference in the IV Model

2.1 The IV Model

Our analysis uses the IV model consisting of (1)–(2). This model is often referred to as the (two-sector) generalized Roy model, and it is a workhorse in labor economics. The observable variables in this model are the outcome Y, the binary treatment D, the vector of observables Z, and a vector of covariates X. Notationally, we regard Z as containing X, which tends to simplify the notation. We denote the support of Z by $\mathcal{Z}$, and we denote the subvector of Z that is not X by Z_0. The Z_0 variables should be thought of as exogenous instruments, whereas X should typically be thought of as consisting of possibly endogenous control variables. The unobservables in the IV model are the potential outcomes (Y_0, Y_1) and the unobservable U that appears in the selection equation (2).

Instead of working with the MTE function (3) directly, we work with the two marginal treatment response (MTR) functions, defined as

$$m_0(u, z) \equiv \mathbb{E}\left[Y_0 \mid U = u, Z = z\right] \quad \text{and} \quad m_1(u, z) \equiv \mathbb{E}\left[Y_1 \mid U = u, Z = z\right]. \tag{7}$$

Of course, each pair $M \equiv (m_0, m_1)$ of MTR functions generates an associated MTE function $m(u, z) \equiv m_1(u, z) - m_0(u, z)$. In general, our method allows the MTR and hence MTE functions to depend on the entire vector Z, including Z_0, in violation of the traditional exclusion restriction. While we will typically impose such an exclusion restriction in applications by only considering MTR functions that do not depend on Z_0, the more general notation is useful for relaxing or testing this restriction.

The following assumptions are maintained throughout the paper and are essential to the analysis.

Assumptions I

I.1 *U is continuously distributed.*

I.2 *$U \perp\!\!\!\perp Z_0 | X$, where $\perp\!\!\!\perp$ denotes (conditional) statistical independence.*

Assumption I.1 is a weak regularity condition. The restrictiveness of Assumptions I comes through I.2, which requires Z_0 to be exogenous to the selection equation. The results of Vytlacil (2002) imply that, given I.2, the assumption that the index of the selection equation is additively separable as in (2) is equivalent to the assumption that Z_0 affects D in the monotonic sense introduced in Imbens and Angrist (1994). Hence, I.2 combined with (2) imposes substantive restrictions on choice behavior. Heckman and Vytlacil (2005, Section 6) show that most of what is known about the IV model depends crucially on these restrictions. We make no attempt to relax them in the current paper.

As is well-known, the function ν in the threshold crossing equation (2) can only be point identified up to a monotonic transformation at best. That is, (2) has the same observable content as the model

$$D = \mathbb{1}[F_{U|Z}(U|Z) \leq F_{U|Z}(\nu(Z)|Z)] \equiv \mathbb{1}[\widetilde{U} \leq F_{U|Z}(\nu(Z)|Z)],$$

where we are using the notation $F_{U|Z}(u|z) \equiv \mathbb{P}[U \leq u | Z = z]$ and we have defined $\widetilde{U} \equiv F_{U|Z}(U|Z)$. Under Assumptions I.1 and I.2, $\widetilde{U} = F_{U|X}(U|X)$ is distributed uniformly over $[0, 1]$, conditional on any realization of Z (or X, which is a subvector of Z). Since working with this normalized model greatly facilitates the analysis without affecting the empirical content of the model, we drop the tilde and maintain throughout the normalization that U itself is distributed uniformly over $[0, 1]$, conditional on Z. A consequence of this normalization is that $p(z) \equiv \mathbb{P}[D = 1|Z = z] = F_{U|Z}(\nu(z)|z) = \nu(z)$, where $p(z)$ denotes the propensity score. Henceforth, we refer to the propensity score $p(z)$ instead of $\nu(z)$.

It is important to observe what is *not* being assumed in our framework under Assumptions I. First, we impose no conditions on the support of Z either here or in the remainder of the paper. Both the control (X) and exogenous (Z_0) components of Z may be either continuous, discrete and ordered, categorical, or binary. Second, the IV model as specified here allows for rich forms of observed and unobserved heterogeneity. In particular, it allows $Y_1 - Y_0$ to vary not only across individuals with different values of X, but also among individuals with the same X. The treatment D may be statistically dependent with Y_0 (indicating selection bias), or $Y_1 - Y_0$ (indicating selection on the gain), or both, even conditional on X. Third, the model does not specify why individuals make the treatment choice that they do, in contrast to the basic Roy model in which $D = \mathbb{1}[Y_1 > Y_0]$. However, it also does not preclude the possibility that individuals choose treatment with full or partial knowledge of (Y_0, Y_1). Any such knowledge will be reflected through the dependence of (Y_0, Y_1) and U.

2.2 What We Want: The Target Parameter

Heckman and Vytlacil (1999, 2005) show that a wide range of treatment parameters can be written for subgroups defined by $Z = z$ as weighted averages of the MTE, i.e. as

$$\beta(z) = \int_0^1 m(u, z)\omega(u, z)\, du, \tag{8}$$

where $\omega(u, z)$ is a weighting function that is either known or identified. In our analysis, we slightly generalize Heckman and Vytlacil (1999, 2005) by writing treatment parameters as the sum of weighted averages of the two MTR functions, m_0 and m_1. This allows us to study parameters that weight m_0 and m_1 asymmetrically. In addition, we allow for the parameter of interest to represent a conditional average over pre-specified subpopulations of Z, whereas Heckman and Vytlacil (1999, 2005) perform all of their analysis conditional on fixed values of covariates $Z = z$.

Formally, we assume that the researcher is interested in a target parameter $\beta^\star$ that can be written for any candidate pair of MTR functions $M \equiv (m_0, m_1)$ as

$$\begin{aligned} \beta^\star \equiv \; & \mathbb{E}\left[\int_0^1 m_0(u, Z)\omega_0^\star(u, Z)\, d\mu^\star(u) \,\middle|\, Z \in \mathcal{Z}^\star \right] \\ + \; & \mathbb{E}\left[\int_0^1 m_1(u, Z)\omega_1^\star(u, Z)\, d\mu^\star(u) \,\middle|\, Z \in \mathcal{Z}^\star \right]. \end{aligned} \tag{9}$$

where $\omega_0^\star$ and $\omega_1^\star$ are identified weighting functions, and $\mathcal{Z}^\star$ is either the entire support of Z, i.e. $\mathcal{Z}$, or a known (measurable) subset of $\mathcal{Z}$ chosen by the researcher. The integrating measure $\mu^\star$ is also chosen by the researcher, and will typically be taken to be Lebesgue measure over $[0, 1]$ as in (8). Our specification of the target parameter $\beta^\star$ nests parameters that can be written in form (8) by taking $\mathcal{Z}^\star = \{z\}$ and $\omega_0^\star(u, z) = -\omega_1^\star(u, z)$. Parameters that average over all covariates Z can be obtained by taking $\mathcal{Z}^\star = \mathcal{Z} \equiv \text{supp}(Z)$.

In Table 1, we provide formulas for the weights $\omega_0^\star$ and $\omega_1^\star$ that correspond to a variety of different treatment parameters, any of which could be taken as the target parameter $\beta^\star$. For example, in the third row, the weights are $\omega_1^\star(u, z) = 1$ and $\omega_0^\star(u, z) = -1$, in which case $\beta^\star$ is the average treatment effect (ATE), conditional on $Z \in \mathcal{Z}^\star$. This is the population average effect of assigning treatment randomly to every individual of type $Z \in \mathcal{Z}^\star$, assuming full compliance. The other rows provide weights that correspond to several other commonly studied treatment parameters. These include (i) the average treatment effect for the treated (ATT), i.e. the average impact of treatment for individuals who actually take the treatment; (ii) the average treatment effect for the untreated (ATU), i.e. the average impact of treatment for individuals who do not take treatment; (iii) LATE$[\underline{u}, \overline{u}]$, i.e. the average treatment effect for individuals who would take the treatment if their realization of the instrument yielded $p(z) = \overline{u}$, but not if it yielded $p(z) = \underline{u}$; and (iv) the policy relevant treatment effect (PRTE), i.e. the average impact on Y (either gross or per net individual shifted) due to a change from the baseline policy to some alternative policy specified by the researcher. We discuss the PRTE further in Section 3.1.

For most of the parameters in Table 1, the integrating measure $\mu^\star$ is taken to be Lebesgue measure on $[0, 1]$. However, researchers are sometimes interested in the MTE function itself, see for example Maestas, Mullen, and Strand (2013) and Carneiro et al. (2011), which both report estimates of the MTEs for various values of u. Our specification of $\beta^\star$ accommodates this by replacing $\mu^\star$ by the Dirac measure (i.e., a point mass) at some specified point $\overline{u}$ and taking $\omega_0^\star(u, z) = -\omega_1^\star(u, z)$. The resulting target parameter is the MTE function averaged over $Z \in \mathcal{Z}^\star$, i.e. $\mathbb{E}[m(\overline{u}, Z) | Z \in Z^\star]$.

2.3 What We Know: IV–Like Estimands

Consider the IV estimand that results from using Z as an instrument for D in a linear instrumental variables regression that includes a constant term, but which does not include any other covariates X. This estimand is given by

$$\beta_{\text{IV}} \equiv \frac{\text{Cov}(Y, Z)}{\text{Cov}(D, Z)}, \tag{10}$$

assuming that $\text{Cov}(D, Z) \neq 0$. For example, if $Z \in \{0, 1\}$ is binary, then β_{IV} reduces to the Wald estimand:

$$\beta_{\text{IV}} = \frac{\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]}{\mathbb{E}[D \mid Z = 1] - \mathbb{E}[D \mid Z = 0]}. \tag{11}$$

Heckman and Vytlacil (2005) show that β_{IV} can also be written in the form (9) as a weighted average of the MTE function, with weights that are identified. This observation forms the foundation for our intuition that useful information about $\beta^\star$ can be extracted from knowledge of β_{IV}. The next proposition shows that, more generally, any cross moment of Y with a known or identified function of (D, Z) can also be expressed as the weighted sum of the two MTR functions m_0 and m_1. We refer to such cross moments as IV–like estimands. A proof of the proposition is contained in Appendix A.

Proposition 1. *Suppose that s is a known or identified, vector-valued function of D and Z that is measurable and has bounded second moments. We refer to such an s as a* IV–like specification *and write $S = s(D, Z)$ for the random vector it generates. For any s, we refer to $\beta_s \equiv \mathbb{E}[SY]$ as an* IV–like estimand*. Then*

$$\beta_s = \mathbb{E}\left[\int_0^1 m_0(u, Z)\omega_{0s}(u, Z)\, du\right] + \mathbb{E}\left[\int_0^1 m_1(u, Z)\omega_{1s}(u, Z)\, du\right], \qquad (12)$$

$$\begin{aligned} \textit{where} \quad & \omega_{0s}(u, Z) \equiv s(0, Z)\mathbb{1}[u > p(Z)] \\ \textit{and} \quad & \omega_{1s}(u, Z) \equiv s(1, Z)\mathbb{1}[u \leq p(Z)]. \end{aligned}$$

It is straightforward to verify that the weights in Proposition 1 reduce to the weights for β_{IV} derived by Heckman and Vytlacil (2005) by taking

$$s(d, z) = \frac{z - \mathbb{E}[Z]}{\text{Cov}(D, Z)}, \qquad (13)$$

which is an identified function of D (trivially) and Z. As we elaborate further in Appendix B, Proposition 1 applies more broadly to include essentially any well-defined weighted linear IV estimand that uses some function of D and Z as included and excluded instruments for a set of endogenous variables also constructed from D and Z.[6] For example, the ordinary least squares (OLS) estimand can be written as an IV–like estimand by taking

$$s(d, z) = \frac{d - \mathbb{E}[D]}{\text{Var}(D)},$$

More generally, the estimands corresponding to the TSLS and optimal GMM estimators can also be written as IV–like estimands. Note that in these cases β_s is typically a vector, with components that correspond to the coefficients on every included regressor.

2.4 From What We Know to What We Want

We now show how to extract information about the target parameter $\beta^\star$ from the general class of IV-like estimands. Throughout this section, we continue to assume that the researcher knows the joint distribution of the observed data (Y, D, Z) without error. We address issues of statistical inference in Section 5.

[6] Here we are using the phrases "included" and "excluded" instrument in the sense typically introduced in textbook treatments of the linear IV model without heterogeneity.

2.4.1 Bounds on the Target Parameter

Let $\mathcal{S}$ denote some finite collection of IV–like specifications (functions s), chosen by the researcher, that each satisfy the conditions set out in Proposition 1. Corresponding to $\mathcal{S}$ is the set of IV–like estimands $\mathcal{B}_{\mathcal{S}} \equiv \{\beta_s : s \in \mathcal{S}\}$ defined as in Proposition 1. Since these IV–like estimands are functions of the observed data, our assumption that the researcher knows the joint distribution of (Y, D, Z) without error implies that they also know $\mathcal{B}_{\mathcal{S}}$ without error. We assume that the researcher has restricted the pair of MTR functions $M \equiv (m_0, m_1)$ to lie in some *admissible space* $\mathcal{M}$. The admissible space incorporates any a priori assumptions that a researcher wishes to maintain about M, such as exclusion or shape restrictions. Our goal is to characterize bounds on values of the target parameter $\beta^\star$ that could have been generated by an $M \in \mathcal{M}$ that could have also delivered the known IV–estimands, $\mathcal{B}_{\mathcal{S}}$.

To this end, we start by denoting the weighting expression in Proposition 1 as a mapping $\tau_s : \mathcal{M} \to \mathbb{R}^{\dim(s)}$, defined for any IV–like specification s as

$$\tau_s(M) = \mathbb{E}\left[\int_0^1 m_0(u, Z)\omega_{0s}(u, Z)\, du\right] + \mathbb{E}\left[\int_0^1 m_1(u, Z)\omega_{1s}(u, Z)\, du\right],$$

where $\dim(s)$ is the dimension of $S \equiv s(D, Z)$ and hence also of β_s. We denote the set of all admissible MTR pairs $M \in \mathcal{M}$ that are consistent with the IV–like estimands $\mathcal{B}_{\mathcal{S}}$ as

$$\mathcal{M}_{\mathcal{S}} \equiv \{M \in \mathcal{M} : \tau_s(M) = \beta_s \ \forall s \in \mathcal{S}\}.$$

In a similar way, we now denote the target parameter more explicitly as a function of M by writing $\beta^\star(M)$. The range of values for the target parameter that are consistent with the set of IV–like estimands $\mathcal{B}_{\mathcal{S}}$ can then be denoted by

$$\mathcal{B}^\star_{\mathcal{S}} \equiv \{\beta^\star(M) : M \in \mathcal{M}_{\mathcal{S}}\}.$$

In other words, $\mathcal{B}^\star_{\mathcal{S}}$ is the image of the target parameter $\beta^\star$ over all $M \in \mathcal{M}_{\mathcal{S}}$, i.e. over the set of all admissible M that could have also generated the observed IV–like estimands $\mathcal{B}_{\mathcal{S}}$. The largest such value of the target parameter, call it $\overline{\beta}^\star$, can be found by solving

$$\overline{\beta}^\star \equiv \sup_{M \in \mathcal{M}} \beta^\star(M) \quad \text{subject to} \quad \tau_s(M) = \beta_s \ \forall s \in \mathcal{S}, \tag{14}$$

where we take $\overline{\beta}^\star = -\infty$ if the program is infeasible, and $\overline{\beta}^\star = +\infty$ if the problem is feasible but unbounded. The smallest such value, $\underline{\beta}^\star$, can be found by solving the analogous minimization problem. The program in (14) is infeasible if and only if $\mathcal{M}_{\mathcal{S}}$ is empty, in which case the IV model is misspecified. We discuss specification tests based on this equivalence in Section 6.3.

Since M is a pair of functions with domain $[0,1] \times \mathcal{Z}$, and is thus an infinite dimensional object, (14) cannot be solved exactly in general. To address this problem, we assume that for every $M \equiv (m_0, m_1) \in \mathcal{M}$, each of m_0 and m_1 can be expressed either exactly or approximately as

$$m_d(u,z) = \sum_{k=1}^{K_d} \theta_{dk} b_{dk}(u,z) \quad \text{for all } u \in [0,1] \text{ and } z \in \mathcal{Z}, \tag{15}$$

for some $\theta_d \equiv (\theta_{d1}, \ldots, \theta_{dK_d}) \in \Theta_d \subseteq \mathbb{R}^{K_d}$, where $\{b_{dk}\}_{k=1}^{K_d}$ are known basis functions that could be different for $d = 0, 1$ if desired, K_0, K_1 are known integers, and Θ_d is a K_d dimensional admissible space specified by the researcher, which we assume is a closed set. Substituting (15) into (14) shows that $\overline{\beta}^\star$ can then be written as the solution to a finite dimensional optimization problem:

$$\begin{aligned} \overline{\beta}^\star &= \max_{\theta \in \Theta} \sum_{k=1}^{K_0} \theta_{0k}\gamma_{0k}^\star + \sum_{k=1}^{K_1} \theta_{1k}\gamma_{1k}^\star \\ \text{subject to } & \sum_{k=1}^{K_0} \theta_{0k}\gamma_{0ks} + \sum_{k=1}^{K_1} \theta_{1k}\gamma_{1ks} = \beta_s \ \forall s \in \mathcal{S}, \end{aligned} \tag{16}$$

where $\theta \equiv (\theta_0, \theta_1)$, $\Theta \equiv \Theta_0 \times \Theta_1$, and we have added the definitions

$$\begin{aligned} \gamma_{dk}^\star &\equiv \mathbb{E}\left[\int_0^1 b_{dk}(u,Z)\omega_d^\star(u,Z)\, d\mu^\star(u) \,\middle|\, Z \in \mathcal{Z}^\star \right] \text{ for } d = 0,1,\ k = 1, \ldots, K_d \\ \text{and } \ \gamma_{dks} &\equiv \mathbb{E}\left[\int_0^1 b_{dk}(u,Z)\omega_{ds}(u,Z)\, du \right] \text{ for } d = 0,1,\ k = 1, \ldots, K_d \text{ and } s \in \mathcal{S}. \end{aligned}$$

Note that since $\omega_d^\star$ and ω_{dks} are identified weighting functions, and b_{dk} are known basis functions, the $\gamma_{dk}^\star$ and γ_{dks} terms are all identified. As a consequence, both $\overline{\beta}^\star$ and the analogous minimum, $\underline{\beta}^\star$, are also identified.

The optimization problem in (16) is finite dimensional, but whether it is linear (or even convex) still depends on the specification of $\Theta \equiv \Theta_0 \times \Theta_1$. If the researcher leaves Θ_d unspecified by setting $\Theta_d = \mathbb{R}^{K_d}$, then (16) is a linear program and can be solved quickly and reliably even if the dimensions of K_d and/or $|\mathcal{S}|$ are quite large. More typically, a researcher may want to place some additional restrictions on Θ_d to ensure that the resulting space of functions satisfy some natural properties such as non-negativity, boundedness or monotonicity. For example, in our empirical application Y is a binary indicator for whether an individual graduates high school, and hence we want to only allow choices of θ such that both m_0 and m_1 are bounded between 0 and 1. As we discuss in Appendix C, boundedness, monotonicity, concavity and some more unusual restrictions can all be imposed by placing linear constraints on θ if the basis functions b_{dk} are taken to be Bernstein polynomials.

One way to interpret the assumption in (15) that M has a linear basis representation is as an approximation to a large class of functions that can be made arbitrarily accurate by taking $K_d \to \infty$. The current discussion presumes that the researcher knows the distribution of (Y, D, Z) without error, in which case the trade-off from taking K_d to be large is only between approximation error and computational costs. Ultimately, however, we want to consider the impact of imperfect knowledge of the distribution of (Y, D, Z) by introducing variation that arises in drawing a random sample. In this case, the trade-off will be between approximation error (or more precisely, bias) and both variance and computational costs, as in the method of sieves (Chen, 2007). We will develop this idea more fully when discussing statistical inference in Section 5.

2.4.2 Computation of Bounds

As discussed in the previous section, (16) can be turned into a linear program either by leaving Θ unconstrained or by selecting an approximating basis such as the Bernstein polynomials for which shape constraints can be imposed through linear restrictions on the basis coefficients. Linear programs are solved routinely in empirical work involving quantile regression (e.g. Abadie, Angrist, and Imbens (2002)), and can be solved about as reliably as matrix inversion. Since we view reliability (i.e. convergence to a solution or determination of no possible solution) as an important quality of a statistical procedure, we constrain ourselves throughout the paper to use only specifications of Θ that render (16) a linear program.

Regardless of whether (16) is a linear program or not, solving it requires one to compute the coefficients $\gamma^{\star}_{dk}$ and γ_{dks}. In general this can be done through one-dimensional numerical integration. However, for all of the cases we consider, the structure of the basis and the weight functions allows us to analytically evaluate these integrals. Hence, computing $\gamma^{\star}_{dk}$ and γ_{dks} merely involves computing sample means of identified functions of Z. We provide more details on the analytical form of the integrals for Bernstein polynomial bases in Appendix C.

2.4.3 Sharpness

The set $\mathcal{M}_{\mathcal{S}}$ represents all MTR functions that are consistent with the set of IV–like estimands $\mathcal{B}_{\mathcal{S}}$ chosen by the researcher. This set does not necessarily exhaust all of the information about M that is contained in the observable data. As a result, $\mathcal{M}_{\mathcal{S}}$ and hence $\mathcal{B}^{\star}_{\mathcal{S}}$ may contain some elements that are inconsistent with other aspects of the data. That is, $\mathcal{M}_{\mathcal{S}}$ and $\mathcal{B}^{\star}_{\mathcal{S}}$ might be non-sharp (or outer) identified sets. In this section we characterize some conditions under which $\mathcal{M}_{\mathcal{S}}$ and hence $\mathcal{B}^{\star}_{\mathcal{S}}$ are actually sharp identified sets. Generally, these conditions require the researcher to choose $\mathcal{S}$ to be a set of functions that is sufficiently rich to exhaust all of the relevant information contained in (Y, D, Z).

As a first step, we show that under fairly weak conditions the identified set $\mathcal{B}^{\star}_{\mathcal{S}}$ resulting from any choice of $\mathcal{S}$ will either be empty or it will be an interval. Hence, assuming that $\mathcal{B}^{\star}_{\mathcal{S}}$ non-empty, every interior point of $[\underline{\beta}^{\star}, \overline{\beta}^{\star}]$ represents a value of the target parameter that is consistent with the observed IV–like estimands $\mathcal{B}_{\mathcal{S}}$.

Proposition 2. *Suppose that $\mathcal{M}$ is convex. Then either $\mathcal{B}^{\star}_{\mathcal{S}}$ is empty, or else the closure of $\mathcal{B}^{\star}_{\mathcal{S}}$ is equal to $[\underline{\beta}^{\star}, \overline{\beta}^{\star}]$.*

To formalize conditions under which $\mathcal{B}^{\star}_{\mathcal{S}}$ is the smallest possible interval, we require some additional notation and concepts. Let F denote a conditional distribution function for $(Y_0, Y_1)|U, Z$, and let $\mathcal{F}$ denote the admissible space of such conditional distribution functions. As with $\mathcal{M}$, the space $\mathcal{F}$ incorporates any a priori assumptions about the conditional distribution of potential outcomes that the researcher wishes to maintain. The sharp identified set for F is defined as

$$\mathcal{F}_{\text{id}} \equiv \Big\{ F \in \mathcal{F} : \mathbb{P}_F[Y \leq y | D = d, Z = z] = \mathbb{P}[Y \leq y | D = d, Z = z] \\ \text{for all } y \in \mathbb{R},\ d \in \{0, 1\}, \text{ and } z \in \mathcal{Z} \Big\},$$

where $\mathbb{P}_F[\cdot | D = d, Z = z]$ denotes the probability of an event, conditional on $D = d$ and $Z = z$, when $(Y_0, Y_1)|U, Z$ has conditional distribution F and (Y, D) are determined through (1)–(2). In words, $\mathcal{F}_{\text{id}}$ is the set of all conditional distributions of potential outcomes that are both admissible and deliver the conditional distribution of observed outcomes. For any $F \in \mathcal{F}$, let

$$M_F(u, z) \equiv (\mathbb{E}_F[Y_0 | U = u, Z = z], \mathbb{E}_F[Y_1 | U = u, Z = z])$$

denote the pair of MTR functions that is generated by F, where $\mathbb{E}_F$ denotes conditional expectation under $\mathbb{P}_F$. The sharp identified set for these MTR functions is defined as

$$\mathcal{M}_{\text{id}} \equiv \{M_F : F \in \mathcal{F}_{\text{id}}\},$$

and the sharp identified set for the target parameter $\beta^\star$ is similarly defined as

$$\mathcal{B}^\star_{\text{id}} \equiv \{\beta^\star(M) : M \in \mathcal{M}_{\text{id}}\}.$$

In terms of these definitions, our objective is to characterize under what circumstances we can conclude that $\mathcal{B}^\star_{\mathcal{S}} = \mathcal{B}^\star_{\text{id}}$. When this is the case, the interval $[\underline{\beta}^\star, \overline{\beta}^\star]$ formed by the minimal and maximal solutions to (14) comprise the best possible bounds on $\beta^\star$ that can be extracted from the data given the maintained assumptions. The following proposition provides one set of conditions under which we can reach this conclusion.

Proposition 3. *Suppose that $Y \in \{0, 1\}$ is binary. For any $d' \in \{0, 1\}$ and $z' \in \mathcal{Z}$ define $s_{d',z'}(d, z) \equiv \mathbb{1}[d \leq d', z \leq z']$. Suppose that $\mathcal{S} = \{s_{d',z'} : d' \in \{0, 1\}, z' \in \overline{\mathcal{Z}}\}$, where $\mathbb{P}[\overline{\mathcal{Z}} \subseteq \mathcal{Z}] = 1$, i.e. $\overline{\mathcal{Z}}$ is a measurable measure 1 subset of $\mathcal{Z}$. Then $\mathcal{B}^\star_{\mathcal{S}} = \mathcal{B}^\star_{id}$.*

Proposition 3 shows that if $\mathcal{S}$ is a sufficiently rich class of functions, then $\mathcal{B}^\star_{\mathcal{S}}$ represents the sharp identified set for $\beta^\star$. For example, if $D \in \{0, 1\}$ and $Z \in \{0, 1\}$ then the set of functions that will lead $\mathcal{B}^\star_{\mathcal{S}}$ to be sharp is

$$\begin{aligned} s_{0,0}(d, z) &= \mathbb{1}[d = 0, z = 0] \qquad & s_{1,0}(d, z) &= \mathbb{1}[z = 0] \\ s_{0,1}(d, z) &= \mathbb{1}[d = 0] & s_{1,1}(d, z) &= 1. \end{aligned}$$

The information contained in the corresponding IV–like estimands is the same as that contained in the coefficients of a saturated regression of Y on D and Z. More generally, if $\mathcal{Z}$ is larger than binary, the set $\mathcal{S}$ may need to be taken to be quite large or even infinite in order to guarantee sharpness. This conclusion is the same as that found in the literature on conditional moment models (e.g. Andrews and Shi (2013)) and follows for the same reasons. There are both computational and statistical reasons why using such a large set of IV–like estimands would be unappealing. Hence in practice a researcher will typically need to balance these considerations against the desire for sharper inference. In most situations one could expect that choosing a finite but well-spaced collection of $s_{d',z'}$ functions will likely provide bounds that are close to sharp, although this type of statement cannot be true in general.

We are in the process of generalizing Proposition 3 to cover cases in which Y is non-binary.

2.4.4 Point Identification

Our method treats partial identification of the MTE and target parameter as the typical case. However, in certain special cases, and depending on the assumptions that are maintained, parts of the MTE function or the target parameter (or both) can be point identified. These cases nest existing results in the literature. For example, Heckman and Vytlacil (1999, 2001c, 2005) and Carneiro et al. (2010, 2011) show that if Z_0 is continuous, then the MTE can be identified over the support of the propensity score by using what they refer to as the local instrumental variables estimator.[7] Consequently, any target parameter $\beta^\star$ whose weighting functions $\omega_d^\star$ have support within the support of the propensity score will be point identified. Those that do not will be partially identified. Brinch et al. (2015) show that if the MTR functions are parameterized as polynomials, then even if Z_0 is only discrete, suitable regressions of Y onto the propensity score can point identify the MTR functions and hence any target parameter $\beta^\star$.

To gain some insight into when these cases are obtained in our framework, we introduce the notation $\boldsymbol{\beta_{\mathcal{S}}} \equiv (\beta_1, \ldots, \beta_{|\mathcal{S}|})'$ for the vector composed of concatenating all IV–like estimands, and $\boldsymbol{\gamma_{ds}} \equiv (\gamma_{d1s}, \ldots, \gamma_{dK_ds})'$ as the vector of coefficients in (16) corresponding to any $d \in \{0, 1\}$ and $s \in \mathcal{S}$. We then define $\boldsymbol{\gamma_d}$ as the $|\mathcal{S}| \times K_d$ matrix with sth row $\boldsymbol{\gamma}'_{\boldsymbol{ds}}$, and let $\boldsymbol{\gamma} \equiv [\gamma_0, \gamma_1]$ and $\boldsymbol{\theta} \equiv [\boldsymbol{\theta}'_0, \boldsymbol{\theta}'_1]$. With this notation, the constraints in (16) can be written more compactly as

$$\boldsymbol{\gamma\theta} = \boldsymbol{\beta_{\mathcal{S}}}, \tag{17}$$

which is a linear system of equations with $|\mathcal{S}|$ equations $K \equiv K_0 + K_1$ unknowns $\boldsymbol{\theta}$.

As is well-known, the linear system (17) either has a unique solution in $\boldsymbol{\theta}$, no solution, or an infinity of solutions. Which case occurs can be determined from the rank of the augmented matrix $[\boldsymbol{\theta}, \boldsymbol{\beta_{\mathcal{S}}}]$, which depends on the number of IV–like estimands, $|\mathcal{S}|$ and the number of parameters, K, as well as any redundancies in $\boldsymbol{\gamma}$ that are introduced by the choice of basis functions and IV–like estimands. In the first case, the feasible set of the program (16) is a singleton if the solution is an element of the admissible space Θ, in which case the MTR functions are point identified, and hence so is any target parameter $\beta^\star$. On the other hand, it may be the case that a unique solution to (17) is not an element of Θ, in which case the program (16) is infeasible, and the IV model is misspecified. This is also true in the second case, regardless of any additional restrictions incorporated in Θ. In the third case, in which there is an

[7] It is straightforward to generalize their results to point identify the MTR functions directly instead of the MTE.

infinity of solutions to (17), $\beta^\star$ will generally be partially identified, except in special cases such as when $\beta^\star$ happens to correspond to one of the IV–like estimands β_s.

2.5 Parametric and Shape Restrictions

Another way to interpret the linear basis representation (15) is to view it as a parametric functional form restriction. For a fixed choice of basis, larger values of K_d represent more flexible functional forms but will necessarily lead to larger optimal values in (16). The researcher's choice of K_d reflects their desire to balance robustness or credibility with the tightness of their conclusions. Recent work by Brinch et al. (2015) has developed sufficient conditions under which this type of basis representation can lead to point identification. Similar parametric assumptions have been employed by French and Song (2014). Kowalski (2016) applies the linear case, studied in depth by Brinch et al. (2015), to an analysis of the Oregon Health Insurance Experiment.

In addition to using (15) as a flexible parametric form (or approximation) for the collection of functions $\mathcal{M}$, one can also impose nonparametric shape restrictions on $\mathcal{M}$. A natural shape restriction to consider is that every $M = (m_0, m_1) \in \mathcal{M}_d$ is such that both m_0 and m_1 takes values within $[\underline{y}, \overline{y}]$, where $\underline{y}$ and $\overline{y}$ are known constants. Such an assumption is implied by a bound on the outcome Y. Most outcome variables in economic applications have a either a logical upper or lower bound (e.g. 0), while some outcome variables, notably binary outcomes, have both. If the specification of M in (15) is viewed as being a nonparametric approximation, and if the supports of the target weights $\omega_1^\star$ and/or $\omega_0^\star$ have some regions that do not overlap with ω_{ds} for any $d = 0, 1$ or $s \in \mathcal{S}$, then such a bounding assumption is usually necessary to ensure that $\underline{\beta}^\star$ and $\overline{\beta}^\star$ are finite. The intuition for this observation is exactly the same as the worst-case bounds of Manski (1989), but here applied to the regions of $[0, 1]$ for which the data places no restrictions on the MTR functions M.

There are other nonparametric shape restrictions that can be imposed on the MTR functions m_0 and m_1, or directly on the MTE function $m = m_1 - m_0$. For example, the monotone treatment response assumption analyzed by Manski (1997) can be imposed by restricting $m_1 - m_0$ to be non-negative, which requires the treatment to have a non-negative effect on Y for all agents. In a similar vein, one could require $m(\cdot, z)$ to be weakly decreasing for every z, which would capture the assumption that those more likely to select into treatment (those will small realizations of U) are also more likely to have larger gains from treatment. This is similar to the "monotone treatment selection" assumption of Manski and Pepper (2000). When using Bernstein polynomials as the approximating basis these types of shape restrictions can all be imposed through linear constraints on the basis coefficients (see Appendix C), which preserves (15) as a linear program.

The specification of $M \in \mathcal{M}$ will typically impose an exclusion restriction with respect to the excluded instrument Z_0. Implementing this is simply a matter of choosing basis functions $b_{dk}(u,z)$ that do not depend on z_0, i.e. $b_{dk}(u,z) = b_{dk}(u,x)$. While nothing in our theory prevents a researcher from not incorporating the exclusion restriction into $\mathcal{M}$, one would expect that the bounds attained from not doing so would be uninformative. Regarding the control variables, X, the linear basis representation (15) provides wide flexibility in the specification of functional form. The issues involved here are the same as in standard linear regression analysis, except that we are specifying $m_d(u,z)$ as a regression in both u and z. Separability between u and z can be imposed by specifying

$$\begin{aligned} & b_{dk}(u,z) = b^U_{dk}(u) \quad \text{for } k = 1,\dots,K^U_d \\ \text{and} \quad & b_{dk}(u,z) = b^Z_{dk}(z) \quad \text{for } k = K^U_d + 1,\dots,K_d, \end{aligned} \tag{18}$$

for some basis functions b^U_{dk} and b^Z_{dk} that (respectively) take only u and z alone as arguments. Separability between u and z has the implication that the slope of the MTR with respect to u does not vary with z. Alternatively, it is straightforward to interact u and z either fully (i.e. with all components) or only partially if full separability is viewed as too strong of a restriction.

3 Applications to Identification, Extrapolation and Testing

This section describes how our method can be used to identify and extrapolate treatment effects as well as to perform specification tests.

3.1 Partial Identification of Policy Relevant Treatment Effects

Heckman and Vytlacil (1999, 2005) consider a class of policies that operate solely on an individual's choice set, and thus affect the probability of treatment (i.e., $p(Z)$) but do not directly affect the MTE. An example from the economics of education is a policy that changes tuition, distance to school, or minimum school leaving age, but does not directly affect the returns to schooling. To evaluate these types of policies, Heckman and Vytlacil (2001a, 2005) introduce a parameter called the policy relevant treatment effect (PRTE) which captures the average causal effect (either aggregate or per net individual affected) on Y of a switch to the new policy from the status quo policy observed in the data. Carneiro et al. (2010, 2011) show that PRTEs corresponding to small policy changes (what they term *marginal* PRTEs) can be point identified if a component of Z_0 is continuously distributed. However, continuous instruments are not available in many applications. Even, when they are, many interesting PRTEs will involve infra– or extra– marginal policy changes. In this section, we show how to use our method to draw inferences on such PRTEs without placing any assumptions on the support of the instruments.

We assume that a policy a can be characterized by a propensity score and instrument pair (p^a, Z^a). Throughout, we maintain the assumption that a policy has no direct effect on the potential outcomes, and in particular that it does not affect the set $\mathcal{M}$ of admissible MTR functions.[8] Treatment choice under policy a is given by

$$D^a \equiv \mathbb{1}[U \leq p^a(Z^a)],$$

where U is the same unobservable term as in the selection equation for the status quo policy, D. The outcome of Y that would be observed under the new policy is therefore given by

$$Y^a = D^a Y_1 + (1 - D^a) Y_0.$$

Carneiro et al. (2010) define the PRTE of policy a_1 relative to another policy a_0 as

$$\frac{\mathbb{E}[Y^{a_1}] - \mathbb{E}[Y^{a_0}]}{\mathbb{P}[D^{a_1} = 1] - \mathbb{P}[D^{a_0} = 1]},$$

which requires the additional assumption that $\mathbb{P}[D^{a_1} = 1] \neq \mathbb{P}[D^{a_0} = 1]$. The PRTE captures the population average effect of this (hypothetical or actual) policy change.

In Table 1, we provide the weights for expressing the PRTE of policy a_1 relative to another policy a_0 in the form of $\beta^\star$. Table 1 also contains the simplified form of these weights that results in the three specific examples of policy comparisons considered by Carneiro et al. (2011). Each of these comparisons is between a hypothetical policy a_1 and the status quo policy a_0, the latter of which is characterized by the pair $(p^{a_0}, Z^{a_0}) = (p, Z)$ observed in the data at hand. The comparisons are: (i) an additive α change in the propensity score, i.e. $p^{a_1} = p + \alpha$; (ii) a proportional $(1 + \alpha)$ change in the propensity score, i.e. $p^{a_1} = (1 + \alpha)p$; and (iii) an additive α shift in the distribution the jth component of Z, i.e. $Z^{a_1} = Z + \alpha e_j$, where e_j is the jth unit vector. The first and second of these represent policies that increase (or decrease) participation in the treated state by a given amount α or a proportional amount $(1 + \alpha)$. The third policy

[8] Heckman and Vytlacil (2005) formalize and discuss the implications of this type of assumption.

represents the effect of shifting the distribution of a variable that impacts treatment choice, such as those described above. In all of these definitions, α is a quantity that could either be hypothesized by the researcher, or estimated from some auxiliary choice model under additional assumptions. Using these weights we can set a given PRTE of interest as the target parameter $\beta^\star$ and directly apply our method to bound it. These bounds can be fully nonparametric, but they can also incorporate a priori parametric or shape restrictions if desired.

Depending on how we specify a_1 and a_0, the PRTE will coincide with alternative treatment parameters used in the evaluation literature. For example, the LATE introduced by Imbens and Angrist (1994) is a leading example of a PRTE. Suppose that Y_0 and Y_1 are mean (or fully) independent of Z, given U, and suppose that there are no covariates X so that $Z = Z_0$. In terms of our admissible set of MTR functions, $M \equiv (m_0, m_1) \in \mathcal{M}$, this amounts to restricting m_0 and m_1 to not vary as functions of z. Then the LATE resulting from a comparison of instrument value $Z = 1$ to $Z = 0$ can written as the PRTE that results from comparing a policy a_1 under which every agent has $Z = 1$ against a policy a_0 under which every agent has $Z = 0$.

To see the link between the LATE and the PRTE, note that treatment choices in policies a_1 and a_0 would be given by

$$D^{a_0} \equiv \mathbb{1}[U \leq p(0)] \quad \text{and} \quad D^{a_1} \equiv \mathbb{1}[U \leq p(1)],$$

where $p(1) > p(0)$ are propensity score values as observed in the data. The PRTE for this policy comparison is given by

$$\frac{\mathbb{E}[Y^{a_1}] - \mathbb{E}[Y^{a_0}]}{\mathbb{E}[D^{a_1}] - \mathbb{E}[D^{a_0}]} = \frac{\mathbb{E}\left[(D^{a_1} - D^{a_0})(Y_1 - Y_0)\right]}{p(1) - p(0)} = \mathbb{E}\left[Y_1 - Y_0 \mid U \in (p(0), p(1)]\right], \quad (19)$$

where we used $D^{a_1} - D^{a_0} = \mathbb{1}[U \in (p(0), p(1)]]$. The right-hand side of (19) is precisely the LATE as defined by Imbens and Angrist (1994).[9] For example, the LATE is the PRTE for a comparison between full implementation (a_1) of a pilot experiment that provided an incentive ($Z = 1$) to take treatment, against the alternative of a status quo (a_0) of not providing such an incentive. Equation (19) shows that the LATE quantity provides the per net person effect on Y from full implementation, which would be an important ingredient in determining whether policy a_1 or a_0 should be preferred.[10]

[9] The derivation leading to (19) is also the same as one that can be used to derive the LATE result. The only minor interpretative difference is that here we are viewing D^{a_1} and D^{a_0} as hypothetical policies rather than potential outcomes.

[10] Unobserved costs or benefits are, of course, other ingredients but this falls outside of the scope of the IV model. We refer to Eisenhauer, Heckman, and Vytlacil (2015) for cost-benefit analysis within the

3.2 Extrapolation of Local Average Treatment Effects

Imbens and Angrist (1994) established that the Wald estimand (11) point identifies the LATE. Our method provides a way to extrapolate from the LATE to average treatment effects for different or larger populations. There are at least two reasons a researcher might be interested in considering such an extrapolation. The first is as a straightforward sensitivity analysis to investigate the fragility or robustness of a given LATE estimate under an expansion (or contraction) of the complier subpopulation. This type of sensitivity analysis can be performed either nonparametrically or parametrically using our general methodology.

The second reason is to consider policies for which the PRTE is equal to a hypothetical LATE that is *not* point identified. For example, suppose that instead of considering full implementation of a pilot experiment as it was conducted, we wish to compare full implementation of a modification of the pilot experiment where the incentive to take treatment has been increased. In particular, suppose that we hypothesize the new incentive will induce participation rates of $p(1) + \alpha$ for some $\alpha \geq 0$. The comparison now is between policies that induce the following treatment choices:

$$D^{a_0} \equiv \mathbb{1}[U \leq p(0)] \quad \text{and} \quad D^{a_1} \equiv \mathbb{1}[U \leq p(1) + \alpha].$$

The per net person PRTE for such a comparison is given by

$$\text{PRTE}(\alpha) \equiv \frac{\mathbb{E}[Y^{a_1}] - \mathbb{E}[Y^{a_0}]}{\mathbb{E}[D^{a_1}] - \mathbb{E}[D^{a_0}]} = \mathbb{E}\left[Y_1 - Y_0 \,|\, U \in (p(0), p(1) + \alpha]\right], \tag{20}$$

which is still a LATE, but one that is not point identified by the IV estimand.

It would be a mistake to conclude that the IV model is uninformative about (20) simply because it is not point identified. As $\alpha \to 0$, the PRTE in (20) will converge to the point identified LATE in (19). Consequently, the strength of the conclusions that can be drawn about (20) will depend on how far the researcher wants to extrapolate, i.e. on the size of α. This will be true regardless of any additional assumptions that are placed on the MTR functions by restricting $\mathcal{M}$. Additional restrictions on $\mathcal{M}$ will tighten the conclusions that can be drawn for any degree of extrapolation α.

Some insight on this process of extrapolation can be gained by decomposing the extrapolated LATE in (20). In particular, suppose that for a given MTR pair $M =$

generalized Roy model.

(m_0, m_1) and its associated MTE $m = m_1 - m_0$ we write for any $\alpha > 0$,

$$\begin{aligned}
\text{PRTE}(\alpha) &= \left(\frac{1}{\alpha + p(1) - p(0)}\right) \int_{p(0)}^{p(1)+\alpha} m(u)\, du \\
&= \left(\frac{p(1) - p(0)}{\alpha + p(1) - p(0)}\right) \left(\frac{\int_{p(0)}^{p(1)} m(u)\, du}{p(1) - p(0)}\right) \\
&\quad + \left(\frac{\alpha}{\alpha + p(1) - p(0)}\right) \left(\frac{\int_{p(1)}^{p(1)+\alpha} m(u)\, du}{\alpha}\right) \\
&= \left(\frac{p(1) - p(0)}{\alpha + p(1) - p(0)}\right) \text{LATE}(p(0), p(1)) \\
&\quad + \left(\frac{\alpha}{\alpha + p(1) - p(0)}\right) \text{LATE}(p(1), p(1) + \alpha). \qquad (21)
\end{aligned}$$

Every term in the final expression in (21) is point identified or known, except for $\text{LATE}(p(1), p(1) + \alpha)$.

The strength of the conclusions that can be drawn about $\text{PRTE}(\alpha)$ depends on how much information the assumptions we maintain provide about $\text{LATE}(p(1), p(1) + \alpha)$. In the extreme case when we maintain no additional assumptions, $\text{LATE}(p(1), p(1)+\alpha)$ could be any real number, and hence $\text{PRTE}(\alpha)$ could also be any real number regardless of how small we take α, as in Manski (1989). The bounds are uninformative, because the MTR functions are allowed to change arbitrarily quickly. Once we are willing to impose some sort of a priori bound on the possible values of Y, the bounds become informative and collapse smoothly to a point as $\alpha \to 0$. Parametric or shape restrictions can be used to tighten the bounds further. Hence, our general framework allows the researcher to transparently tradeoff their preferences for the degree of extrapolation (α) and the strength of their assumptions ($\mathcal{M}$) with the strength of their conclusions, i.e. the width of the resulting bounds.

A natural question is whether, and in what settings, shape restrictions or parametric assumptions can be motivated by or is consistent with economic theory or what is known or typically assumed about technology or preferences. Consider, for example, the literature on production function estimation. Suppose we think of m_j as a production function in state j, with Y_j as the output and Z denoting the observed input factors. Additive separability between observed and unobserved factors in m_j is then implied by perfect substitutability between Z and unobserved inputs factors. By comparison, if output and input factors are measured in logs, additive separability between observed and unobserved factors in m_j is implied by unit-elasticity between observable and unobservable inputs, as in the Cobb-Douglas technology. More generally, additive separability between unobserved and observed factors in m_j is compatible with a production technology in which unobserved productivity differences across agents are factor neutral, which is a standard assumption in methods of production function estimation.

For restrictions other than additive separability, we refer to Chernozhukov et al. (2015). They discuss a number of empirical studies that replace parametric assumptions with shape restrictions implied by economic theory.

3.3 Specification Tests

The program (14) is infeasible if and only if the set $\mathcal{M}_{\mathcal{S}}$ is empty. This indicates that the model is misspecified since there does not exist a pair of MTR functions M that are both admissible ($M \in \mathcal{M}$) and which could have generated the observed data. If $\mathcal{M}$ is a very large, unrestricted class of functions, then this is attributable to a rejection of selection equation (2) together with Assumptions I. That this type of rejection is possible is well-known for the IV model, see e.g. Kitagawa (2015). On the other hand, if other restrictions have been placed on $\mathcal{M}$, then misspecification could be due either to the failure of Assumptions I, or to the rejection of restrictions in $\mathcal{M}$, or both. Hence, the infeasibility of (14) serves as an omnibus specification test. What the researcher concludes from rejecting the hypothesis correct specification depends on how strongly the researcher believes different aspects of the assumptions. A rejection could also arise from the type of sampling error that we have abstracted from so far. In Section 5, we explicitly allow for such errors by measuring how closely the constraint set is to being infeasible, in a sense which we make more precise in that discussion.

Our framework can be used to test other interesting hypotheses besides model misspecification. For example, Table 1 reports the weights that correspond to the quantity $\mathbb{E}[Y_0|X, D = 1] - \mathbb{E}[Y_0|X, D = 0]$. This quantity is often considered as a measure of selection bias, since it captures the extent to which average untreated outcomes differ solely on the basis of treatment status, conditional on observables. As with any of the other quantities in Table 1, we can set the target parameter $\beta^\star$ to be average selection bias and then construct bounds on it. If these bounds do not contain 0, then we reject the null hypothesis that there is no selection bias. The hypothesis of no selection on the unobserved gain from treatment can be tested in a similar way.

Alternatively, one may be interested in testing the joint null hypothesis that there is no selection on unobservables; that is, no selection bias and no selection on gain. While this is no longer a hypothesis concerning only a scalar quantity, we can harness the general specification test to evaluate this hypothesis as well. To see this, let $\beta^\star_{\text{sel}}(M)$

denote the average selection bias for MTR pair M, and let $\beta^{\star}_{\text{gain}}(M)$ denote the average selection on the gain. Then replace $\mathcal{M}$ with the smaller subset $\mathcal{M}_{\text{test}} \equiv \{M \in \mathcal{M} : \beta^{\star}_{\text{sel}}(M) = \beta^{\star}_{\text{gain}}(M) = 0\}$. If (14) is infeasible when $\mathcal{M}$ is replaced by $\mathcal{M}_{\text{test}}$, and if neither $\mathcal{M}$ nor Assumptions I are suspected to be the cause of this infeasibility, then the researcher can reject the hypothesis that there is both no selection bias and no selection on the gain.

Using a similar idea, one can also evaluate the null hypothesis that Y_0 and Y_1 are mean independent of Z_0, conditional on X and U. This is the minimal exclusion condition typically imposed in analysis of mean marginal treatment effects (Heckman and Vytlacil, 2005). There are many ways to implement such a test. If Z_0 is binary, then the correct specification of the model can be tested when adding the constraints that $\mathbb{E}[Y_d|U, X, Z_0 = 1] - \mathbb{E}[Y_d|U, X, Z_0 = 0] = 0$ for $d = 0$ and 1 to the definition of $\mathcal{M}$, just as in the joint test of no selection bias and no selection on the gain discussed above. For distributions of Z_0 that have more support points, one could perform a collection of pairwise comparisons, or weight the absolute value of multiple comparisons together in any of a variety of ways.[11] Regardless of how one measures a violation of the exclusion restriction, it is important to remember to specify $\mathcal{M}$ so that it does *not* restrict the MTR functions to be invariant to the Z_0 component of Z. Such a restriction imposes precisely the exclusion restriction that is being tested, and hence the misspecification test will always fail to reject by force of assumption, at least as long as there are no other sources of misspecification.

4 Numerical Illustration

We illustrate the procedure described in the previous sections with the following data generating process:

$$\begin{aligned} Y_0 &= \mathbb{1}[U \leq V_0] & (V_0, V_1) &\sim N([-.2, .4], I_2) \\ Y_1 &= \mathbb{1}[U \leq V_1] & U &\sim \text{Unif}[0,1],\ U \perp\!\!\!\perp (V_0, V_1) \\ D &= \mathbb{1}[U \leq p(Z)] & p(z) &\equiv \Phi(z - 1), \end{aligned} \tag{22}$$

where Φ is the cumulative distribution function of a standard normal distribution and I_2 is the two dimensional identity matrix. The first distribution of Z that we consider is discrete with support $\mathcal{Z} \equiv \{-.5, -.4, \ldots, .4, .5\}$ having 11 elements. The density

[11] Note that constraints involving absolute values of linear functions can be reformulated as a collection of standard linear constraints, see e.g. Bertsimas and Tsitsiklis (1997).

(with respect to counting measure) of Z is set to

$$f_Z(z_l) = \frac{\Phi_Z(z_l) - \Phi_Z(z_{l-1})}{\Phi_Z(z_L)}, \tag{23}$$

where z_l is the lth largest element of $\mathcal{Z}$ (with $z_0 \equiv -\infty$ and $z_L \equiv .5$), and Φ_Z is the cumulative distribution function for a $N(0, .3^2)$ distribution. Since $Y \in \{0, 1\}$, we restrict $\mathcal{M}$ to contain only pairs of MTR functions that are bounded below by 0 and above by 1. The MTR functions implied by this data generating process are given by

$$m_d(u) = 1 - \Phi\left(u - \mathbb{E}[V_d]\right). \tag{24}$$

for $d = 0, 1$.

We take the target parameter to be the average treatment on the treated (ATT), i.e. $\mathbb{P}[Y_1 - Y_0|D = 1]$, the weights for which are given in Table 1. Figure 1a plots these weights (i.e. $\omega^\star_1$) against the weights for the IV–like estimand corresponding to the IV estimator that uses $p(Z)$ as an instrument for D. The weights for this IV–like estimand are given through Proposition 1 with

$$s(d, z) = \frac{p(z) - \mathbb{E}[p(Z)]}{\text{Cov}(D, p(Z))}. \tag{25}$$

Note that we include the denominator for this exercise because it ensures the weights integrate to 1, but as previously noted it can be omitted without impacting $\mathcal{M}_\mathcal{S}$ or $\mathcal{B}^\star_\mathcal{S}$. Heckman and Vytlacil (2005) showed that for any instrument Z, using $p(Z)$ as the instrument generates IV weights that have the largest possible support. Evidently, for this DGP this support still does not completely overlap with that of $\omega^\star_1$, which places much of its weight on u near 0. Consequently, one should expect that, barring any additional assumptions, the ATT will not be point identified. The true value of the ATT (i.e., the value implied by (22)) is approximately .235.

We model m_0 and m_1 with the Bernstein polynomials (see Appendix C) and consider different choices of basis length $K \equiv K_0 = K_1$, i.e. of order $K - 1$. Columns (1)–(4) of Table 2a report the bounds on the ATT ($\underline{\beta}^\star$ and $\overline{\beta}^\star$) derived from solving (16) for increasing values of K when $\mathcal{S}$ contains only the specification s defined in (25). The bounds increase with K, since this increases the number of variables in the program (16). More intuitively, the more flexible the parametric specification is, the fewer assumptions are imposed on $\mathcal{M}$, and the larger the bounds. The bounds can be viewed as approaching nonparametric as $K \to \infty$.

In our numerical illustration, the specification of K materially affects the size of the bounds on the ATT. Figure 2a offers some intuition, reporting the MTE function obtained at the lower and upper bounds when $K = 5$ (column (2) of Table 2a). The lower bound function, displayed in solid blue, tries to become as small as possible on the region near 0 where the ATT weights are large, c.f. Figure 1a. At the same time, this function needs to still achieve a certain value of the IV–like estimand β_s, and so must be sufficiently large over the region where the weights for β_s are non-zero. The optimal lower bound function tries to decrease as quickly as possible to 0 while still maintaining this restriction. The larger more flexible that K is, the faster it can plunge, since larger values of K allow for functions that adjust more quickly. For comparison, Figure 2b shows the same plot with $K = 15$ (column (4) of Table 2a). The dotted blue line in both figures shows an MTE function that delivers a value of the ATT that is smaller than $\underline{\beta}^{\star}$, but which does not deliver the required value of the IV–like estimand β_s, and is therefore logically inconsistent with the observed data. Analogous intuition applies to the optimal and infeasible MTE functions for the upper bound, which are shown in red. Note that because we are optimizing functionals of the MTR functions M, rather than the MTE at a given point, the lower and upper bound MTE functions can cross, as they do in Figures 2a–2b.

Columns (5)–(8) of Table 2a report analogous bounds when $\mathcal{S}$ is expanded to also contain the specification of s in (13) that uses Z itself as the instrument. As expected, the bounds are uniformly smaller than for columns (1)–(4), since more restrictions are maintained in the optimization problem (16). In fact, with $K = 2$, $\mathcal{B}^{\star}_{\mathcal{S}}$ shrinks to a singleton, since there is only one choice of θ that is consistent with both β_s in $\mathcal{B}_{\mathcal{S}}$. This singleton is not equal to the true value of the ATT (.235) because the model is misspecified: The true MTR functions in (24) is not an order 1 (linear) polynomial. Intuitively, using a larger set of specifications corresponds to exploiting more of the observed data. This shrinks the class $\mathcal{M}_{\mathcal{S}}$ of MTR functions that could have generated the observed IV–like estimands, and therefore also shrinks $\mathcal{B}^{\star}_{\mathcal{S}}$.

Columns (9)–(12) of Table 2a push this observation further by including specifications that correspond to IV–like estimands defined analogously to (25) but with different functions of Z. The weights corresponding to these additional specifications are shown in Figure 1b. The impact on the bounds is so dramatic that even with $K = 5$ the bounds are empty, i.e. the program in (16) is infeasible. This yields two observations. First, the five specifications included in $\mathcal{S}$ have tremendous identifying content for the ATT. Second, the true MTE function corresponding to the data generating process, while nearly linear (see Figure 2a), is actually the difference of two normal cumulative distribution functions (see (24)) and so does not have an exact representation as a Bernstein polynomial with any finite number of terms K. Hence, not only do the bounds derived from (16) have substantial identifying content, they can also be used to falsify a given parametric specification, as suggested in Section 6.3. When K is increased to 10 and 15, the parametric form becomes sufficiently flexible as to be able to rationalize all of the specifications in $\mathcal{S}$ simultaneously. However, the set of MTR functions that deliver all of these IV–like estimands simultaneously is sufficiently restricted as to be logically consistent with only a small range of values of ATT, thereby leading to very tight bounds.

Table 3a repeats the exercise of Table 2a when the support of Z is reduced to the three element set $\mathcal{Z} \equiv \{-.5, 0, .5\}$. The data generating process remains otherwise unchanged, although this modification of the distribution of Z changes the ATT slightly (at the fourth digit) through its compositional effect on which individuals select into treatment. Figure 3a provides a counterpart to Figure 1a for this case, and shows that the weights are constant over larger regions. The effect that this has on the bounds varies considerably depending on which column one considers. Perhaps surprisingly, in some columns this reduction of the support markedly shrinks the bounds. This is likely due to the fact that as the support is shrunk while keeping the density of Z at (23), there will be relatively more individuals with $Z = -.5$, i.e. with a propensity score value towards 0. The interesting takeaway is that having an instrument with more points of support is not necessarily beneficial in terms of empirical content.

In Table 3b we instead modify the support of Z to be $\mathcal{Z} \equiv \{0, .1, \ldots, .9, 1\}$, i.e. we add .5 to each support point in the specification for Table 2a. This has the effect of increasing the support of the propensity score and consequently shifts the weights for β_s when s is given by (25) to be farther away from those for the ATT, c.f. Figures 1a and 3b. As one would expect, the bounds on the ATT become wider, reflecting the fact that the parameter being extrapolated to has become increasingly different from those that are being estimated. Intuitively, the process of extrapolating is easier when the degree of extrapolation required is smaller.

5 Statistical Inference

While our discussion so far has centered on population parameters, we now turn to the problem of statistical inference on these parameters of interest. Before proceeding, however, it will prove helpful to introduce some additional notation. Towards this end we define $\gamma_{0s} = (\gamma_{01s}, \ldots, \gamma_{0K_0s})'$ and $\gamma_{1s} = (\gamma_{11s}, \ldots, \gamma_{1K_1s})'$ for any $1 \leq s \leq |\mathcal{S}|$ and set $\gamma_s = (\gamma_{0s}', \gamma_{1s}')'$. Analogously letting $\gamma_0^\star = (\gamma_{01}^\star, \ldots, \gamma_{0K_0}^\star)'$, $\gamma_1^\star = (\gamma_{11}^\star, \ldots, \gamma_{1K_1}^\star)'$, and $\gamma^\star = ((\gamma_0^\star)', (\gamma_1^\star)')'$, we then note the linear program in (16) may be written as

$$\bar{\beta}^\star = \max_{\theta\in\Theta}\{(\gamma^\star)'\theta \text{ s.t. } \gamma_s'\theta = \beta_s \text{ for all } s \in \mathcal{S}\} \tag{26}$$

with a parallel expression defining $\underline{\beta}^\star$. As previously discussed, the unknown population quantities in (26) comprise of $\gamma^\star$, γ_s, and β_s for all $1 \le s \le |\mathcal{S}|$, for which we assume:

Assumptions II

There are $(\hat{\gamma}^\star, \{\hat{\gamma}_s, \hat{\beta}_s\}_{s\in\mathcal{S}})$ *and bootstrap analogues* $(\hat{\gamma}^{\star,*}, \{\hat{\gamma}_s^*, \hat{\beta}_s^*\}_{s\in\mathcal{S}})$ *such that*

II.1 $(\sqrt{n}(\hat{\gamma}^\star - \gamma^\star), \{\sqrt{n}(\hat{\gamma}_s - \gamma_s), \sqrt{n}(\hat{\beta}_s - \beta_s)\}_{s\in\mathcal{S}}) \xrightarrow{L} \mathbb{G}_0$ *for a nondegenerate Gaussian random variable* $\mathbb{G}_0$ *in* $\mathbb{R}^{K_0+K_1} \times \mathbb{R}^{|\mathcal{S}|\times(K_0+K_1+1)}$.

II.2 $(\sqrt{n}(\hat{\gamma}^{\star,*} - \hat{\gamma}^\star), \{\sqrt{n}(\hat{\gamma}_s^* - \hat{\gamma}_s), \sqrt{n}(\hat{\beta}_s^* - \hat{\beta}_s)\}_{s\in\mathcal{S}}) \xrightarrow{L} \mathbb{G}_0$ *conditional on the data.*

The unknown quantities in (26) may be estimated by a variety of different methods, including both parametric and nonparametric approaches for estimating the terms $\gamma^\star$ and γ_s for $1 \le s \le |\mathcal{S}|$. Rather than taking a stand on a particular modeling choice, we therefore simply impose the requirement that an estimator be available in Assumption II.1 and that we be able to bootstrap their asymptotic distributions in Assumption II.2. It is worth noting that Assumption II.1 in fact immediately implies that the sample analogues to $\bar{\beta}^*$ and $\underline{\beta}^\star$ are consistent provided that the population constraints set have non-empty interiors. Regrettably, while estimation is straightforward, inference is more challenging. In particular, a naive "plug-in" bootstrap approach fails due to $\beta^\star$ not being (fully) differentiable in the constraints; see Shapiro (1989) for the lack of differentiability and Fang and Santos (2014) for the implication of bootstrap failure.

Assumptions II.1 and II.2 nonetheless enable us to establish the validity of a simple but admittedly conservative method for constructing a $1-\alpha$ confidence region for any $\beta^\star \in \mathcal{B}_{\mathcal{S}}^\star$. Specifically, let $\hat{c}_{1-\alpha}$ denote the $1-\alpha$ quantile conditional on the data of the expression

$$\sup_{\theta\in\Theta} \max\left\{|\sqrt{n}(\hat{\gamma}^{\star,*} - \hat{\gamma}^\star)'\theta|, \max_{1\le s\le|\mathcal{S}|}\left|\sqrt{n}(\hat{\gamma}_s^* - \hat{\gamma}_s)'\theta - \sqrt{n}(\hat{\beta}_s^* - \hat{\beta}_s)\right|\right\} \; . \tag{27}$$

We observe that, assuming (16) is a linear program, (27) can also be written as a linear program in θ. Consequently, computing $\hat{c}_{1-\alpha}$ just requires us to solve a linear program for each bootstrap iteration and then obtain the $1-\alpha$ empirical quantile of the optimal value across these iterations. Given $\hat{c}_{1-\alpha}$ we may then obtain the upper endpoint of a confidence region for $\beta^\star$ by solving the optimization problem

$$\bar{c}_n \equiv \sup_{\lambda\in\mathbb{R},\theta\in\Theta} \lambda \quad \text{subject to} \quad \begin{array}{ll} \text{(i)} & -\frac{\hat{c}_{1-\alpha}}{\sqrt{n}} \leq \hat{\gamma}_s'\theta - \beta_s \leq \frac{\hat{c}_{1-\alpha}}{\sqrt{n}} \\ \text{(ii)} & -\frac{\hat{c}_{1-\alpha}}{\sqrt{n}} \leq (\hat{\gamma}^\star)'\theta - \lambda \leq \frac{\hat{c}_{1-\alpha}}{\sqrt{n}} \end{array} . \tag{28}$$

This is also a linear program whenever (16) is a linear program. In turn, we obtain a lower endpoint for our confidence region by solving the following

$$\underline{c}_n \equiv \inf_{\lambda\in\mathbb{R},\theta\in\Theta} \lambda \quad \text{subject to} \quad \begin{array}{ll} \text{(i)} & -\frac{\hat{c}_{1-\alpha}}{\sqrt{n}} \leq \hat{\gamma}_s'\theta - \beta_s \leq \frac{\hat{c}_{1-\alpha}}{\sqrt{n}} \\ \text{(ii)} & -\frac{\hat{c}_{1-\alpha}}{\sqrt{n}} \leq (\hat{\gamma}^\star)'\theta - \lambda \leq \frac{\hat{c}_{1-\alpha}}{\sqrt{n}} \end{array} . \tag{29}$$

In the next proposition we establish that the resulting confidence region provides valid coverage.

Proposition 4. *If Assumption II.1, II.2 hold, Θ is a bounded set, and $\mathcal{B}^\star_\mathcal{S} \neq \emptyset$, then*

$$\liminf_{n\to\infty} P(\underline{c}_n \leq \beta^\star \leq \bar{c}_n) \geq 1-\alpha \tag{30}$$

for any $\beta^\star \in \mathcal{B}^\star_\mathcal{S}$.

It is of course possible that in attempting to compute $\underline{c}_n$ and $\bar{c}_n$ we find that the constraint sets in (28) and (29) are empty. Such an event provides evidence rejecting the proper specification of the model – formally, a specification test that rejects whenever the constraint sets in (28) and (29) are empty controls size at level α. We also note that the analysis in Proposition 4 is pointwise in the underlying distribution of the data, but trivially holds uniformly among any class of distributions for which Assumptions II.1 and II.2 hold uniformly. On the other hand, we expect that the confidence region in Proposition 4 may be quite conservative. We are currently developing a different, less conservative procedure that more directly exploits the geometry of our linear programming problem.

6 Empirical Application

This section illustrates one of the several applications of our method, namely extrapolation of the average causal effects for compliers to the average causal effect for larger populations. In particular, we use Norwegian administrative data to draw inference about the causal effects of family size for individuals other than those affected by the available instruments.

6.1 Motivation and Related Literature

Motivated by the seminal Becker and Lewis (1973) quantity-quality (QQ) model of fertility, a large and growing body of empirical research has examined the effects of family size on child outcomes. Much of the early literature that tested the QQ model found that larger families reduced observable measures of child quality, such as educational attainment. However, recent studies from several developed countries have used instruments, such as twin births and same-sex sibship, to address the problem of selection bias in family size. The estimated LATEs suggest that family size has little effect on children's outcomes. For example, the widely cited study of Black, Devereux, and Salvanes (2005) uses administrative data from Norway and concludes "there is little if any family size effect on child education" (p.697).[12]

In interpreting these findings, a natural question is whether we are interested in the average causal effects for children whose sibship size is affected by these instruments. An obvious concern is that families induced to have another child because of twin birth or same-sex sibship may differ from families induced to have additional children by a given policy change. In particular, tax and transfer policies would typically affect the households budget constraint, and households would optimally choose family size considering any number of factors. By comparison, everyone who have twins will have another child, whereas the same-sex instrument isolates shifts in family size due to parental preferences for variety in the sex composition.

Arguing that the estimated LATEs of family size may not be the parameters of interest, Brinch et al. (2015) re-examine the analysis by Black et al. (2005). Imposing functional structure on $m(u, x)$, they can point identify the MTEs of family size. Their findings suggest the effects of family size are both more varied and more extensive than what the LATEs imply. It is not clear, however, whether this conclusion is warranted, as the functional structure Brinch et al. (2015) impose aids identification by allowing interpolation between different values of $P(Z)$ in the data or extrapolation beyond the support of $P(Z)$ given X.

The goal of our empirical analysis is to offer a middle ground between the polar cases of i) only reporting the LATEs, with their high degree of internal validity but limited external validity, and ii) invoking the full set of assumptions necessary to point identify the MTEs (and hence the treatment effects for the entire population).

[12] Using data from the US and Israel, Caceres-Delpiano (2006) and Angrist, Lavy, and Schlosser (2010) come to a similar conclusion.

6.2 Data and Summary Statistics

As in Black et al. (2005) and Brinch et al. (2015), our data are based on administrative registers from Statistics Norway covering the entire resident population of Norway who were between 16 and 74 of age at some point during the period 1986-2000. The family and demographic files are merged by unique individual identifiers with detailed information about educational attainment reported annually by Norwegian educational establishments. The data also contains identifiers that allow us to match parents to their children. As we observe each child's date of birth, we are able to construct birth order indicators for every child in each family. We refer to Black et al. (2005) for a more detailed description of the data as well as of relevant institutional details for Norway.

We follow the sample selection used in Black et al. (2005) and Brinch et al. (2015). We begin by restricting the sample to children who were aged at least 25 in 2000 to make it likely that most individuals in our sample have completed their education. Twins at first parity are excluded from the estimation sample because of the difficulty of assigning birth order to these children. To increase the chances of measuring completed family size, we drop families with children aged less than 16 in 2000. We exclude a small number of children with more than 5 siblings as well as a handful of families where the mother had a birth before she was aged 16 or after she was 49. In addition, we exclude a few children where information about the characteristics of the mother is missing.

As in Black et al. (2005) and Brinch et al. (2015), our measure of family size is the number of children born to each mother. Throughout the empirical analysis, we follow much of the previous literature in focusing on the treatment effect on a first born child from being in a family with 2 or more siblings rather than 1 sibling. The outcome of interest is the whether the child completed high school or dropped out. We are in the process of accessing data with information on more outcome variables. This will be included in a future draft of the paper.

In line with much of the previous literature on family size and child outcomes, we consider the following two instruments: twin birth and same-sex sibship. The twins instrument is a dummy for a multiple second birth (2nd and 3rd born children are twins). The validity of this instrument rests on the assumptions that the occurrence of a multiple birth is as good as random, and that a multiple birth affects child development solely by increasing fertility. The same-sex instrument is a dummy variable equal to one if the two first children in a family have the same sex. This instrument is motivated by the fact that parents with two children are more likely to have a third child if the first two are of the same sex than if sex-composition is mixed. The validity of the same-sex instrument rests on the assumptions that sibling sex composition is essentially random and that it affects child development solely by increasing fertility. It should be emphasized that our focus is not on the validity of these instruments: Our aim is to move beyond the LATE of family size, applying commonly used instruments. We refer to Black et al. (2005) and Angrist et al. (2010) for empirical evidence in support of the validity of the instruments.

Our sample consists of 514,004 first-born children with at least one sibling. Table 5 displays basic descriptive statistics. In 50 percent of the sample, there are at least three children in the family, and the average family size is 2.7 children. There are a few noticeable differences between children from a family with only one sibling as compared to those with two or more siblings. As expected, parents with two children are more likely to have a third child if the first two are of the same sex than if sex-composition is mixed. Furthermore, the second and third born children are twins in about 1 percent of the families. It is also evident that first born children with one sibling have higher educational attainment than first born children with two or more siblings, pointing to a negative association between family size and child quality. Specifically, 68 percent of children with one sibling completes high school while the completion rate is only 60 percent among children with more two ore more siblings. However, mothers with more children are also different. For instance, they are younger at first birth. Differences in observables by family size suggest we need to be cautious in giving the correlation between number of children and child quality a causal interpretation.

6.3 Empirical Results

We begin by presenting OLS and IV estimates of the effect on a first born child from being in a family with 2 or more siblings ($D = 1$) rather than 1 sibling ($D = 0$), with and without covariates (X). Table 5 display the results. For now, we only use the same-sex sibship instrument (Z_0). In the specification with covariates, we saturate the regression models allowing for a separate treatment effect for every possible combination of values that X can take on. To obtain the unconditional LATE, we compute the complier weighted average of the covariate-specific LATEs. For now, standard errors are constructed by bootstrap. As expected, the OLS estimates suggest that larger family size reduces children's educational attainment, as measured by high school completion. However, controlling for observables lower this estimate, and once we instrument for family sizen the point estimate is close to zero.

Our method exploits that both the IV estimand and many treatment parameters of interest can be expressed as weighted averages of the same underlying MTEs. A natural first step is therefore to estimate these weights. Figure 9 displays the estimated distribution of weights for the LATE, the ATT, the ATE, and the ATUT. The formulas for these weights are provided in Table 4. The y-axis measures the density of the distribution of weights, whereas the x-axis measures the unobserved component U of parents' net gain from having 3 or more children rather than 2 children. Recall that a high value of U means that a family is less likely to have 3 or more children. There are clear patterns in the distribution of weights. IV estimates using the same-sex instrument is based on a small number of compliers, assigning weights only to the MTEs of families with values of U between 0.471 and 0.529. By comparison, the ATE is a simple unweighted average of all MTEs, whereas the ATT assigns more weight to the MTEs of families who are likely to have another child while the ATUT assigns most of the weight to MTEs of families unlikely to have another child.

The local nature of the same-sex LATE raises concerns about its external validity and policy relevance. To address these concerns, we can use our method to investigate the fragility or robustness of the LATE to expanding the complier subpopulation. This type of analysis can performed non-parametrically or parametrically. In our analysis, we model m_0 and m_1 with Bernstein polynomials and consider three linear, cubic and quintic choices for basis length K. As in Table 5, we estimate separate bounds for every possible combination of values X can take on. We then obtain bounds on the unconditional treatment parameters by taking the complier weighted average of the covariate-specific bounds.

Figure 7 presents upper and lower bounds on what the LATE would have been if we increase the size of the complier group. We begin by expanding the complier group to include families with values of U between 0.466 and 0.524. This amounts to adding 1 percent of the population to the complier group, increasing its size by 17 percent. We then compute the bounds on the LATE for this new and larger set of families. Next, we add another percent of the population to the complier group, and recompute the bounds. We continue in this fashion, until we have added nearly 23 percent of the population, increasing the size of the complier group by 400 percent.

The strength of the conclusions that can be drawn from this extrapolation depends on two aspects. First, bounds based on the class of IV-like estimates are more informative than those that are based on the IV estimate alone. In the upper graph of Figure 7, we only use the IV estimate in the extrapolation. By comparison, the lower graph exploits our theoretical prediction that both the OLS and IV estimate carry independent information about the MTE, which helps to sharpen the bounds considerably.

Secondly, the weaker the restrictions that are imposed on m_0 and m_1, the larger the bounds. However, even with a flexible quintic function for K, the bounds remain quite informative. For example, if we double the size of the complier group, the average causal effect is bounded between approximately 0 (lower bound)and 0.01 (upper bound) (see the lower graph of Figure 7). Taken together, the results in Figure 7 highlight how our method allows researcher to transparently tradeoff their preferences for the degree of extrapolation (α) and the strength of their assumptions with the strength of their conclusions, i.e. the width of the resulting bounds.

Figure 8 complements by exploring the identifying content of alternative specifications of $\mathcal{S}$. In this figure, we focus attention on the bounds based on the quintic function for K, The upper graph compares the bounds based on the same-sex IV estimate alone to what one can learn from the joint distribution of the outcome, the treatment, and the same-sex instrument. The lower graph examines how the bounds change by including both the same-sex IV estimate and the TSLS estimate using twins, same-sex and the interaction as instruments. Both graphs shows the usefulness of taking advantage of data in addition to the same-sex IV estimate.

7 Conclusion

In this paper, we proposed a method for using IV to draw inference about causal effects for individuals other than those affected by the instrument at hand. The question of external validity and policy relevance turns on our ability to do this reliably. Our method exploits the observation that both the IV estimand and many treatment parameters can be expressed as weighted averages of the same underlying marginal treatment effects. Since the weights are known or identified, knowledge of the IV estimand generally places some restrictions on the unknown marginal treatment effects, and hence on the logically permissible values of the treatment parameters of interest. We showed how to extract the information about the average effect of interest from the IV estimand, and more generally, from a class of IV-like estimands which includes the TSLS and OLS estimands, among many others.

Our method has several applications. First, it can be used to construct nonparametric bounds on the average causal effects of an actual or hypothetical policy change. Second, our method allows the researcher to flexibly incorporate shape restrictions and parametric assumptions, thereby enabling extrapolation of the average effects for compliers to the average effects for different or larger populations. Third, our method provides various specification tests for an IV model. In addition to testing the null of correctly specified model, we can use our method to test null hypotheses of no selection bias, no selection on gains and instrument validity. Importantly, specification tests using our method do not require the treatment effect to be constant over individuals with the same observables.

To illustrate the applicability of our method, we used Norwegian administrative data to draw inference about the causal effects of family size on children's outcomes. To date, existing research on this topic has chosen corner solutions, being unwilling or unable to make a tradeoff between external and internal validity. The studies have either only reported the LATEs, with their high degree of internal validity but limited external validity, or invoked the full set of assumptions necessary to point identify the MTEs (and hence the treatment effects for the entire population). Our empirical results suggest a middle ground, showing that bounds on treatment parameters other than the LATEs are informative. Additionally, we show that only weak auxiliary assumptions are necessary to extrapolate the average causal effects for a small complier group to the average causal effect for a much larger population.

A Proofs

Proof of Proposition 1. Using (1), note that

$$\beta_s = \mathbb{E}[SDY_1] + \mathbb{E}[S(1-D)Y_0]. \tag{31}$$

Then using (2), observe that the first term of (31) can be written as

$$\mathbb{E}[SDY_1] = \mathbb{E}\left[S\mathbb{1}[U \leq p(Z)]\,\mathbb{E}\left(Y_1 | U, Z\right)\right] \equiv \mathbb{E}\left[s(1,Z)\mathbb{1}[U \leq p(Z)] m_1(U,Z)\right], \tag{32}$$

where the first equality follows because SD is a deterministic function of (U, Z) and second equality uses the definition of m_1, together with the identity that

$$S\mathbb{1}[U \leq p(Z)] \equiv s\left(\mathbb{1}[U \leq p(Z)], Z\right)\mathbb{1}[U \leq p(Z)] = s(1,Z)\mathbb{1}[U \leq p(Z)].$$

Using the normalization of $U|Z$ as uniformly distributed on $[0,1]$, for any realization of Z it follows from (32) that

$$\begin{aligned}\mathbb{E}[SDY_1] &= \mathbb{E}\left[\mathbb{E}\left(s(1,Z)\mathbb{1}[U \leq p(Z)] m_1(U,Z)|Z\right)\right] \\ &= \mathbb{E}\left[\int_0^1 s(1,Z)\mathbb{1}[u \leq p(Z)] m_1(u,Z)\,du\right] \equiv \mathbb{E}\left[\int_0^1 \omega_{1s}(u,Z) m_1(u,Z)\,du\right].\end{aligned}$$

The claimed result follows after applying a symmetric argument to the second term of (31). *Q.E.D.*

Proof of Proposition 2. Observe that the mapping τ_s is linear in M. Hence, $\mathcal{M}_{\mathcal{S}}$ is either convex or empty. Since $\beta^\star$ is also linear in M, the image of $\mathcal{M}_{\mathcal{S}}$ under $\beta^\star$, i.e. $\mathcal{B}^\star_{\mathcal{S}}$, is also either convex or empty. Hence, $\mathcal{B}^\star_{\mathcal{S}}$ is either empty or an extended real-valued interval. Since every point in $\mathcal{B}^\star_{\mathcal{S}}$ is by definition larger than $\underline{\beta}^\star$ and smaller than $\overline{\beta}^\star$, it follows that the closure of $\mathcal{B}^\star_{\mathcal{S}}$ is equal to $[\underline{\beta}^\star, \overline{\beta}^\star]$. (If $\underline{\beta}^\star$ and $\overline{\beta}^\star$ are actually attainable by some $M \in \mathcal{M}_{\mathcal{S}}$, then $\mathcal{B}^\star_{\mathcal{S}}$ is exactly equal to this interval.) *Q.E.D.*

Proof of Proposition 3. If $\beta \in \mathcal{B}^\star_{\text{id}}$, then there exists an $F \in \mathcal{F}_{\text{id}}$ such that $\beta^\star(M_F) = \beta$. Since $F \in \mathcal{F}_{\text{id}}$, it follows that $\tau_s(M_F) = \beta_s$ for all $s \in \mathcal{S}$, and hence that $\beta \in \mathcal{B}^\star_{\mathcal{S}}$.

Conversely, suppose that $\beta \in \mathcal{B}^\star_{\mathcal{S}}$, so that there exists an $M = (m_0, m_1) \in \mathcal{M}$ such that $\beta^\star(M) = \beta$ and $\tau_s(M) = \beta_s$ for all $s \in \mathcal{S}$. Since $Y \in \{0,1\}$ is binary, m_d

determines a marginal distribution F_d for Y_d conditional on U and Z, i.e.

$$F_d(y|u,z) = \begin{cases} 0, & \text{if } y < 0 \\ 1 - m_d(u,z), & \text{if } y \in [0,1) \\ 1, & \text{if } y \geq 1. \end{cases}$$

Combine these two conditional marginal distributions together with an arbitrary copula to form a conditional joint distribution F for $(Y_0, Y_1)|U, Z$.[13] From Proposition 1 it follows that

$$\mathbb{E}[Ys(D,Z)] = \beta_s = \tau_s(M) = \mathbb{E}_F[Ys(D,Z)]$$

for all $s \in \mathcal{S}$. With $\mathcal{S}$ following the specification given in the statement of the proposition, this implies that

$$\mathbb{E}[Y|D=d, Z=z] = \mathbb{E}_F[Y|D=d, Z=z]$$

for all $d \in \{0,1\}$ and a.e. $z \in \mathcal{Z}$. However, since Y is binary, this is equivalent to the statement that

$$\mathbb{P}[Y=1|D=d, Z=z] = \mathbb{P}_F[Y=1|D=d, Z=z],$$

for a.e. $z \in \mathcal{Z}$, which implies that $F \in \mathcal{F}_{\text{id}}$. Since $\beta^\star(M_F) = \beta^\star(M) = \beta$, it follows that $\beta \in \mathcal{B}^\star_{\text{id}}$, and hence that $\mathcal{B}^\star_{\mathcal{S}} = \mathcal{B}^\star_{\text{id}}$. *Q.E.D.*

Proof of Proposition 4. First note that if $\beta^\star = c_0$, then there exists a $\bar{\theta} \in \Theta$ with

$$\gamma_s'\bar{\theta} = \beta_s \text{ for all } s \in \mathcal{S} \text{ and } \gamma^\star\bar{\theta} = c_0 \ . \tag{33}$$

Moreover, by construction it also follows that $\underline{c}_n \leq \beta^\star \leq \bar{c}_n$ holds whenever the point $(c_0, \bar{\theta}) \in \mathbb{R} \times \Theta$ satisfies the constraints in the optimization problems defining $\bar{c}_n$ and $\underline{c}_n$ (see (28) and (29)). Therefore, rearranging terms we obtain the bounds

$$\begin{aligned} P(\underline{c}_n \leq \beta^\star \leq \bar{c}_n) &\geq P(|\sqrt{n}\{\hat{\gamma}_s'\bar{\theta} - \hat{\beta}_s\}| \vee |\sqrt{n}\{(\hat{\gamma}^\star)'\bar{\theta} - c_0\}| \leq \hat{c}_{1-\alpha} \text{ for all } s \in \mathcal{S}) \\ &\geq P(\sup_{\theta \in \Theta} \max_{s \in \mathcal{S}} |\sqrt{n}\{(\hat{\gamma}_s - \gamma_s)'\theta - (\hat{\beta}_s - \beta_s)\}| \vee |\sqrt{n}\{(\hat{\gamma}^\star - \gamma^\star)'\theta\}| \leq \hat{c}_{1-\alpha}) \end{aligned} \tag{34}$$

where in the final inequality we exploited (33) and that $\bar{\theta} \in \Theta$. The conclusion of the

[13] For example, use the product copula to define $F(y_0, y_1|u,z) = F_0(y_0|u,z)F_1(y_1|u,z)$.

Proposition then follows by noting that since Θ is bounded by hypothesis, the map

$$\sup_{\theta\in\Theta}\max_{s\in\mathcal{S}}|\sqrt{n}\{(\hat{\gamma}_s'-\gamma_s)\theta-(\hat{\beta}_s-\beta_s)\}|\vee|\sqrt{n}(\hat{\gamma}^\star-\gamma^\star)'\theta| \tag{35}$$

is a Lipschitz continuous transformation of $(\sqrt{n}(\hat{\gamma}^\star-\gamma^\star),\{\sqrt{n}(\hat{\gamma}_s-\gamma_s),\sqrt{n}(\hat{\beta}_s-\beta_s)\}_{s\in\mathcal{S}})$ and thus (30) holds by Proposition 10.7 in Kosorok (2008) and Assumptions II.1 and II.2. *Q.E.D.*

B MTR Weights for Linear IV Estimands

In this appendix we show that linear IV estimands are a special case of our notion of an IV–like estimand. For the purpose of this discussion, we adopt some of the standard textbook terminology regarding "endogenous variables" and "included" and "excluded" instruments in the context of linear IV models without heterogeneity. Consider a linear IV specification with endogenous variables $\widetilde{X}_1$, included instruments $\widetilde{Z}_1$, and excluded instruments $\widetilde{Z}_2$. We let $\widetilde{X}\equiv[\widetilde{X}_1,\widetilde{Z}_1]'$ and $\widetilde{Z}\equiv[\widetilde{Z}_2,\widetilde{Z}_1]'$. We assume that both $\mathbb{E}[\widetilde{Z}\widetilde{Z}']$ and $\mathbb{E}[\widetilde{Z}\widetilde{X}']$ have full rank.

As long as these two conditions hold, all of the variables in $\widetilde{X}$ and $\widetilde{Z}$ can be functions of (D,Z). Usually, one would expect that $\widetilde{X}_1$ would include D and possibly some interactions between D and other covariates X. The instruments, $\widetilde{Z}$, would usually consist of functions of the vector Z, which contains X, by notational convention. The included portion of $\widetilde{Z}$, i.e. $\widetilde{Z}_1$, would typically also include a constant term as one of its components. However, whether $\widetilde{Z}$ is actually "exogenous" in the usual sense of the linear instrumental variables model is not relevant to our definition of an IV–like estimand or the derivation of the weighting expression (12). In particular, OLS is nested as a linear IV specification through the case in which $\widetilde{Z}_1=[1,D]'$ and both $\widetilde{X}_1$ and $\widetilde{Z}_2$ are empty vectors.

It may be the case that $\widetilde{Z}$ has dimension larger than $\widetilde{X}$, as in "overidentified" linear models. In such cases, a positive definite weighting matrix Π is used to generate instruments $\Pi\widetilde{Z}$ that have the same dimension as $\widetilde{X}$. A common choice of Π is the two-stage least squares weighting $\Pi_{\text{TSLS}}\equiv\mathbb{E}[\widetilde{X}\widetilde{Z}']\,\mathbb{E}[\widetilde{Z}\widetilde{Z}']^{-1}$ which has as its rows the first stage coefficients corresponding to linear regressions of each component of $\widetilde{X}$ on the entire vector $\widetilde{Z}$. We assume that Π is known or identified non-stochastic matrix with full rank, which covers Π_{TSLS} and the optimal weighting under heteroskedasticity (optimal GMM) as particular cases.

The instrumental variables estimator that uses $\Pi\widetilde{Z}$ as an instrument for $\widetilde{X}$ in a regression of Y on $\widetilde{X}$ has corresponding estimand

$$\beta_{\text{IV},\Pi} \equiv \left(\Pi\,\mathbb{E}[\widetilde{Z}\widetilde{X}']\right)^{-1}\left(\Pi\,\mathbb{E}[\widetilde{Z}Y]\right) = \mathbb{E}\left[\left(\Pi\,\mathbb{E}[\widetilde{Z}\widetilde{X}']\right)^{-1}\Pi\widetilde{Z}Y\right],$$

which is an IV–like estimand with $s(D,Z) \equiv (\Pi\,\mathbb{E}[\widetilde{Z}\widetilde{X}'])^{-1}\Pi\widetilde{Z}$. As in the case of simple IV discussed in the main text, the "denominator" $(\Pi\,\mathbb{E}[\widetilde{Z}\widetilde{X}'])^{-1}$ will cancel out of the constraints in (16), given that it enters as a constant multiple in the definitions of both β_s and the weights ω_{ds}. As a consequence, it can be omitted from these definitions without affecting the empirical content of the model.

C Bernstein Polynomials

The kth Bernstein basis polynomial of degree K is defined as

$$b_k : [0,1] \to \mathbb{R} : b_k(u) \equiv \binom{K}{k} u^k(1-u)^{K-k}$$

for $k = 0, 1, \ldots, K$. A degree K Bernstein polynomial B is a linear combination of these $K+1$ basis polynomials:

$$B(u) : [0,1] \to \mathbb{R} : B(u) \equiv \sum_{k=0}^{K} \theta_k b_k(u),$$

for some constants $\theta_0, \theta_1, \ldots, \theta_K$. As is well-known, any continuous function on $[0,1]$ can be uniformly well approximated by a Bernstein polynomial of sufficiently high order.

The shape of B can be constrained by imposing linear restrictions on $\theta_0, \theta_1, \ldots, \theta_K$. This computationally appealing property of the Bernstein polynomials has been noted elsewhere by Chak, Madras, and Smith (2005), Chang, Chien, Hsiung, Wen, and Wu (2007) and McKay Curtis and Ghosh (2011), among others. The following constraints are especially useful in the current application. Derivations of these properties can be found in Chang et al. (2007) and McKay Curtis and Ghosh (2011).

Shape Constraints

S.1 *Bounded below by* 0*:* $\theta_k \geq 0$ *for all* k.

S.2 *Bounded above by* 1*:* $\theta_k \leq 1$ *for all* k.

S.3 *Monotonically increasing:* $\theta_0 \leq \theta_1 \leq \cdots \leq \theta_K$.

S.4 *Concave:* $\theta_k - 2\theta_{k+1} + \theta_{k+2} < 0$ *for* $k = 0, \ldots, K-2$.

Each Bernstein polynomial basis is itself an ordinary degree K polynomial. The coefficients on this ordinary polynomial representation (i.e. the power basis representation) can be computed by applying the binomial theorem:

$$b_k(u) = \sum_{i=k}^{K} (-1)^{i-k} \binom{K}{i} \binom{i}{k} u^i. \tag{36}$$

Representation (36) is useful for computing the terms $\gamma^{\star}_{dk}$ and γ_{dks} that appear in the finite-dimensional program (16). To see this note for example that with $d = 1$,

$$\gamma_{1ks} \equiv \mathbb{E}\left[\int_0^1 b_{1k}(u, Z)\omega_{1s}(u, Z)\, du\right] = \mathbb{E}\left[s(0, Z)\int_0^{p(Z)} b_{1k}(u, Z)\, du\right]$$

If $b_{1k}(u, Z) = b_{1k}(u)$ is a Bernstein polynomial, then $\int_0^{p(Z)} b_{1k}(u)\, du$ can be computed analytically through elementary calculus using (36). The result of this integral is a known function of $p(Z)$. The coefficient γ_{1ks} is then simply the population average of the product of this known function with $s(0, Z)$, which is also known or identified. Thus, no numerical integration is needed to compute or estimate the γ_{dks} terms. This conclusion depends on the form of the weights, and may not hold for all target weights $\omega^{\star}_{dk}$, although it holds for all of the parameters listed in Table 1. When it does not, one-dimensional numerical integration can be used instead.

References

Abadie, A., J. Angrist, and G. Imbens (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91–117. 15

Andrews, D. W. K. and X. Shi (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81, 609–666. 17

Angrist, J., 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. 32, 34

Angrist, J. D. and I. Fernandez-Val (2013): "ExtrapoLATE-ing: External Validity and," in *Advances in Economics and Econometrics: Volume 3, Econometrics: Tenth World Congress*, Cambridge University Press, vol. 51, 401–. 7

Angrist, J. D., K. Graddy, and G. W. Imbens (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *The Review of Economic Studies*, 67, 499–527. 2

Angrist, J. D. and G. W. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 90, 431–442. 2

Angrist, J. D. and A. B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics*, 106, 979–1014. 2

Balke, A. and J. Pearl (1997): "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1176. 6

Becker, G. S. and H. G. Lewis (1973): "On the Interaction between the Quantity and Quality of Children," *Journal of Political Economy*, 81, S279–S288. 32

Bertsimas, D. and J. N. Tsitsiklis (1997): *Introduction to linear optimization*, vol. 6, Athena Scientific Belmont, MA. 26

Black, S. E., P. J. Devereux, and K. G. Salvanes (2005): "The More the Merrier? The Effect of Family Size and Birth Order on Children's Education," *The Quarterly Journal of Economics*, 120, 669–700. 32, 33, 34

Brinch, C. N., M. Mogstad, and M. Wiswall (2015): "Beyond LATE with a Discrete Instrument," *Journal of Political Economy*, forthcoming. 5, 6, 18, 19, 32, 33

Caceres-Delpiano, J. (2006): "The Impacts of Family Size On Investment in Child Quality," *Journal of Human Resources*, 41, 738–754. 32

Carneiro, P., J. J. Heckman, and E. Vytlacil (2010): "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin," *Econometrica*, 78, 377–394. 6, 18, 20, 21

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011): "Estimating Marginal Returns to Education," *American Economic Review*, 101, 2754–81. 6, 11, 18, 20, 21

Chak, P. M., N. Madras, and B. Smith (2005): "Semi-nonparametric estimation with Bernstein polynomials," *Economics Letters*, 89, 153–156. 41

Chang, I.-S., L.-C. Chien, C. A. Hsiung, C.-C. Wen, and Y.-J. Wu (2007): "Shape restricted regression with random Bernstein polynomials," in *Lecture Notes–Monograph Series*, ed. by R. Liu, W. Strawderman, and C.-H. Zhang, Beachwood, Ohio, USA: Institute of Mathematical Statistics, vol. Volume 54, 187–202. 41

Chen, X. (2007): "Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. Volume 6, Part 2, 5549–5632. 4, 15

Chesher, A. (2003): "Identification in Nonseparable Models," *Econometrica*, 71, 1405–1441. 2

Eisenhauer, P., J. J. Heckman, and E. Vytlacil (2015): "The Generalized Roy Model and the Cost-Benefit Analysis of Social Programs," *Journal of Political Economy*, 123, 413–443. 22

Fang, Z. and A. Santos (2014): "Inference on directionally differentiable functions," *arXiv preprint arXiv:1404.3763*. 30

Florens, J. P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of Treatment Effects Using Control Functions in Models With Continuous, Endogenous Treatment and Heterogeneous Effects," *Econometrica*, 76, 1191–1206. 2

French, E. and J. Song (2014): "The Effect of Disability Insurance Receipt on Labor Supply," *American Economic Journal: Economic Policy*, 6, 291–337. 19

Gurobi Optimization, I. (2015): "Gurobi Optimizer Reference Manual," . 4

Heckman, J. J. and D. Schmierer (2010): "Tests of hypotheses arising in the correlated random coefficient model," *Economic Modelling*, 27, 1355–1367. 7

Heckman, J. J., D. Schmierer, and S. Urzua (2010): "Testing the correlated random coefficient model," *Journal of Econometrics*, 158, 177–203. 7

Heckman, J. J. and S. Urzua (2010): "Comparing IV with structural models: What simple IV can and cannot identify," *Journal of Econometrics*, 156, 27–37. 2

Heckman, J. J., S. Urzua, and E. Vytlacil (2006): "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics*, 88, 389–432. 2

Heckman, J. J. and E. Vytlacil (2001a): "Policy-Relevant Treatment Effects," *The American Economic Review*, 91, 107–111. 3, 6, 20

——— (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, 669–738. 2, 3, 6, 7, 9, 10, 11, 12, 18, 20, 21, 26, 27

Heckman, J. J. and E. J. Vytlacil (1999): "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences of the United States of America*, 96, 4730–4734. 2, 3, 10, 18, 20

——— (2001b): "Instrumental Variables, Selection Models, and Tight Bounds on the Average Treatment Effect," in *Econometric Evaluations of Active Labor Market Policies in Europe*, ed. by M. Lechner and F. Pfeiffer, Heidelberg and Berlin: Physica. 3, 5

——— (2001c): "Local Instrumental Variables," in *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, ed. by K. M. C Hsiao and J. Powell, Cambridge University Press. 3, 18

——— (2007a): "Chapter 70 Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. Volume 6, Part 2, 4779–4874. 3

——— (2007b): "Chapter 71 Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. Volume 6, Part 2, 4875–5143. 2, 3

Huber, M. and G. Mellace (2014): "Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints," *Review of Economics and Statistics*, 97, 398–411. 6

IBM (2010): *IBM ILOG AMPL Version 12.2*, International Business Machines Corporation. 4

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475. 2, 3, 9, 22, 23

Imbens, G. W. and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77, 1481–1512. 2

Imbens, G. W. and D. B. Rubin (1997): "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," *The Review of Economic Studies*, 64, 555–574. 6

Kirkeboen, L., E. Leuven, and M. Mogstad (2015): "Field of Study, Earnings and Self-Selection," *The Quarterly Journal of Economics*, forthcoming. 2

Kitagawa, T. (2015): "A Test for Instrument Validity," *Econometrica*, 83, 2043–2063. 6, 25

Kosorok, M. (2008): "Introduction to empirical processes and semiparametric inference. 2008," . 40

Kowalski, A. (2016): "Doing More When You're Running LATE: Applying Marginal Treatment Effect Methods to Examine Treatment Effect Heterogeneity in Experiments," *NBER Working paper 22363.* 19

Lee, S. and B. Salanié (2016): "Identifying Effects of Multivalued Treatments," *Working paper.* 2

Maestas, N., K. J. Mullen, and A. Strand (2013): "Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt," *The American Economic Review*, 103, 1797–1829. 11

Manski, C. (1994): "The selection problem," in *Advances in Econometrics, Sixth World Congress*, vol. 1, 143–70. 5

Manski, C. F. (1989): "Anatomy of the Selection Problem," *The Journal of Human Resources*, 24, 343–360. 5, 19, 24

——— (1990): "Nonparametric Bounds on Treatment Effects," *The American Economic Review*, 80, 319–323. 5

——— (1997): "Monotone Treatment Response," *Econometrica*, 65, 1311–1334. 5, 19

——— (2003): *Partial identification of probability distributions*, Springer. 5

Manski, C. F. and J. V. Pepper (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68, 997–1010. 6, 19

Masten, M. A. (2015): "Random Coefficients on Endogenous Variables in Simultaneous Equations Models," *cemmap working paper 25/15.* 2

Masten, M. A. and A. Torgovitsky (2016): "Identification of Instrumental Variable Correlated Random Coefficients Models," *The Review of Economics and Statistics*, forthcoming. 2

McKay Curtis, S. and S. K. Ghosh (2011): "A variable selection approach to monotonic regression with Bernstein polynomials," *Journal of Applied Statistics*, 38, 961–976. 41

MOURIFIÉ, I. AND Y. WAN (2015): "Testing Local Average Treatment Effect Assumptions," *Working paper.* 6

SHAPIRO, A. (1989): "Asymptotic properties of statistical estimators in stochastic programming," *The Annals of Statistics*, 841–858. 30

TORGOVITSKY, A. (2015): "Identification of Nonseparable Models Using Instruments With Small Support," *Econometrica*, 83, 1185–1197. 2

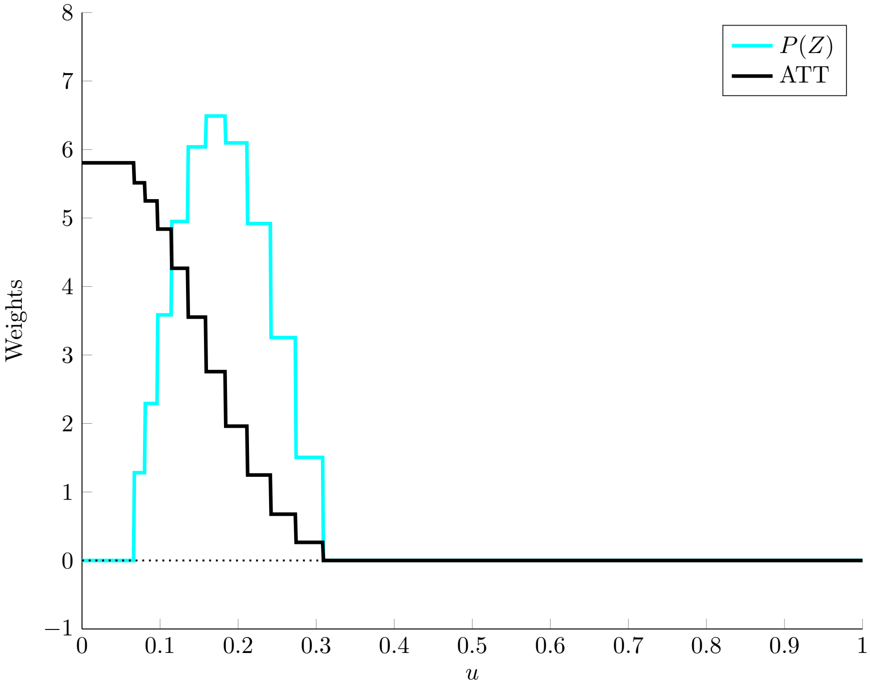
VYTLACIL, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331–341. 3, 9

Table 1: Weights for a Variety of Target Parameters and Estimands

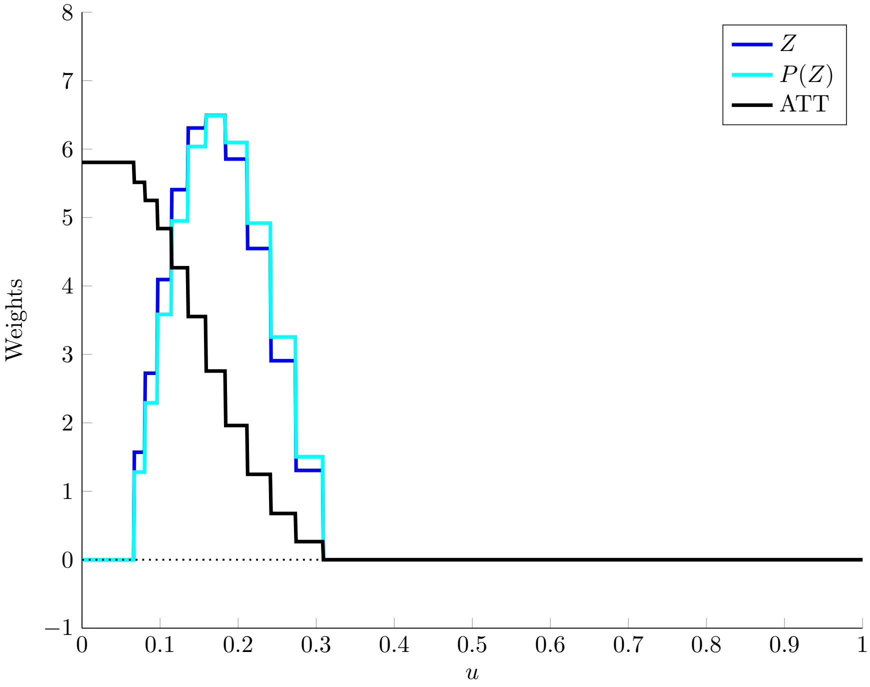
Note: take $\mathcal{Z}^\star = \mathcal{Z}$ for unconditional averages		**Weights**		
Quantity	**Expression**	$\boldsymbol{\omega_0(u,z)}$	$\boldsymbol{\omega_1(u,z)}$	**Measure $\boldsymbol{\mu^\star}$**
Average Untreated Outcome given $Z \in \mathcal{Z}^\star$	$\mathbb{E}[Y_0 \vert Z \in \mathcal{Z}^\star]$	1	0	Leb.$[0,1]$
Average Treated Outcome given $Z \in \mathcal{Z}^\star$	$\mathbb{E}[Y_1 \vert Z \in \mathcal{Z}^\star]$	0	1	Leb.$[0,1]$
Average Treatment Effect (ATE) given $Z \in \mathcal{Z}^\star$	$\mathbb{E}[Y_1 - Y_0 \vert Z \in \mathcal{Z}^\star]$	-1	1	Leb.$[0,1]$
Average Treatment on the Treated (ATT) given $Z \in \mathcal{Z}^\star$	$\mathbb{E}[Y_1 - Y_0 \vert D = 1, Z \in \mathcal{Z}^\star]$	$-\omega_1^\star(u,z)$	$\frac{\mathbb{1}[u \leq p(z)]}{\mathbb{P}[D=1 \vert Z \in \mathcal{Z}^\star]}$	Leb.$[0,1]$
Average Treatment on the Untreated (ATU) given $Z \in Z^\star$	$\mathbb{E}[Y_1 - Y_0 \vert D = 0, Z \in \mathcal{Z}^\star]$	$-\omega_1^\star(u,z)$	$\frac{\mathbb{1}[u > p(z)]}{\mathbb{P}[D=0 \vert Z \in \mathcal{Z}^\star]}$	Leb.$[0,1]$
Local Average Treatment Effect for $U \in [\underline{u}, \overline{u}]$ (LATE$(\underline{u}, \overline{u})$)	$\mathbb{E}[Y_1 - Y_0 \vert U \in [\underline{u}, \overline{u}]]$	$-\omega_1^\star(u,z)$	$\frac{\mathbb{1}[u \in [\underline{u}, \overline{u}]]}{(\overline{u} - \underline{u})}$	Leb.$[0,1]$

Selection bias	$\mathbb{E}[Y_0 \mid D=1]$ $- \mathbb{E}[Y_0 \mid D=0]$	$\frac{\mathbb{1}[u \leq p(z)]}{\mathbb{P}[D=1]} - \frac{\mathbb{1}[u > p(z)]}{\mathbb{P}[D=0]}$	0	Leb.$[0,1]$
Selection on the gain	$\mathbb{E}[Y_1 - Y_0 \mid D=1]$ $- \mathbb{E}[Y_1 - Y_0 \mid D=0]$	$-\omega_1^\star(u,z)$	$\frac{\mathbb{1}[u \leq p(z)]}{\mathbb{P}[D=1]} - \frac{\mathbb{1}[u > p(z)]}{\mathbb{P}[D=0]}$	Leb.$[0,1]$
IV–Like estimand	$\mathbb{E}[s(D,Z)Y]$	$s(0,Z)\mathbb{1}[u > p(Z)]$	$s(1,Z)\mathbb{1}[u \leq p(Z)]$	Leb.$[0,1]$
Average Marginal Treatment Effect at $\overline{u}$	$\mathbb{E}[m_1(\overline{u},Z) - m_0(\overline{u},Z)]$	-1	1	Dirac($\overline{u}$)
Policy Relevant Treatment Effect (PRTE) for new policy $(p^\star, Z^\star)$	$\frac{\mathbb{E}[Y^\star]-\mathbb{E}[Y]}{\mathbb{E}[D^\star]-\mathbb{E}[D]}$	$-\omega_1^\star(u,z)$	$\frac{\mathbb{1}[u \leq p^\star(z^\star)]-\mathbb{1}[u \leq p(z)]}{\mathbb{E}[p^\star(Z^\star)]-\mathbb{E}[p(Z)]}$	Leb.$[0,1]$
Additive PRTE with magnitude α	PRTE with $Z^\star = Z$ and $p^\star(z) = p(z) + \alpha$	$-\omega_1^\star(u,z)$	$\frac{\mathbb{1}[u \leq p(z)+\alpha]-\mathbb{1}[u \leq p(z)]}{\alpha}$	Leb.$[0,1]$
Proportional PRTE with magnitude α	PRTE with $Z^\star = Z$ and $p^\star(z) = (1+\alpha)p(z)$	$-\omega_1^\star(u,z)$	$\frac{\mathbb{1}[u \leq (1+\alpha)p(z)]-\mathbb{1}[u \leq p(z)]}{\alpha\,\mathbb{E}[p(Z)]}$	Leb.$[0,1]$
PRTE for an additive α shift of the j^{th} component of Z	PRTE with $Z^\star = Z + \alpha e_j$ and $p^\star(z) = p(z)$	$-\omega_1^\star(u,z)$	$\frac{\mathbb{1}[u \leq p(z+\alpha e_j)]-\mathbb{1}[u \leq p(z)]}{\mathbb{E}[p(Z+\alpha e_j)]-\mathbb{E}[p(Z)]}$	Leb.$[0,1]$
Sum of two quantities $\beta_A^\star$, $\beta_B^\star$ with common measure $\mu^\star$	$\beta_A^\star + \beta_B^\star$	$\omega_{A,0}^\star(u,z) + \omega_{B,0}^\star(u,z)$	$\omega_{A,1}^\star(u,z) + \omega_{B,1}^\star(u,z)$	Common $\mu^\star$

Figure 1: Numerical Illustration of Propensity

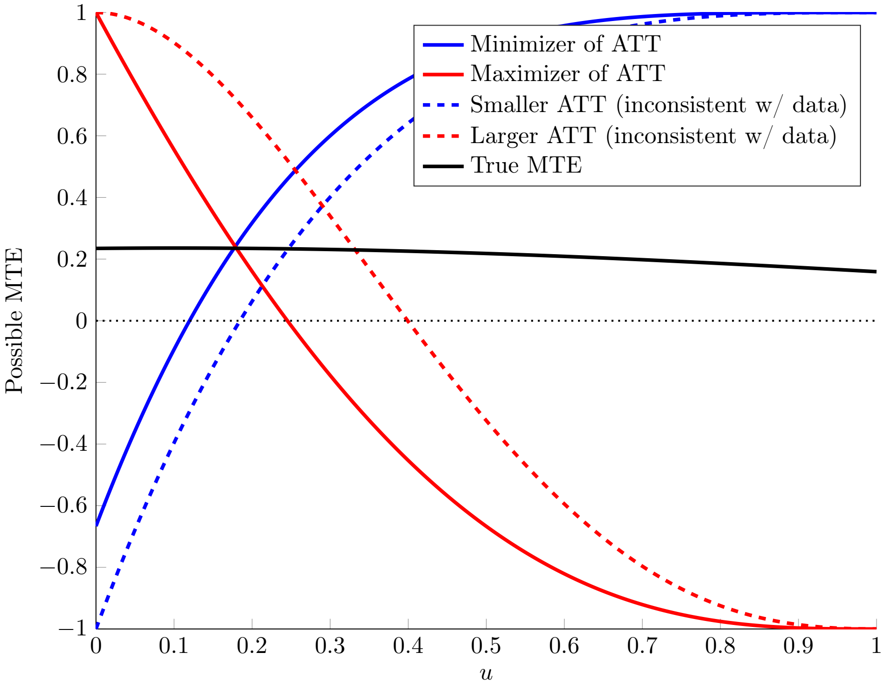


(a) Weights for columns (1)–(4) of Table 2a.

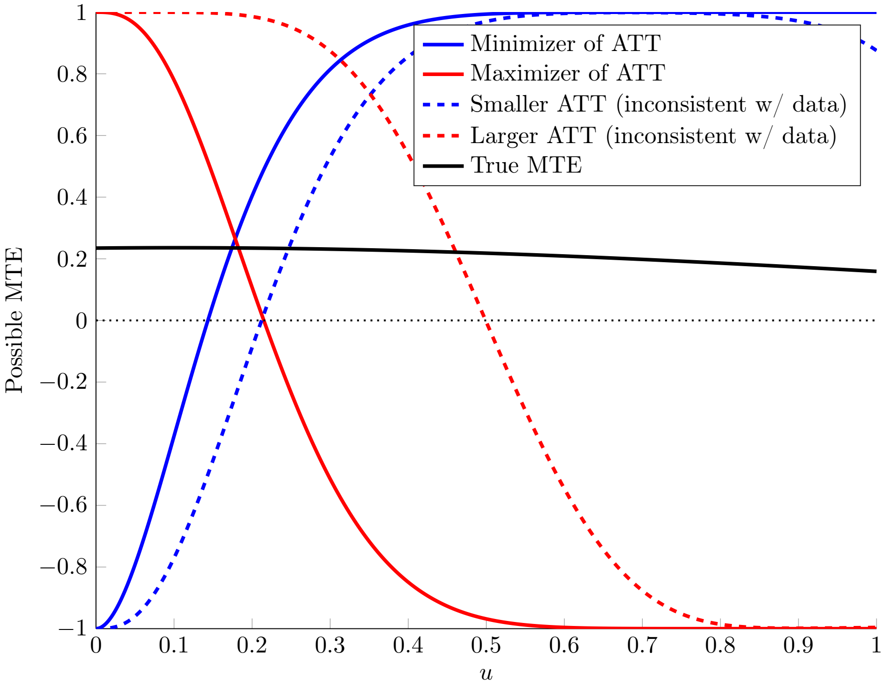


(b) Weights for columns (5)–(8) of Table 2a.

Figure 2: Numerical Illustration of MTE

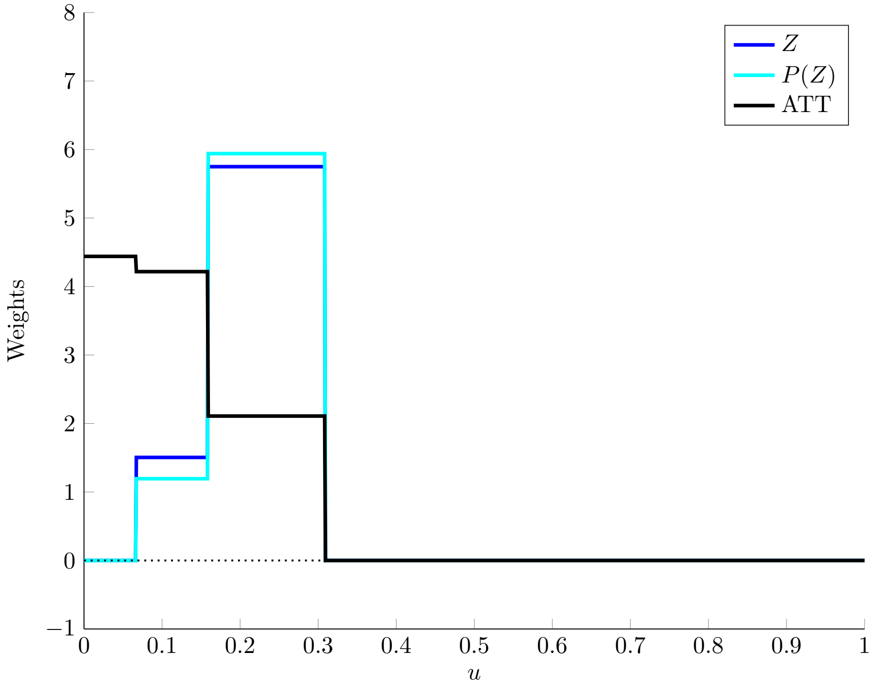


(a) Optimal and infeasible MTEs for column (2) of Table 2a.

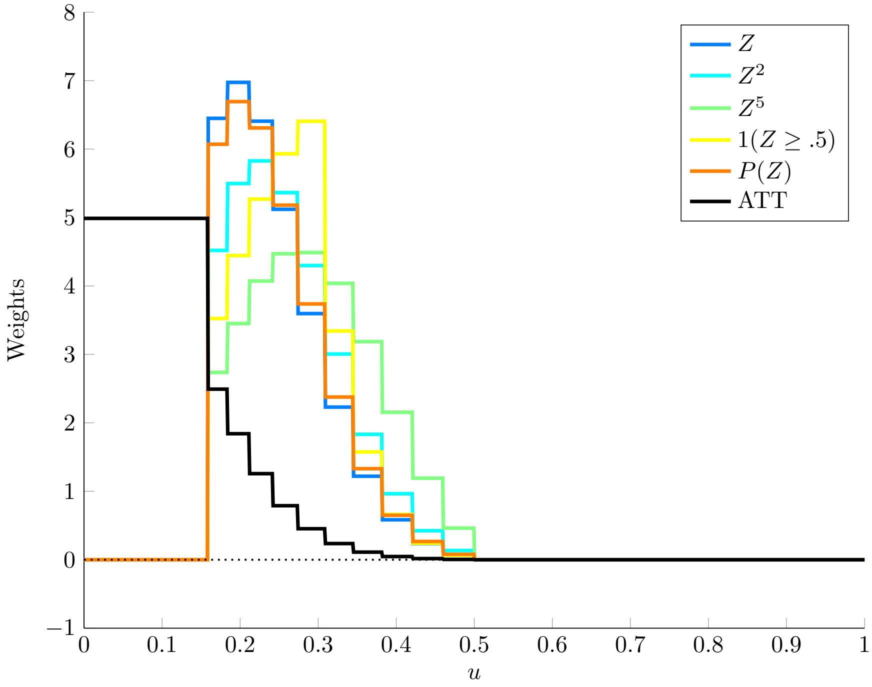


(b) Optimal and infeasible MTEs for column (4) of Table 2a.

Figure 3: Numerical Illustration of Weights



(a) Weights for columns (5)–(8) of Table 3a.



(b) Weights for columns (9)–(12) of Table 3b.

Table 2: Numerical Illustration of Bounds on the ATT

Instrument	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
$p(Z)$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Z					✓	✓	✓	✓	✓	✓	✓	✓
Z^2									✓	✓	✓	✓
Z^5									✓	✓	✓	✓
$1(Z \geq 0)$									✓	✓	✓	✓
K	2	5	10	15	2	5	10	15	5	10	15	20
$\underline{\beta}^\star$	.156	-.138	-.300	-.387	.236	.154	-.074	-.113	$\emptyset$	.231	.198	.157
$\overline{\beta}^\star$	.362	.575	.657	.695		.302	.437	.499		.239	.267	.324
$\overline{\beta}^\star - \underline{\beta}^\star$	.206	.713	.957	1.082	0	.148	.511	.612	—	.007	.069	.167

(a) Bounds on the ATT in the numerical illustration of Section 4. The checkmarks indicate different collections of IV–like estimands, while K is the number of linear basis terms.

Table 3: Comparison of Numerical Illustrations

Instrument	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
$p(Z)$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Z					✓	✓	✓	✓	✓	✓	✓	✓
Z^2									✓	✓	✓	✓
Z^5									✓	✓	✓	✓
$1(Z \geq 0)$									✓	✓	✓	✓
K	2	5	10	15	2	5	10	15	5	10	15	20
$\underline{\beta}^\star$	.142	-.226	-.332	-.391	.235	.186	.051	.012	.186	.051	.012	-.031
$\overline{\beta}^\star$	.381	.559	.639	.675		.277	.368	.403	.277	.368	.403	.416
$\overline{\beta}^\star - \underline{\beta}^\star$	.238	.785	.970	1.066	0	.091	.317	.390	.091	.317	.390	.447

(a) This table provides the same information as in Table 2a for the numerical illustration when Z has only three points of support.

Instrument	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
$p(Z)$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Z					✓	✓	✓	✓	✓	✓	✓	✓
Z^2									✓	✓	✓	✓
Z^5									✓	✓	✓	✓
$1(Z \geq .5)$									✓	✓	✓	✓
K	2	5	10	15	2	5	10	15	2	5	10	15
$\underline{\beta}^\star$	.089	-.381	-.536	-.612	.238	-.003	-.293	-.423	$\emptyset$	.231	.107	-.062
$\overline{\beta}^\star$	.463	.685	.799	.828		.513	.625	.688		.241	.378	.467
$\overline{\beta}^\star - \underline{\beta}^\star$	.374	1.066	1.335	1.440	0	.516	.918	1.111	—	.010	.270	.529

(b) This table provides the same information as in Table 2a for the numerical illustration when the support of Z is shifted to the right by .5.

Table 4: Descriptive Statistics and Regression Estimates

Panel A. Descriptive Statistics

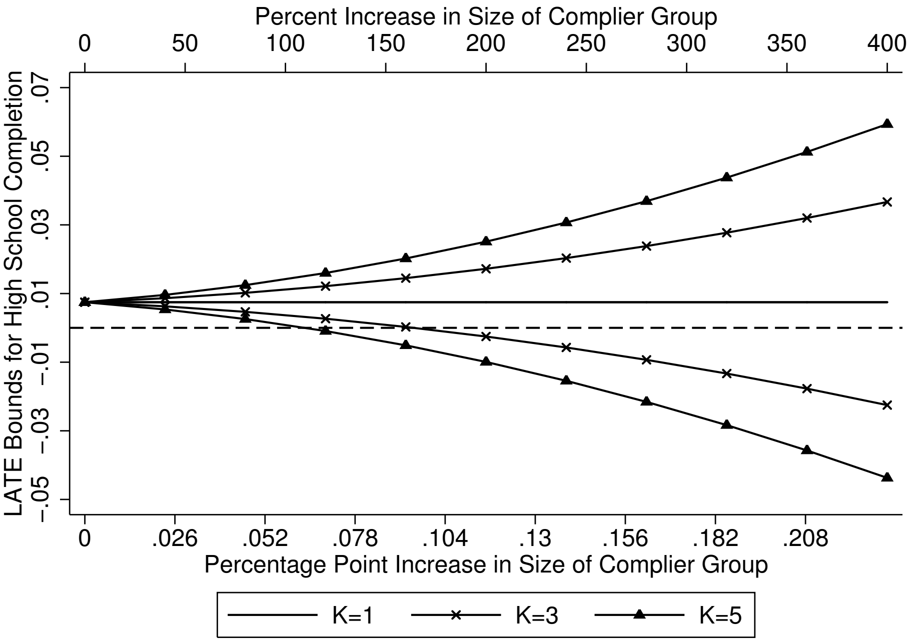
	All	2 children	3+ children
	Mean	Mean	Mean
Outcomes:			
High School Completion	0.638	0.675	0.602
Instruments:			
Same sex, 1st and 2nd child	0.500	0.471	0.529
Twins at second birth	0.010	0.00	0.019
Endogenous regressor:			
At least three children	0.502	0	1
Covariates:			
Female child	0.473	0.475	0.471
Older mother	0.482	0.543	0.422
Number of observations	514,004	255,933	258,071

Panel B. Regression Estimates

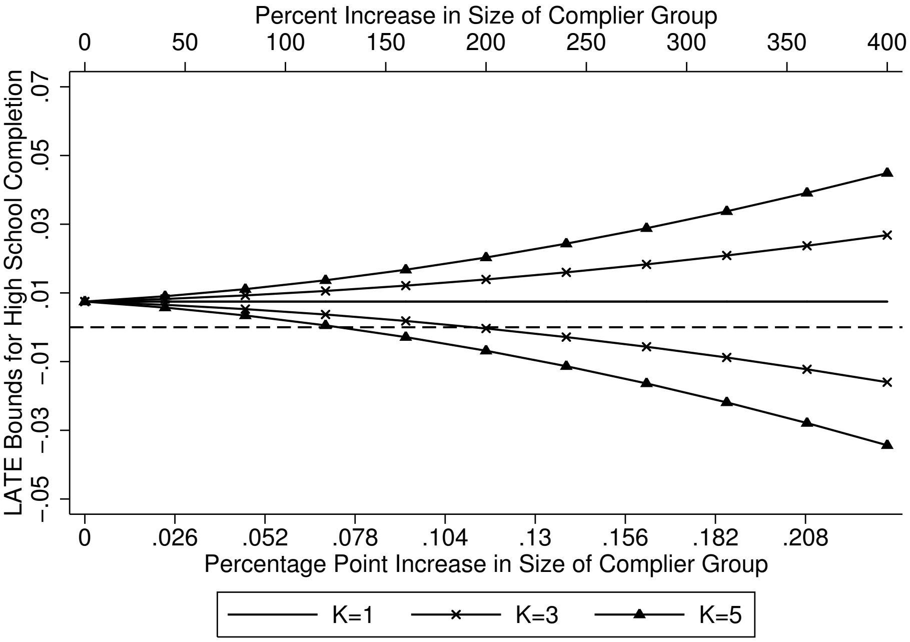
	Estimate
OLS:	
No covariates	-0.073
	(0.001)
Saturated in covariates	-0.045
	(0.001)
TSLS with same-sex instrument:	
No covariates	0.030
	(0.023)
Saturated in covariates	0.007
	(0.024)

Notes: The covariates are female child, older mother, and female child × older mother. In the regression model that is saturated in covariates, the standard errors are computed using 200 bootstrap samples.

Figure 4: LATE Extrapolation of the Effect of Family Size



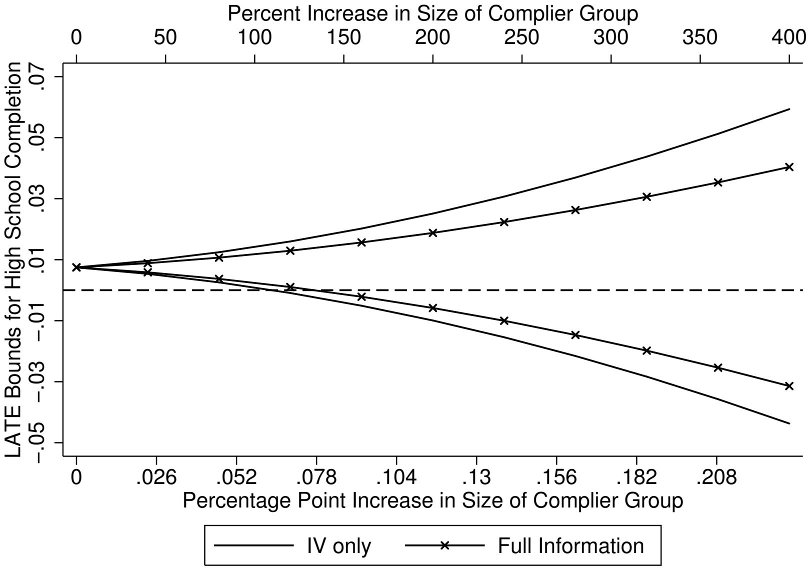
(a) Extrapolation based on IV estimate only



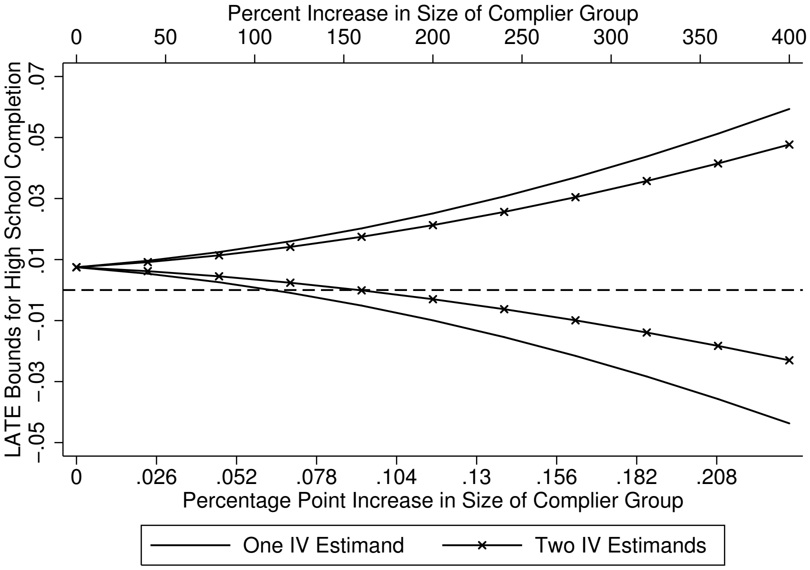
(b) Extrapolation based on IV and OLS estimates

Notes: Panel (a) constructs bounds based on the same-sex IV estimate alone. Panel (b) constructs bounds based on the OLS estimate and the same-sex IV estimate.

Figure 5: Sensitivity of LATE Extrapolation to the Choice of Information Set



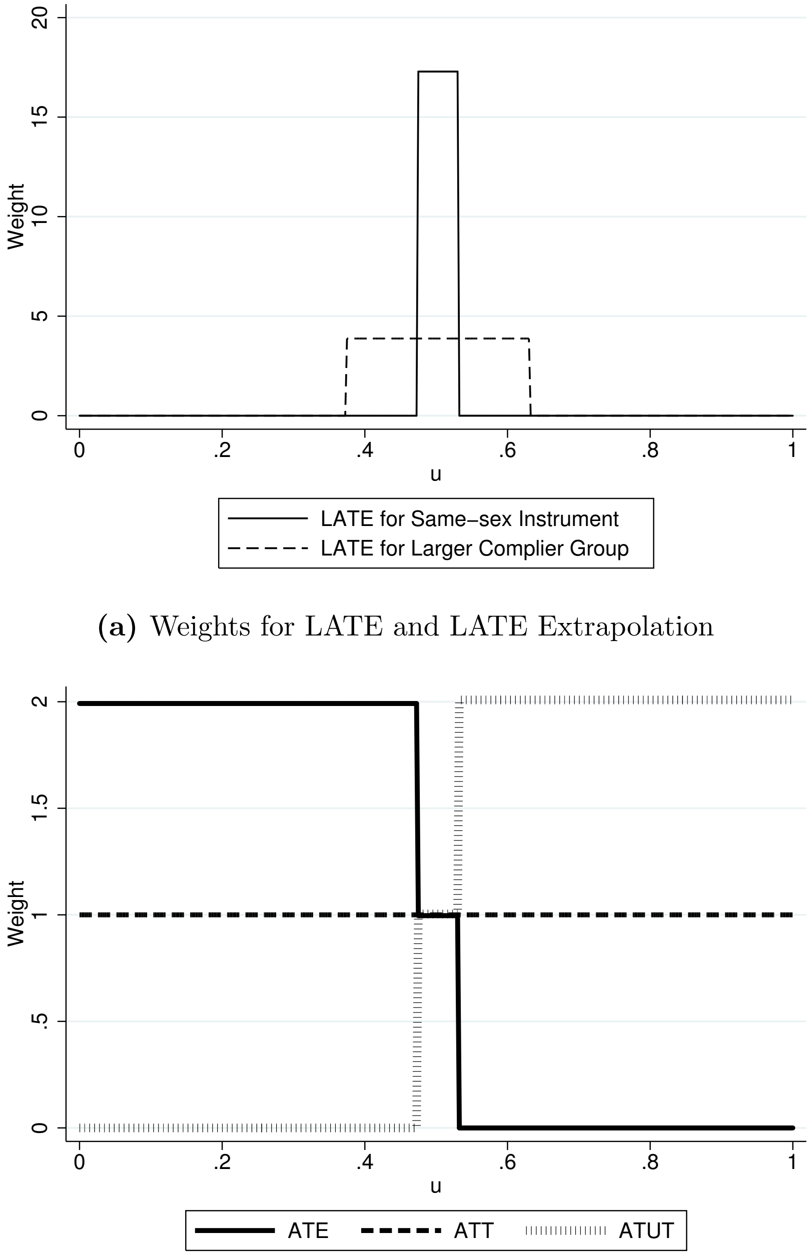
(a) IV only compared to full information



(b) One instrument compared to two instruments and their interaction

Notes: Panel (a) compares the bounds based on the same-sex IV estimate alone versus bounds based on the joint distribution of the outcome, treatment, and same-sex instrument. Panel (b) compares bounds based on the same-sex IV estimate alone versus bounds based on the same-sex IV estimate and TSLS estimate with same-sex, twin, and their interaction as instruments. In both panels, the polynomial order is $K = 5$.

Figure 6: Weights on Treatment Effects



(a) Weights for LATE and LATE Extrapolation

(b) Weights for ATE, ATT, and ATUT

Notes: The larger complier group adds 20 percent of the population to the initial complier group.