Identification of Causal Effects with Multiple Instruments: Problems and Some Solutions[*]

Magne Mogstad[†] Alexander Torgovitsky[‡] Christopher R. Walters[§]

February 11, 2019

Abstract

Empirical researchers often combine multiple instruments for a single treatment using two stage least squares (2SLS). When treatment effects are heterogeneous, a common justification for including multiple instruments is that the 2SLS estimand can still be interpreted as a positively-weighted average of local average treatment effects (LATEs). This justification requires the well-known monotonicity condition. However, we show that with more than one instrument, this condition can only be satisfied if choice behavior is effectively homogenous. Based on this finding, we consider the use of multiple instruments under a weaker, partial monotonicity condition. This condition is implied by standard choice theory and allows for richer heterogeneity. First, we show that the weaker partial monotonicity condition can still suffice for the 2SLS estimand to be a positively-weighted average of LATEs. We characterize a simple sufficient and necessary condition that empirical researchers can check to ensure positive weights. Second, we develop a general method for using multiple instruments to identify a wide range of causal parameters other than LATEs. The method allows researchers to combine multiple instruments to obtain more informative empirical conclusions than one would obtain by using each instrument separately.

[*] We thank Stéphane Bonhomme, Vishal Kamat, and Ed Vytlacil for helpful comments, and Christine Blandhol and John Bonney for outstanding research assistance.

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

[‡] Department of Economics, University of Chicago. Research supported in part by National Science Foundation grant SES-1530538.

[§] Department of Economics, University of California, Berkeley; NBER.

1 Introduction

Instrumental variables (IVs) are widely used to estimate causal relationships. In practice, researchers often combine multiple instruments using two stage least squares (2SLS). In Appendix A, we document this practice with a survey of empirical research published in leading journals since 2000. We find that among the papers using an IV, more than half report results from a specification with multiple instruments, typically combined using 2SLS.

The textbook motivation for combining multiple instruments is greater statistical efficiency. However, this requires an assumption of constant treatment effects. In contrast, allowing for heterogeneous treatment effects is a key motivation in the modern program evaluation literature, and one which is supported by a large body of empirical work.[1] In an influential paper, Imbens and Angrist (1994, "IA" hereafter) provided an alternative justification for using 2SLS with multiple instruments, which allows for heterogeneous treatment effects. They show that the 2SLS estimand can be interpreted as a positively-weighted average of local average treatment effects (LATEs) for subpopulations whose treatment status would be affected by the instrument. This result holds for any number of instruments, as long as IA's "monotonicity" condition is satisfied.[2]

The fact that widespread empirical practice rests on the monotonicity condition raises a number of questions. What requirements does this condition place on behavior when there are multiple instruments? Can we expect these requirements to be satisfied? If not, then how else can one use multiple instruments for causal inference and policy evaluation while still allowing for heterogeneous treatment effects? The contribution of our paper is to answer these questions and, by doing so, develop a general blueprint for extracting and aggregating information about treatment effects from multiple controlled or natural experiments while still allowing for rich heterogeneity in both treatment effects and choice behavior.

We begin, in Section 2, by showing that the monotonicity condition cannot be satisfied with more than one instrument without restricting choice behavior to be ef-

[1] Heckman (2001) compiled a list of empirical evidence on heterogeneous treatment effects prior to 2001. More recent papers that find evidence of heterogeneity include Bitler, Gelbach, and Hoynes (2006), Doyle Jr. (2007), Moffitt (2008), Carneiro and Lee (2009), Firpo, Fortin, and Lemieux (2009), Carneiro, Heckman, and Vytlacil (2011), Maestas, Mullen, and Strand (2013), Bitler, Hoynes, and Domina (2014), Walters (2014), Felfe and Lalive (2014), French and Song (2014), Havnes and Mogstad (2015), Kirkeboen, Leuven, and Mogstad (2016), Kline and Walters (2016), Hull (2016), Carneiro, Lokshin, and Umapathi (2016), Cornelissen, Dustmann, Raute, and Schönberg (forthcoming), Nybom (2017), and Brinch, Mogstad, and Wiswall (2017).

[2] In discussing the use of multiple instruments instead of a single binary instrument, Angrist and Pischke (2009, p. 173) write that "The econometric tool remains 2SLS and the interpretation remains fundamentally similar to the basic LATE result, with a few bells and whistles."

fectively homogenous.[3] For example, if the treatment is college attendance and the instruments are tuition and proximity, the monotonicity condition requires all individuals to respond more to tuition than to proximity, or vice versa. This is a concerning implication; it shows that appealing to IA monotonicity to justify combining multiple instruments using 2SLS comes at the cost of assuming homogeneity in choice behavior.

Motivated by this observation, we then develop ways to to use multiple instruments under a strictly weaker, *partial* version of the monotonicity condition. The partial monotonicity condition is that the IA monotonicity condition is satisfied for each instrument separately, holding all of the other instruments fixed. We show that partial monotonicity is satisfied if each instrument makes every individual weakly more likely to choose treatment. For example, a sufficient condition for partial monotonicity is that all individuals are at least as likely to attend college if they live closer to a college or face lower tuitions. However, unlike the IA monotonicity condition, partial monotonicity does not restrict heterogeneity in the relative impacts of different instruments; it allows for some individuals to respond more to tuition than to proximity, and for others to respond more to proximity than to tuition.

In Section 3, we show that even though partial monotonicity permits heterogeneous choice behavior, it can still be sufficient to give the 2SLS estimand an interpretation as a positively-weighted average of LATEs. Moreover, we characterize sufficient and necessary conditions for this interpretation. We show that the conditions can be checked empirically by verifying that the *unconditional* correlations between the treatment and each instrument have the expected sign. These results provide a simple empirical check that researchers can report alongside 2SLS estimates formed from multiple instruments.

Weighted averages of LATEs do not generally answer interesting policy counterfactuals, so in Section 4 we develop a new general framework for conducting inference on specific target parameters. The framework is based on insights from the literature on marginal treatment effects (Heckman and Vytlacil, 1999, 2001, 2005, 2007a,b). While that literature always maintains the IA monotonicity condition, we develop our framework under only partial monotonicity so that multiple instruments can be considered without imposing choice homogeneity.[4]

There are two inherent identification problems that arise when considering inference on specific, policy-relevant target parameters. First, there is the problem of extrapo-

[3] Our analysis here builds upon points made by Heckman and Vytlacil (2005, Section 6) and Heckman, Urzua, and Vytlacil (2006, Section III.D).

[4] Our framework therefore provides a middle-ground between approaches rooted in IA monotonicity and approaches that remain agnostic on treatment choice, such as Manski (1990, 1994), Manski and Pepper (2000, 2009), and Chernozhukov and Hansen (2005).

lating from the individuals whose treatment choices would be affected by the variation in the data to the individuals relevant for the counterfactual question posed by the desired target parameter. This problem arises even if there is only a single instrument (Mogstad and Torgovitsky, 2018). Second, combining multiple instruments requires aggregating across the different choice margins generated by each instrument.

We address the first identification problem by using the ideas developed in Mogstad, Santos, and Torgovitsky (2018). Their method allows researchers to extract identifying information about a specific target parameter using each instrument separately. We solve the second problem by showing that using each instrument separately generates implications about certain "instrument-invariant" parameters, such as the average treatment effect. This suggests a "logical consistency" condition that requires these implications to not be contradictory. We show that this logical consistency condition allows identifying information to be aggregated across different instruments. Thus, our new method provides a general blueprint for extracting and aggregating variation across different sources of exogenous variation, including controlled or natural experiments, without imposing strong restrictions on choice behavior such as the IA monotonicity condition.

2 The Monotonicity Condition with Multiple Instruments

We begin by considering the interpretation of the IA monotonicity condition when there are multiple instruments. To do this, we first develop an equivalent characterization of the condition which facilitates a graphical interpretation of its content. Then, we argue that the condition severely restricts choice heterogeneity. This motivates interest in the weaker, partial monotonicity assumption that we use in the rest of the paper. Throughout the paper, we focus on the workhorse case of a binary treatment, and we condition on covariates nonparametrically.

2.1 Definition of the IA Monotonicity Condition

Consider a population of individuals $i \in \mathcal{I}$. Denote individual i's potential treatment status if some instrument Z_i were set to z by $D_i(z) \in \{0, 1\}$, where z takes values in some subset $\mathcal{Z}$ of $\mathbb{R}^L$. We assume that the support of Z_i is contained in $\mathcal{Z}$, possibly as a proper subset. When $L > 1$, we view each component of the vector Z_i as comprising an economically distinct quantity. That is, if $L = 2$ then $Z_{i,1}$ and $Z_{i,2}$ will denote the two distinct instruments, each of which can take two or more values.

Imbens and Angrist (1994, "IA") introduced the following assumption on the potential treatment states, which they described as "monotonicity."

Assumption IAM. ***(IA Monotonicity)*** *For all $z, z' \in \mathcal{Z}$ either $D_i(z) \geq D_i(z')$ for all $i \in \mathcal{I}$, or $D_i(z) \leq D_i(z')$ for all $i \in \mathcal{I}$.*

Heckman and Vytlacil (2005, pp. 715-716) observed that Assumption IAM requires uniformity across individuals, not monotonicity in the instrument. The results ahead provide further justification of their observation. Nevertheless, to conform with the existing literature, we still refer to Assumption IAM as "IA *monotonicity*." For clarity, we refer to the usual definition of monotonicity as "*actual* monotonicity."

Assumption AM. ***(Actual Monotonicity)*** *If $z' \geq z$ in the vector sense (componentwise), then $D_i(z') \geq D_i(z)$ for all $i \in \mathcal{I}$.*

We show below that IA monotonicity (Assumption IAM) neither implies nor is implied by actual monotonicity (Assumption AM).

2.2 A Graphical Characterization of IA Monotonicity

Assumption IAM is a comparison across all individuals for any two values of Z_i. To interpret this condition when Z_i is a vector, it is useful to rephrase it as a comparison across all values of Z_i for any two individuals. The equivalent condition is that for any two individuals j and k, either j must take treatment under all instrument values that k does, or the opposite. This is the content of the following proposition.[5]

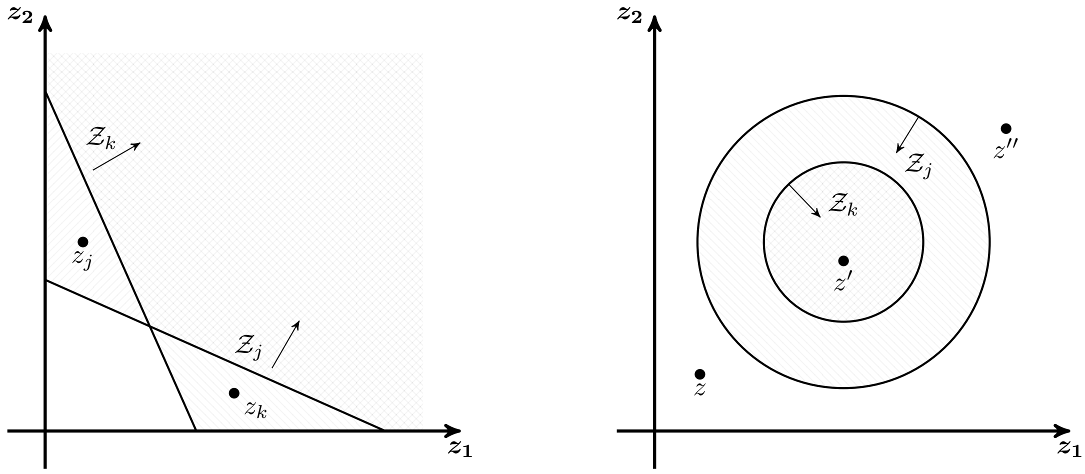
Proposition 1. *For any $i \in \mathcal{I}$, define $\mathcal{Z}_i \equiv \{z \in \mathcal{Z} : D_i(z) = 1\}$ as the set of all instrument values for which individual i would take treatment. Then Assumption IAM holds if and only if for all $j, k \in \mathcal{I}$, either $\mathcal{Z}_j \subseteq \mathcal{Z}_k$, or $\mathcal{Z}_k \subseteq \mathcal{Z}_j$.*

Proposition 1 shows that Assumption IAM can be interpreted as a "nesting condition" among the sets of instrument values that induce different agents to take treatment.[6] This means that with two instruments one can gain intuition about the content of Assumption IAM by drawing sets in $\mathbb{R}^2$.

For example, in Figure 1a, we have drawn two sets $\mathcal{Z}_j$ and $\mathcal{Z}_k$ that are not nested. Proposition 1 says that Assumption IAM fails, which can be verified by comparing the choices individuals j and k would make at the points marked z_j and z_k. Yet, for both individuals j and k, the instrument has a monotonic effect in the sense that $D_i(z)$ is

[5] Proofs for all propositions are contained in Appendix B.

[6] This nesting condition is different than the "equivalent monotonicity condition" used by Vytlacil (2002, p. 335), although it shares a superficial resemblance. Vytlacil (2002, p. 336) used the sets $\mathcal{Z}_i$ for his proof of the existence of an equivalent latent index model.



(a) *Sets $\mathcal{Z}_j$ and $\mathcal{Z}_k$ are not nested, so Proposition 1 implies that Assumption IAM does not hold. For example, compare z_j and z_k: $D_j(z_j) = 1 > 0 = D_k(z_j)$, while $D_k(z_k) = 1 > 0 = D_j(z_k)$. Yet, $D_i(z)$ is monotone in z for both $i = j, k$.*

(b) *If $\mathcal{I} = \{j, k\}$, then Proposition 1 shows that Assumption IAM would hold. However, neither $D_j(z)$ nor $D_k(z)$ are monotone in z. For example, $z \leq z'$, and $z' \leq z''$, but $D_i(z) = D_i(z'') = 0 < D_i(z') = 1$ for $i = j, k$.*

Figure 1: *Assumption IAM neither implies nor is implied by monotonicity of $D_i(z)$ in z.*

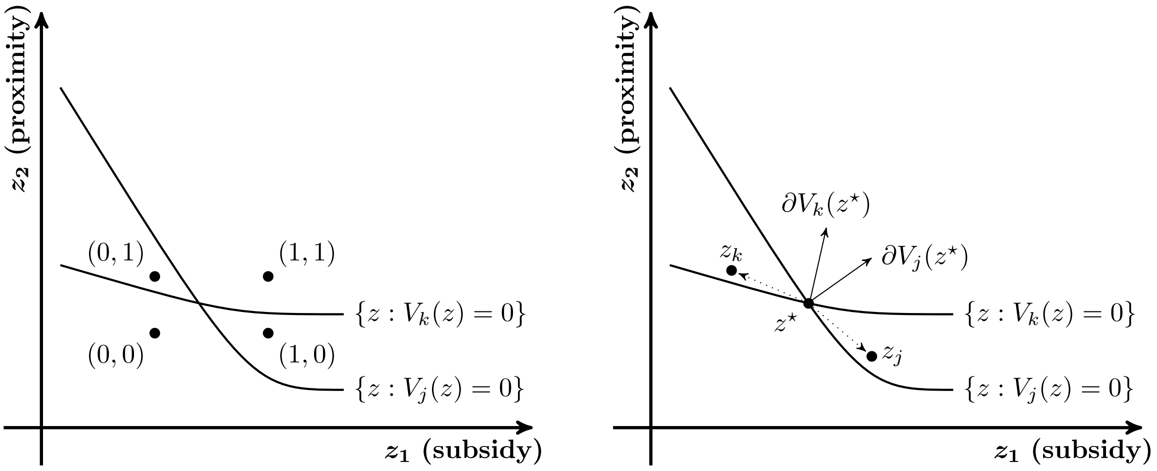
increasing in z. That is, if $z' \geq z$ as a vector (component-wise), then $D_i(z') \geq D_i(z)$. This shows that Assumption AM does not imply Assumption IAM.[7]

Figure 1b depicts the opposite case, in which $\mathcal{Z}_j \subseteq \mathcal{Z}_k$. If $\mathcal{I}$ only consists of individuals like j and k, then Proposition 1 implies that Assumption IAM is satisfied. However, the instrument does not have a monotonic effect on treatment choice. For example, moving from z to $z' \geq z$ moves both individuals' choices from 0 to 1, but moving from z' to $z'' \geq z'$ moves their choices back to 0. This shows that Assumption IAM does not imply monotonicity in the usual sense of Assumption AM.

2.3 Implications for Heterogeneity in Choice Behavior

In this section, we examine the restrictions that Assumption IAM places on choice behavior. To do this, we use a random utility model. Assume that individual i's indirect utility from choosing d when the instrument is z is given by $V_i(d, z)$. The

[7] This finding contrasts with the assertion of Heckman et al. (2006, p. 400) that "If $[D_i(z)]$ is monotonic in the usual usage of this term, and response are in the same direction for all people, then the 'monotonicity' or 'uniformity' condition IV-3 will be satisfied." The context of their statement is somewhat ambiguous. While it is true for a scalar instrument (see Proposition 3 ahead), Figure 1a shows that it is false in general.



(a) *Each instrument takes two binary values. Individual j has $\mathcal{Z}_j = \{(1,0),(1,1)\}$ and individual k has $\mathcal{Z}_k = \{(0,1),(1,1)\}$. Points $(1,0)$ and $(0,1)$ violate the nesting condition in Proposition 1.*

(b) *An illustration of Proposition 2. When (3) fails to hold for two individuals $j, k \in \mathcal{I}$, one can find points z_j and z_k which violate the nesting condition in Proposition 1 by taking small steps in the directions of the dotted arrows.*

Figure 2: *Assumption IAM requires homogenous choice behavior.*

individual chooses $D_i(z) = 1$ if and only if $V_i(1,z) \geq V_i(0,z)$:

$$D_i(z) = \underset{d \in \{0,1\}}{\arg\max}\ V_i(d,z) \equiv \begin{cases} 1, & \text{if } V_i(z) \geq 0 \\ 0, & \text{if } V_i(z) < 0 \end{cases}, \tag{1}$$

where $V_i(z) \equiv V_i(1,z) - V_i(0,z)$ and ties are resolved in favor of treatment.

For concreteness, consider the familiar setting of the returns to schooling in which $D_i(z)$ represents a binary decision to attend college. Suppose that $z = (z_1, z_2)$, where z_1 is a tuition subsidy, and z_2 is proximity to a college. Larger values of either instrument encourages college attendance, so that $D_i(z)$ is a monotonic function of z and Assumption AM is satisfied. As just noted, this neither implies nor is implied by Assumption IAM. In Figure 2, we draw two possible indifference curves along which individuals j and k would be on the margin between attending and not attending college.

Suppose that z only takes the values $\{(0,0),(0,1),(1,0),(1,1)\}$ shown in Figure 2a. Then individual j would attend college if and only if they received a tuition subsidy, regardless of whether they lived in close proximity. Individual k would attend college if and only if they lived in close proximity, regardless of whether they received a tuition subsidy. That is, $(1,0)$ is in $\mathcal{Z}_j$ but not $\mathcal{Z}_k$, and $(0,1)$ is in $\mathcal{Z}_k$ but not $\mathcal{Z}_j$, so that

Assumption IAM fails. Thus, Assumption IAM does not permit individuals to differ in their response to different incentives to attend college: All individuals must find either a tuition subsidy or distance to be more compelling. This suggests a strong form of preference homogeneity.

We can sharpen this statement when the instrument is continuous and the net indirect utility function is differentiable. This is shown in the next proposition, which is illustrated in Figure 2b.

Proposition 2. *Suppose that $D_i(z)$ is determined by* (1). *Let $z^\star$ be a point in the interior of $\mathcal{Z}$, and let $\mathcal{I}(z^\star) \equiv \{i \in \mathcal{I} : V_i(z^\star) = 0\}$ denote the set of individuals who are indifferent between treatment and non-treatment when faced with $z^\star$. Suppose further that V_i is continuously differentiable in a neighborhood of $z^\star$. Then Assumption IAM implies that*

$$\partial_1 V_j(z^\star)\partial_2 V_k(z^\star) = \partial_1 V_k(z^\star)\partial_2 V_j(z^\star) \tag{2}$$

for all $j, k \in \mathcal{I}(z^\star)$, where $\partial_\ell V_i(z) \equiv \frac{\partial}{\partial z_\ell} V_i(z)$ for $\ell = 1, 2$.

Proposition 2 says that if Assumption IAM holds, then any two individuals who are indifferent between treatment and non-treatment at $z^\star$ must have the same marginal rate of substitution between the two components of the instrument. That is, assuming that the second component has an impact at $z^\star$ (so that $\partial_2 V_i(z^\star) \neq 0$), Assumption IAM implies

$$\frac{\partial_1 V_j(z^\star)}{\partial_2 V_j(z^\star)} = \frac{\partial_1 V_k(z^\star)}{\partial_2 V_k(z^\star)} \tag{3}$$

for *all* individuals j and k who are indifferent at $z^\star$. This is a strong statement about preference homogeneity.

For example, suppose that individual i's net utility from attending college is given by the random coefficients specification

$$V_i(z) = B_{i,0} + B_{i,1} z_1 + z_2 \quad \text{so that} \quad D_i(z) = \mathbb{1}[B_{i,0} + B_{i,1} z_1 + z_2 \geq 0], \tag{4}$$

where $B_{i,1} \geq 0$ controls variation in the taste for subsidies relative to proximity. Proposition 2 shows that Assumption IAM does not hold if $B_{i,1}$ varies with i.[8] Thus, the college attendance decision of every individual is either affected more by tuition subsidies (if $b_1 \equiv B_{i,1} \geq 1$), or by proximity (if $b_1 < 1$), and all individuals trade off these

[8] Heckman et al. (2006, p. 399) note that Assumption IAM *can* fail in random coefficient specifications like (4). Our analysis shows that it *must* fail when the instruments are continuous.

incentives at the same rate. Assumption IAM does not permit heterogeneity in these behavioral responses.

2.4 Partial IA Monotonicity

Assumption IAM creates unattractive implications for choice behavior because it requires cross-instrument comparisons, such as the comparison between $(0, 1)$ and $(1, 0)$ in Figure 2a. We can eliminate these comparisons by considering a condition that compares only values of a single component of the instrument, holding all other components fixed. To state such a condition, we divide vectors $z \in \mathcal{Z}$ into their ℓth component, z_ℓ, and all other $(L - 1)$ components, $z_{-\ell}$. We write $z = (z_\ell, z_{-\ell})$ to emphasize the separation of the ℓth component.

Consider the following assumption:[9]

Assumption PM. ***(Partial Monotonicity)*** *Take any $\ell = 1, \ldots, L$, and let $(z_\ell, z_{-\ell})$ and $(z'_\ell, z_{-\ell})$ be two points in $\mathcal{Z}$. Then either $D_i(z_\ell, z_{-\ell}) \geq D_i(z'_\ell, z_{-\ell})$ for all $i \in \mathcal{I}$, or $D_i(z_\ell, z_{-\ell}) \leq D_i(z'_\ell, z_{-\ell})$ for all $i \in \mathcal{I}$.*

Assumption IAM clearly implies Assumption PM. When $L = 1$, Assumption PM is equivalent to Assumption IAM; when $L > 1$, it is strictly weaker. To see this, recall Figure 2a, where we determined that Assumption IAM fails. Holding $z_2 = 0$ fixed, both individuals j and k are weakly induced to treatment by switching z_1 from 0 to 1. The same is true when the roles of z_1 and z_2 are swapped. If $\mathcal{I}$ consisted of only individuals like j and k, then Assumption PM would be satisfied.

Figure 2 suggests that a simple sufficient condition for Assumption PM is monotonicity in the usual sense of Assumption AM. This is the content of the following proposition.

Proposition 3. *Assumption AM implies Assumption PM.*

Unlike Assumption IAM, Assumption AM can be easily satisfied in random utility models with heterogeneous preferences. For example, if $V_i(z)$ follows the random coefficients specification (4), then it will be satisfied if $B_{i,1}$ is positive for all i. This is easy to interpret and justify: All individuals are more likely to attend college if tuition is lower, even if they differ in their preferences for tuition relative to proximity. More generally, Proposition 3 shows that Assumption PM is satisfied whenever $V_i(z)$ is weakly increasing in z.[10]

[9] Mountjoy (2018) uses a similar assumption in a setting with multiple unordered treatments.

[10] More generally still, Theorem 4 of Milgrom and Shannon (1994) implies that Assumption PM will be satisfied if $V_i(d, z)$ has the single-crossing property in $(d; z)$. The single-crossing property is ordinal, and is strictly weaker than monotonicity in the usual sense.

Assumption AM is a sufficient but not necessary condition for Assumption PM. To see the difference, consider the interacted random coefficient specification

$$V_i(z) = B_{i,0} + B_{i,1}z_1 + z_2 + B_{i,2}z_1z_2, \tag{5}$$

and suppose for simplicity that $\mathcal{Z} = \{0,1\}^2$. If for all individuals $i \in \mathcal{I}$, $B_{i,1} \geq 0$, $B_{i,2} \leq -1$, but $B_{i,1} \leq -B_{i,2}$, then Assumption AM fails while Assumption PM is satisfied. The reason is due to the strong negative interaction effect between z_1 and z_2 on indirect net utility, which is controlled here by $B_{i,2}$. This implies that $D_i(z_1, z_2)$ is increasing as a function of z_1 when $z_2 = 0$, but decreasing when $z_2 = 1$, and similarly when the roles of z_1 and z_2 are reversed. This violates Assumption AM, even though Assumption PM is satisfied.[11]

3 Interpreting 2SLS under Partial IA Monotonicity

Imbens and Angrist (1994, Theorem 2) showed that under standard instrument exogeneity and relevance conditions, Assumption IAM ensures that the 2SLS estimand can be written as a weighted average of causal effects for complier subpopulations. The weights are convex in that they are non-negative and sum to one. Their result holds regardless of the number of instruments, as long as the first stage for the 2SLS estimand is fully saturated, and the instruments satisfy Assumption IAM. However, we have shown that Assumption IAM requires a strong form of preference homogeneity with two or more distinct instruments. In this section, we show that their result can be partially salvaged when one replaces Assumption IAM with the weaker Assumption PM.

3.1 Potential Outcomes and Exogeneity Condition

Before continuing, we need to introduce an outcome variable, Y_i. We write potential outcomes for Y_i as $Y_i(1)$ and $Y_i(0)$ to correspond to setting D_i to treatment ($D_i = 1$) and non-treatment states ($D_i = 0$). The observed outcome is $Y_i = D_iY_i(1) + (1 - D_i)Y_i(0) = Y_i(D_i)$. The observed treatment state is related to the potential treatment states analyzed in the previous section as

$$D_i = \sum_{z \in \mathcal{Z}} \mathbb{1}[Z_i = z]D_i(z) = D_i(Z_i).$$

[11] To see that the above configuration satisfies Assumption PM, note that $B_{i,1} \geq 0$ implies that $D_i(0,0) \leq D_i(1,0)$, $B_{i,1} \leq -B_{i,2}$ implies that $D_i(0,1) \geq D_i(1,1)$, $D_i(0,0) \leq D_i(0,1)$ by virtue of the normalized coefficient on z_2, and $B_{i,2} \leq -1$ implies that $D_i(1,0) \geq D_i(1,1)$.

G_i (group)	$D_i(0,0)$	$D_i(0,1)$	$D_i(1,0)$	$D_i(1,1)$
Always-taker (at)	1	1	1	1
Eager complier (ec)	0	1	1	1
Reluctant complier (rc)	0	0	0	1
Never-taker (nt)	0	0	0	0
Z_1 complier (1c)	0	0	1	1
Z_2 complier (2c)	0	1	0	1

Table 1: *Types of individuals under Assumption AM with two binary instruments.*

Throughout the paper, we maintain the following exogeneity condition:

Assumption E. ***(Exogeneity)*** $(Y_i(0), Y_i(1), \{D_i(z)\}_{z\in\mathcal{Z}}) \perp\!\!\!\perp Z_i$.

Assumption E is common in nonparametric IV models, and identical to Condition 1 of Imbens and Angrist (1994). For simplicity, we state the condition in terms of full independence, although our analysis will be about mean outcomes, so only requires a weaker form of Assumption E.[12]

3.2 Two Binary Instruments

To build intuition, we first consider a special case in which $Z_i = (Z_{i,1}, Z_{i,2})$ consists of two binary instruments $Z_{i,1} \in \{0,1\}$ and $Z_{i,2} \in \{0,1\}$, so that $\mathcal{Z} = \{0,1\}^2$. We also assume that Assumption AM holds, instead of the weaker Assumption PM that we consider in the subsequent sections. Thus, each instrument weakly encourages all individuals to participate in treatment, so that $D_i(1, z_2) \geq D_i(0, z_2)$ and $D_i(z_1, 1) \geq D_i(z_1, 0)$ for all (z_1, z_2) and all i. Under these conditions the population $\mathcal{I}$ can be partitioned into six mutually exclusive and exhaustive groups.

Proposition 4. *If $\mathcal{Z} = \{0,1\}^2$ and Assumption AM is satisfied, then each $i \in \mathcal{I}$ belongs to exactly one of the six groups in Table 1.*

The terminology in Table 1 modifies that of Angrist, Imbens, and Rubin (1996). Always and never takers always exhibit the same behavior regardless of either instrument. Z_1 compliers take treatment if and only if $Z_{i,1}$ is switched on, while Z_2 compliers take treatment if and only if $Z_{i,2}$ is switched on. Eager compliers participate in treatment if either instrument is on, and reluctant compliers only participate if both instruments are on. For any of the six groups, an increase in either instrument weakly increases

[12] The weaker form still requires that $\{D_i(z)\}_{z\in\mathcal{Z}} \perp\!\!\!\perp Z_i$, but relaxes full joint independence to the restriction that $\mathbb{E}[Y_i(d)|\{D_i(z)\}_{z\in\mathcal{Z}}, Z_i = z']$ not depend on z' for both $d = 0$ and $d = 1$.

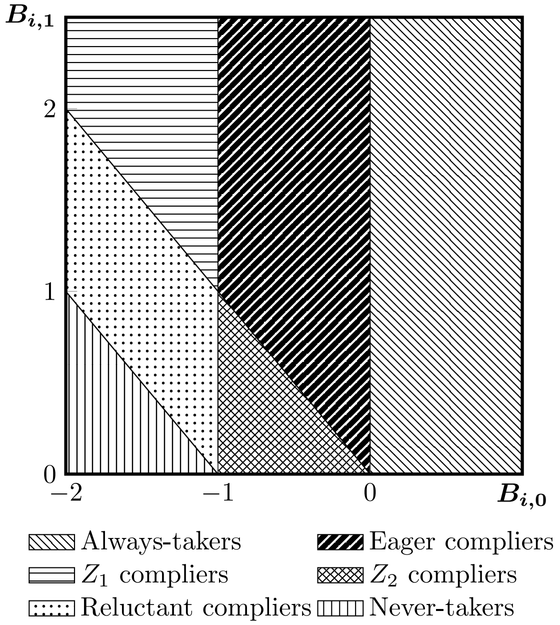


Figure 3: *Correspondence between random coefficients and behavioral types if treatment were determined by the binary choice model (4) with $B_{i,1} \geq 0$. For example, a Z_2 complier has $D_i(0,0) = 0$, $D_i(1,0) = 0$, $D_i(0,1) = 1$, and $D_i(1,1) = 1$. The first three choices imply that $B_{i,0} < 0$, $B_{i,0} + B_{i,1} < 0$, and $B_{i,0} + 1 > 0$, with the fourth choice implied by the third. The region of such $(B_{i,0}, B_{i,1})$ realizations is shown in cross-hatches.*

treatment. Assumption IAM will be violated if the population includes both Z_1 and Z_2 compliers, since in this case a change in Z_i from $(0,1)$ to $(1,0)$ would induce Z_1 compliers to enter treatment and Z_2 compliers to exit treatment. Figure 3 shows how realizations of random coefficients would map into these six types if potential treatment states were generated by (4).

As in Imbens and Angrist (1994, Theorem 2), we consider the 2SLS estimand with a saturated first stage, and a second stage with D_i and a constant. That is, the 2SLS estimand is formed by using 1 (a constant), $Z_{i,1}$, $Z_{i,2}$, and $Z_{i,1}Z_{i,2}$ as instruments for 1 and D_i. Since the first stage is saturated, this 2SLS procedure generates the same coefficient estimate on D_i as the IV estimator that uses the propensity score $p(Z_i) \equiv \mathbb{P}[D_i = 1|Z_i]$ as the sole instrument for D_i. Let β_{2sls} denote this coefficient.[13] Let $G_i \in \{\text{at}, \text{nt}, \text{1c}, \text{2c}, \text{ec}, \text{rc}\}$ denote individual i's group among the six shown in Table 1, and define $\pi_g \equiv \mathbb{P}[G_i = g]$ and $\Delta_g \equiv \mathbb{E}[Y_i(1) - Y_i(0)|G_i = g]$ to be the population shares and average treatment effects for each group. The following proposition establishes the relationship between these quantities and β_{2sls}.

[13] That is, $\beta_{\text{2sls}} = \frac{\text{Cov}(Y_i, p(Z_i))}{\text{Cov}(D_i, p(Z_i))}$.

Proposition 5. *Suppose Z_i has support $\{0,1\} \times \{0,1\}$ and that Assumption AM is satisfied. Suppose in addition that Assumption E is satisfied, and that β_{2sls} exists. Then*

$$\beta_{2sls} = \sum_{g \in \{1c, 2c, ec, rc\}} \tau_g \Delta_g,$$

where the τ_g are weights that sum to 1. Both τ_{ec} and τ_{rc} are always non-negative. If $\pi_{1c} \geq \pi_{2c}$, then τ_{1c} is also non-negative, while the sign of τ_{2c} is given by

$$\operatorname{sgn}(\tau_{2c}) = \mathbb{1}[\pi_{2c} > 0] \times \operatorname{sgn}\Big(\mathbb{P}[D_i = 1 | Z_{i,2} = 1] - \mathbb{P}[D_i = 1 | Z_{i,2} = 0]\Big).$$

If $\pi_{2c} \geq \pi_{1c}$, then τ_{2c} is non-negative, and the sign of τ_{1c} is given by

$$\operatorname{sgn}(\tau_{1c}) = \mathbb{1}[\pi_{1c} > 0] \times \operatorname{sgn}\Big(\mathbb{P}[D_i = 1 | Z_{i,1} = 1] - \mathbb{P}[D_i = 1 | Z_{i,1} = 0]\Big).$$

Proposition 5 shows that the 2SLS estimand is a linear combination of average treatment effects for the four groups that change treatment status in response to one or both of the instruments. The groups are non-overlapping, and the weights on the groups (τ_g) sum to unity.[14] The 2SLS estimand might not be a convex average of treatment effects for the four groups, however, because some of the weights might be negative. The weights for reluctant compliers and eager compliers are always non-negative. If $\pi_{1c} \geq \pi_{2c}$, the weight given to the Z_1 compliers is also non-negative, but the weight given to the Z_2 compliers can be either positive or negative.

The intuition is that a shift of Z_i from $(0,1)$ to $(1,0)$ induces Z_1 compliers to enter treatment and Z_2 compliers to exit treatment. If $\pi_{1c} \geq \pi_{2c}$, then the net effect of this shift is still more participation in treatment. However, the Z_2 compliers act as "defiers," and therefore receive negative weight for this binary contrast. Whether the overall weight given to the Z_2 compliers is positive or negative depends on whether this negative weight is outweighed by the positive weight given to the Z_2 compliers in the two other instrument contrasts for which they enter treatment: $Z_i = (0,0)$ to $Z_i = (0,1)$, and $Z_i = (1,0)$ to $Z_i = (1,1)$. If instead $\pi_{2c} \geq \pi_{1c}$, then the roles of the Z_1 and Z_2 compliers are reversed.

3.3 When is the 2SLS Estimand a Positive Weighted Average?

Proposition 5 shows that one can check whether the 2SLS estimand is a positive weighted average of causal effects by examining observable relationships between treat-

[14] Formulas for τ_g are given in the proof.

ment and instruments. Specifically, if $\pi_{1c} \geq \pi_{2c}$, then τ_{2c} will be negative if and only if the coefficient on $Z_{i,2}$ in a regression of D_i on $Z_{i,2}$ (and a constant) is negative. Likewise, if $\pi_{2c} \geq \pi_{1c}$, then τ_{1c} will be negative if and only if the coefficient on $Z_{i,1}$ in a regression of D_i on $Z_{i,1}$ (and a constant) is negative. One can either check both cases, or check only the relevant case, which is identified by the sign of $p(1,0) - p(0,1) = \pi_{1c} - \pi_{2c}$.[15]

Finding negative weights in either case represents a situation in which the unconditional relationship between an instrument and the treatment has a different sign than the *ceteris paribus* impact of the instrument, which is positive under Assumption AM. This may be rare in practice, since researchers often have prior beliefs regarding the impacts of the instruments (e.g. if each instrument is an encouragement to take treatment), and a researcher may be unlikely to use an instrument if the raw correlation with the treatment contradicts the theoretically expected sign. A necessary condition for such a contradiction is that the instruments are negatively correlated.

Proposition 6. *Suppose Z_i consists of two binary instruments that satisfy Assumption AM, Assumption E is satisfied, and β_{2sls} exists. If $\text{Cov}(Z_{i,1}, Z_{i,2}) \geq 0$, then both τ_{1c} and τ_{2c} are non-negative.*

An important special case of Proposition 6 is when the instruments are independent, so that $\text{Cov}(Z_{i1}, Z_{i,2}) = 0$ and the 2SLS weights are guaranteed to be positive. The leading scenario in which the instruments would be negatively correlated is when $Z_{i,j} = 1$ tends to imply $Z_{i,k} = 0$ for $j \neq k$. For example, $Z_{i,1}$ and $Z_{i,2}$ may indicate two different arms in an experiment corresponding to different types of encouragement to take the treatment. In this setting, $\text{Cov}(Z_{i,1}, Z_{i,2}) < 0$ and negative weights are possible.

3.4 Multivalued Instruments

Suppose that Z_i consists of two or more distinct, discrete instruments, and that its support has K elements total. Label these elements as $\text{supp}(Z_i) \equiv \{z^1, \ldots, z^K\}$ in increasing order of the propensity score, so that $p(z^k) \geq p(z^{k-1})$ for all $k \geq 2$. In the case considered in Section 3.2, $K = 4$ and the instrument values would be ordered as $z^1 = (0,0)$, $z^2 = (0,1)$, $z^3 = (1,0)$, and $z^4 = (1,1)$ if $p(1,0) \geq p(0,1)$, with the roles of z^2 and z^3 switched in the opposite case.

Suppose that Assumption PM is satisfied. Let $\mathcal{G}$ represent the set of all realizations of $\{D_i(z)\}_{z \in \mathcal{Z}}$ that are consistent with Assumption PM. In Section 3.2, there were two binary instruments that satisfied Assumption AM, so $\mathcal{G}$ was composed of the six groups

[15] Except for π_{at} and π_{nt}, the group shares are not themselves separately identified, since there are five linearly independent unknown shares π_g and only four values of $p(Z_i)$.

in Table 1. As before, let G_i denote individual i's group membership, let π_g denote the proportion of the population in each group g, and let Δ_g denote group g's average treatment effect.

For each $g \in \mathcal{G}$, define the set:

$$\mathcal{C}_g = \{k \in \{2, \ldots, K\} : \; D_i(z^k) = 1 \text{ and } D_i(z^{k-1}) = 0 \text{ for all } i \text{ with } G_i = g\}.$$

This is the set of instrument values k at which individuals in group g are compliers in the sense that they would not take treatment if $Z_i = z^{k-1}$, but would take treatment if $Z_i = z^k$. Similarly, define

$$\mathcal{D}_g = \{k \in \{2, \ldots, K\} : \; D_i(z^k) = 0 \text{ and } D_i(z^{k-1}) = 1 \text{ for all } i \text{ with } G_i = g\}$$

as the set of instrument values at which individuals in group g act as defiers. For example, in Section 3.2 with $p(1,0) \geq p(0,1)$, we had $\mathcal{C}_{1\text{c}} = \{3\}$, $\mathcal{D}_{1\text{c}} = \emptyset$, $\mathcal{C}_{2\text{c}} = \{2, 4\}$, and $\mathcal{D}_{2\text{c}} = \{3\}$. We also had $\mathcal{C}_{\text{ec}} = \{1\}$, $\mathcal{C}_{\text{rc}} = \{4\}$, $\mathcal{C}_{\text{at}} = \mathcal{C}_{\text{nt}} = \emptyset$, and $\mathcal{D}_g = \emptyset$ for each $g \in \{\text{at}, \text{nt}, \text{ec}, \text{rc}\}$.

As before, consider the same 2SLS specification used by Imbens and Angrist (1994, Theorem 2) with a saturated first stage, and a second stage that contains D_i and a constant. Let β_{2sls} denote the 2SLS estimand corresponding to the coefficient on D_i. The following proposition provides an interpretation of the 2SLS estimand.

Proposition 7. *Suppose Z_i takes K values $\{z^1, \ldots, z^K\}$ labeled so that the propensity score is increasing and suppose that the support of Z_i is rectangular, that is* $\text{supp}(Z_i) = \text{supp}(Z_{i,1}) \times \cdots \times \text{supp}(Z_{i,L})$. *If Assumptions PM and E are satisfied, and if β_{2sls} exists, then*

$$\beta_{2sls} = \sum_{g \in \mathcal{G} : \mathcal{C}_g \neq \emptyset} \tau_g \Delta_g,$$

where τ_g are weights such that $\sum_{g \in \mathcal{G} : \mathcal{C}_g \neq \emptyset} \tau_g = 1$, and

$$\text{sgn}(\tau_g) = \mathbb{1}[\pi_g > 0] \times \text{sgn}\left(\sum_{k=2}^{K} \left(\mathbb{1}[k \in \mathcal{C}_g] - \mathbb{1}[k \in \mathcal{D}_g] \right) \text{Cov}\left(D_i, \mathbb{1}[p(Z_i) \geq p(z^k)] \right) \right).$$

Proposition 7 shows that under Assumption PM, the 2SLS estimator produces a weighted average of treatment effects for groups that comply with some instrument change. The weights on each group could be positive or negative, but this can be checked empirically. To do this, one must generate the sets $\mathcal{C}_g$ and $\mathcal{D}_g$ by applying

Assumption PM to the set of K instrument values. The sign of the weight for group g is then determined by the overall sum of $\text{Cov}(D_i, \mathbb{1}[p(Z_i) \geq p(z^k)]$ for instrument values k at which they comply less the sum of these terms at values k for which they defy. The additional rectangular support condition is necessary to ensure that the contrasts picked up in β_{2sls} are restricted by Assumption PM. For example, if in the special case in Section 3.2 the support of Z_i were only $\{(0,1),(1,0)\}$, then either the Z_1 compliers or Z_2 compliers would always have negative weight. This is intuitive since Assumption PM does not place any direct restrictions on behavior in the contrast between $(0,1)$ and $(1,0)$.

4 Identification with Multiple Instruments

In the previous section, we showed that under Assumption PM, the 2SLS estimand can still identify a positive weighted average of causal effects over exhaustive and mutually exclusive subpopulations. However, even when this is so, the 2SLS estimand will not necessarily answer a policy counterfactual of interest to the researcher. This is a consequence of restricting oneself to the 2SLS estimator, which weights groups based on statistical considerations rather than economic considerations. In this section, we develop a new general framework for causal inference with multiple instruments under Assumption PM.

4.1 Instruments and Treatment Choices

As before, we let $Z_i \equiv (Z_{i,1}, \ldots, Z_{i,L})$ denote a vector of L instruments. In this section, we assume for notational simplicity that $\text{supp}(Z_i) = \mathcal{Z}$. However, we will not place any restrictions on the joint distribution of Z_i, and in particular its components may be discrete or continuous, and dependent with each other in arbitrary ways. As in Section 2.4, we let $Z_{i,-\ell}$ denote the $(L-1)$–dimensional random vector formed by removing the ℓth component $(Z_{i,\ell})$ from Z_i. We denote the supports of $Z_{i,\ell}$ and $Z_{i,-\ell}$ by $\mathcal{Z}_\ell$ and $\mathcal{Z}_{-\ell}$, respectively.

For each ℓ, the potential treatments $\{D_i(z)\}_{z \in \mathcal{Z}}$ generate a collection of marginal potential treatments, defined as

$$D_{i,\ell}(z_\ell) \equiv \sum_{z_{-\ell} \in \mathcal{Z}_{-\ell}} \mathbb{1}[Z_{i,-\ell} = z_{-\ell}] D_i(z_\ell, z_{-\ell}) \equiv D_i(z_\ell, Z_{i,-\ell}). \tag{6}$$

The marginal potential treatments represent individual i's treatment choice if the ℓth component of Z_i had been set to z_ℓ, but all other components had remained at their observed realizations. Conditional on $Z_{i,-\ell}$, each marginal potential treatment is equal to a single joint potential treatment:

$$\mathbb{P}[D_{i,\ell}(z_\ell) = D_i(z_\ell, z_{-\ell}) | Z_{i,-\ell} = z_{-\ell}] = 1. \tag{7}$$

Thus, if Assumption PM is satisfied, then each set of potential treatments $\{D_{i,\ell}(z_\ell)\}_{z_\ell \in \mathcal{Z}_\ell}$ satisfies Assumption IAM conditional on any realization of $Z_{i,-\ell}$.

In addition to Assumption PM, we continue to maintain Assumption E. Notice that Assumption E *does not* imply that $\{D_{i,\ell}(z_\ell)\}_{z_\ell \in \mathcal{Z}_\ell} \perp\!\!\!\perp Z_i$. This is because $D_{i,\ell}(z_\ell)$ depends on $Z_{i,-\ell}$, which is itself a subvector of Z_i. However, Assumption E does imply that

$$\{D_{i,\ell}(z_\ell)\}_{z_\ell \in \mathcal{Z}_\ell} \perp\!\!\!\perp Z_{i,\ell} | Z_{i,-\ell} \text{ for every } \ell. \tag{8}$$

This observation combined with (7) suggests considering L different models of treatment choice, where in the ℓth model the instrument is the scalar $Z_{i,\ell}$, and the other components, $Z_{i,-\ell}$, are conditioned on as control variables.

4.2 Selection Equations

We formulate the models of treatment choice using Vytlacil's (2002) equivalence result. This result shows that for every ℓ there exists a scalar latent variable $V_{i,\ell}$ that satisfies $V_{i,\ell} \perp\!\!\!\perp Z_{i,\ell} | Z_{i,-\ell}$, and such that

$$D_{i,\ell}(z_\ell) = \mathbb{1}\left[V_{i,\ell} \leq \eta_\ell(z_\ell, Z_{i,-\ell})\right], \tag{9}$$

where η_ℓ is a function that does not vary across i. Using the standard normalization, this model is equivalent to

$$D_{i,\ell}(z_\ell) = \mathbb{1}\left[U_{i,\ell} \leq p(z_\ell, Z_{i,-\ell})\right], \tag{10}$$

where $U_{i,\ell}$ is distributed uniformly over $[0, 1]$, conditional on any realization of $Z_{i,-\ell}$, and $p(z) \equiv \mathbb{P}[D_i = 1 | Z_i = z]$ is the propensity score. We maintain this normalization for each model $\ell = 1, \ldots, L$.

The unobservable $U_{i,\ell}$ can be interpreted as individual i's latent propensity to take treatment as measured against the incentive (or disincentive) created by the ℓth instrument. Individuals with lower values of $U_{i,\ell}$ are more prone to take treatment than those with higher values. Individuals with intermediate values of $U_{i,\ell}$ are compliers whose treatment decisions are impacted by shifts in $Z_{i,\ell}$. The choice behavior of each individual is characterized by a vector of unobservables, $(U_{i,1}, \ldots, U_{i,L})$, which measures the propensity to take treatment along different margins.

It is important to emphasize that $U_{i,\ell}$ is only directly comparable within strata defined by the other instruments, $Z_{i,-\ell}$. To see this, return to the example with two binary instruments, and recall the six latent groups shown in Table 1. Conditional on $Z_{i,2} = z_2$, the propensity score $p(Z_{i,1}, z_2)$ takes two values, which divides the unit interval into three sets, $(0, p(0, z_2)]$, $(p(0, z_2), p(1, z_2)]$, and $(p(1, z_2), 1]$. An individual with $U_{i,1}$ in the first set takes treatment even if $Z_{i,1} = 0$. They must be an always-taker if $z_2 = 0$. However, if $z_2 = 1$, then they could be an always-taker, an eager complier, a reluctant complier, or a Z_2 complier. This underscores the importance of conditioning on $Z_{i,-\ell}$, which is a consequence of using Assumption PM instead of Assumption IAM.

Another novel consequence of considering Assumption PM is that each selection model provides a different representation of the same observed treatment status. That is:

$$D_i = \sum_{z_\ell \in \mathcal{Z}_\ell} \mathbb{1}[Z_{i,\ell} = z_\ell] D_{i,\ell}(z_\ell) = \mathbb{1}\left[U_{i,\ell} \leq p(Z_i)\right] \text{ for every } \ell = 1, \ldots, L. \tag{11}$$

Thus, replacing Assumption IAM with Assumption PM requires replacing threshold crossing models like (9) with multiple hurdle models like (11).[16] Notice that (11) will only be satisfied if the joint distribution of $(U_{i,1}, \ldots, U_{i,L})$ satisfies some particular properties. For example, if $L = 2$, then it necessary for $\mathbb{P}[U_{i,1} \leq p(z), U_{i,2} > p(z) | Z_i = z] = 0$ for any z, since otherwise we would have the logical contradiction that $D_i = 1$ and $D_i = 0$. These properties get incorporated into our analysis through the concept of "logical consistency" which is discussed in the next section.

Instead of deriving (11) from Assumption PM, one can also derive it directly from a threshold crossing equation that satisfies Assumption PM. For example, suppose that $L = 2$ with potential outcomes determined by a random coefficients specification of

[16] See Poirier (1980) and the related models in Heckman (1978). Lee and Salanié (2018) have recently considered marginal treatment effect methods when the selection equation is assumed to have a double hurdle model form like (11) with $L = 2$. Their approach takes the double hurdle model as primitive, and requires maintaining assumptions sufficiently strong to point identify the joint distribution of its latent variables. In contrast, our approach derives the double hurdle model from Assumption PM, and we will not need such assumptions. However, our work can be viewed as sharing their motivation of allowing for richer unobserved heterogeneity in treatment choice than is allowed by Assumption IAM. In this regard, our work is also related to a recent active literature that focuses on providing weaker models of choice behavior for use with discrete, unordered treatments. For example, see Heckman, Urzua, and Vytlacil (2008), Kirkeboen et al. (2016), Kline and Walters (2016), Mountjoy (2018), Heckman and Pinto (2018), and Kamat (2018). None of this literature addresses the restrictions on choice behavior implied by having multiple instruments.

indirect utility as in (4). That is,

$$D_i(z_1, z_2) = \mathbb{1}[B_{i,0} + B_{i,1} z_1 + z_2 \geq 0], \tag{12}$$

where $(B_{i,0}, B_{i,1}) \perp\!\!\!\perp (Z_{i,1}, Z_{i,2})$. As shown in Proposition 2, this specification leads to a violation of Assumption IAM unless $\text{Var}(B_{i,1}) = 0$. However, Proposition 3 implies that Assumption PM will be satisfied as long as $B_{i,1} \geq 0$ (almost surely).

From (12), the two pre-normalized selection equations (9) can be derived as

$$D_{i,1}(z_1) = \mathbb{1}\Big[\overbrace{-\frac{(B_{i,0} + Z_{i,2})}{B_{i,1}}}^{\equiv V_{i,1}} \leq \overbrace{z_1}^{\equiv \eta_1}\Big] \quad \text{and} \quad D_{i,2}(z_2) = \mathbb{1}\Big[\overbrace{-(B_{i,0} + B_{i,1} Z_{i,1})}^{\equiv V_{i,2}} \leq \overbrace{z_2}^{\equiv \eta_2}\Big].$$

Notice in particular that even though $(B_{i,0}, B_{i,1})$ is independent of $(Z_{i,1}, Z_{i,2})$, this will not be the case for $(V_{i,1}, V_{i,2})$. Instead, $V_{i,1}$ is dependent with $Z_{i,2}$, and in general only independent with $Z_{i,1}$ after conditioning on $Z_{i,2}$. Similarly, $V_{i,2}$ is dependent with $Z_{i,1}$ with independence between $V_{i,1}$ and $Z_{i,2}$ only guaranteed after conditioning on $Z_{i,1}$. In addition, $V_{i,1}$ and $V_{i,2}$ are clearly dependent, since they are both functions of $B_{i,0}$ and $B_{i,1}$. The joint distribution of the normalized selection unobservables, $(U_{i,1}, U_{i,2})$, is effectively the copula of $(V_{i,1}, V_{i,2})$, and so will also be dependent.

4.3 Marginal Treatment Response Functions

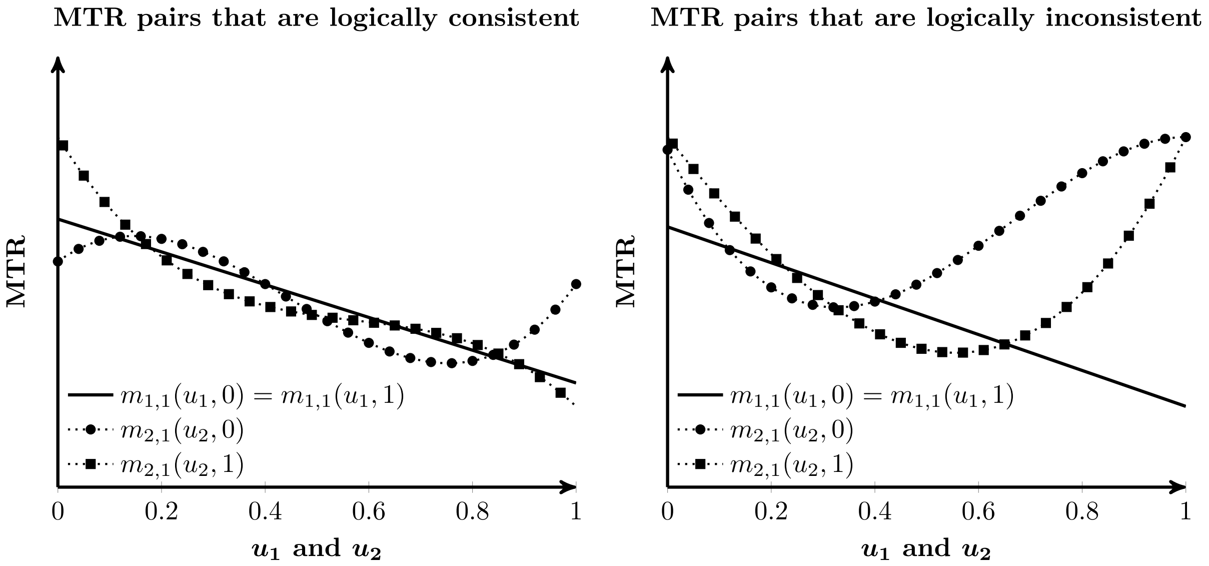
For each ℓ, we define a pair of marginal treatment response (MTR) functions,

$$m_{\ell,d}(u_\ell, z_{-\ell}) \equiv \mathbb{E}\left[Y_i(d) | U_{i,\ell} = u_\ell, Z_{i,-\ell} = z_{-\ell}\right] \quad \text{for } d = 0, 1. \tag{13}$$

The MTR functions describe variation in outcomes as a function of the propensity to take treatment along the ℓth margin, $U_{i,\ell}$, again conditioning on all other instruments, $Z_{i,-\ell} = z_{-\ell}$. We denote each pair of MTR functions as $m_\ell \equiv (m_{\ell,0}, m_{\ell,1})$. Each pair m_ℓ generates a marginal treatment effect (MTE) function (Heckman and Vytlacil, 1999, 2001, 2005, 2007a,b) formed as $m_{\ell,1}(u_\ell, z_{-\ell}) - m_{\ell,0}(u_\ell, z_{-\ell})$. We let $m \equiv (m_1, \ldots, m_L)$, and assume that m belongs to a known parameter space, $\mathcal{M} \equiv \mathcal{M}_1 \times \cdots \times \mathcal{M}_L$, which reflects prior information (assumptions) that the researcher wants to impose about the MTR pairs.[17]

Each MTR pair and corresponding MTE function is defined in terms of a different margin of selection, $U_{i,\ell}$, which is itself defined by the ℓth instrument component. Since the MTR pairs are instrument-specific, they are not directly comparable. However, a

[17] We assume throughout that each $\mathcal{M}_\ell$ is contained in a vector space.



(a) *Both $m_{1,1}$ and $m_{2,1}$ imply the same value of $\mathbb{E}[Y_i(1)]$. These MTR pairs are logically consistent.*

(b) *The value of $\mathbb{E}[Y_i(1)]$ implied by $m_{1,1}$ differs from that implied by $m_{2,1}$. These MTR pairs are logically inconsistent.*

Figure 4: *MTR pairs along different margins of selection are not directly comparable. Nevertheless, they are not completely unrelated, since both pairs provide a description of the entire population.*

key point for our discussion ahead is that each MTR pair still describes the entire population, just organized along a different dimension of choice behavior. Thus, while the MTR pairs for different ℓ will typically be different, they cannot be arbitrarily different.

For example, Figure 4a plots $m_{1,1}(\cdot, z_2)$ for a case in which this function does not vary with z_2. It also plots $m_{2,1}(\cdot, z_1)$ for both $z_1 = 0$ and $z_1 = 1$. While $m_{1,1}$ is not directly comparable to $m_{2,1}$, both functions are a conditional mean for the same random variable, $Y_i(1)$. To be logically consistent, it should be the case that

$$\mathbb{E}\left[m_{1,1}(U_{i,1}, Z_{i,2})\right] = \mathbb{E}[Y_i(1)] = \mathbb{E}\left[m_{2,1}(U_{i,2}, Z_{i,1})\right],$$

so that both functions generate the same mean for $Y_i(1)$. In Figure 4a this is the case, since the integrals of both dotted curves are the same as the integral of the solid line.

In contrast, the functions $m_{2,1}(\cdot, z_1)$ in Figure 4b are not logically consistent. The areas under $m_{2,1}(\cdot, 0)$ and $m_{2,1}(\cdot, 1)$ are clearly greater than the area under $m_{1,1}$. These MTR pairs cannot both be describing conditional means for $Y_i(1)$, since they would imply different values of $\mathbb{E}[Y_i(1)]$. In the following, we will develop a method that *requires* logical consistency, so that pairs like those in Figure 4b are excluded from consideration.

4.4 The Target Parameter

We assume that the researcher has a well-posed empirical question that can be informed by a specific target parameter, $\beta^\star$. We require the target parameter to be expressed as a weighted average of the L MTR pairs in the form

$$\beta^\star(m) = \sum_{\ell=1}^{L} \beta^\star_\ell(m_\ell) \equiv \sum_{\ell=1}^{L} \sum_{d\in\{0,1\}} \mathbb{E}\left[\int_0^1 m_{\ell,d}(u, Z_{i,-\ell})\omega^\star_{\ell,d}(u, Z_i)\, du\right], \tag{14}$$

where $\omega^\star_{\ell,d}$ are weighting functions. The weighting functions are assumed to be known given knowledge of the joint distribution of (Y_i, D_i, Z_i). For catalogues of common weighting functions, see Heckman and Vytlacil (2005, 2007b), Mogstad et al. (2018, "MST" hereafter), and Mogstad and Torgovitsky (2018). When $L = 1$, (14) reduces to the form used for the target parameter by MST.

When $L > 1$, there might be several ways to express the same target parameter. For example, if $\beta^\star$ is the population average treatment effect (ATE), $\mathbb{E}[Y_i(1) - Y_i(0)]$, then one could take $\omega^\star_{\ell,0}(u, z_{-\ell}) = -1$ and $\omega^\star_{\ell,1}(u, z_{-\ell}) = 1$ for any ℓ, while setting all other weight functions to 0. This is another manifestation of the logical consistency issue illustrated in Figure 4. When using multiple instruments, the way we incorporate this idea will require the implied value of the ATE to be the same for any ℓ. Thus, as a practical matter, any choice of ℓ will yield the same inference on an instrument-invariant parameter like the ATE or the average effect of treatment on the treated (ATT).

Other target parameters might be instrument-specific. For example, the class of policy-relevant treatment effects (PRTE) introduced by Heckman and Vytlacil (1999, 2005) includes parameters that measure the impact of changing the incentive associated with a given instrument. A special case of a PRTE is an extrapolated LATE, for example

$$\text{LATE}_{1,+\delta\%} \equiv \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, p(0, Z_{i,2}) < U_{i,1} \leq \left(1 + \frac{\delta}{100}\right) \times p(1, Z_{i,2})\right], \tag{15}$$

which is the LATE that would result if the $Z_{i,1}$ instrument were changed sufficiently to cause a $\delta\%$ increase in participation under $Z_{i,1} = 1$. This target parameter can be used to gauge the sensitivity of point identified IV estimates to the definition of the complier group. See Heckman and Vytlacil (2005), Carneiro, Heckman, and Vytlacil

(2010), and MST for further discussion and additional examples of PRTEs.

When the definition of the target parameter depends on the instrument as in (15), there will only be a single set of weights that delivers the desired target parameter. Nevertheless, we still want to require instrument-invariant parameters to be logically consistent across different MTR pairs. As we demonstrate ahead, this requirement will allow information to flow from one model to another. The surprising implication is that even if the target parameter is instrument-specific, there can still be a benefit from combining multiple instruments.

4.5 Using One Instrument at a Time

We briefly recall the procedure in MST by describing how to use each instrument separately for inference about $\beta^\star$.

If $(Y_i(0), Y_i(1), D_i)$ were generated by (9) for any ℓ, with MTR pair $m_\ell \equiv (m_{\ell,0}, m_{\ell,1})$, then it must be the case that

$$\mathbb{E}[s(D_i, Z_i)Y_i] = \sum_{d\in\{0,1\}} \mathbb{E}\left[\int_0^1 m_{\ell,d}(u, Z_{i,-\ell})\omega_{d,s}(u, Z_i)\, du\right] \tag{16}$$

$$\text{where} \quad \omega_{0,s}(u, Z_i) \equiv s(0, Z_i)\mathbb{1}[u > p(Z_i)] \quad \text{and} \quad \omega_{1,s}(u, Z_i) \equiv s(1, Z_i)\mathbb{1}[u \le p(Z_i)],$$

for any (measurable) function, s. MST refer to a choice of s as an IV–like specification, and show that by choosing s appropriately, one can reproduce any linear IV estimand on the left-hand side of (16). Given a collection $\mathcal{S}$ of IV–like specifications, we say that an MTR pair m_ℓ is observationally equivalent under $\mathcal{S}$ if it satisfies (16) for every $s \in \mathcal{S}$ and each $\ell = 1, \dots, L$. We denote the set of such pairs by

$$\mathcal{M}_\ell^{\text{obs}} \equiv \{m_\ell : m_\ell \text{ satisfies (16) for each } s \in \mathcal{S}\}.$$

The identified set for the ℓth MTR pair is defined as

$$\mathcal{M}_\ell^{\text{id}} \equiv \mathcal{M}_\ell \cap \mathcal{M}_\ell^{\text{obs}}.$$

$\mathcal{M}_\ell^{\text{id}}$ is the collection of MTR pairs for the ℓth instrument that satisfy the researcher's prior assumptions ($m_\ell \in \mathcal{M}_\ell$) and are observationally equivalent ($m_\ell \in \mathcal{M}_\ell^{\text{obs}}$) for the choice of IV–like estimands in $\mathcal{S}$. The identified set for the ℓth component of the target parameter is

$$\mathcal{B}_\ell^{\text{id}} \equiv \left\{\beta_\ell^\star(m_\ell) : m_\ell \in \mathcal{M}_\ell^{\text{id}}\right\}.$$

If $\mathcal{M}_\ell$ is a convex set, then $\mathcal{B}^{\text{id}}_\ell$ is an interval, $[\underline{\beta}^\star_\ell, \overline{\beta}^\star_\ell]$, with endpoints given by

$$\underline{\beta}^\star_\ell \equiv \inf_{m_\ell \in \mathcal{M}^{\text{id}}_\ell} \beta^\star_\ell(m_\ell) \qquad \text{and} \qquad \overline{\beta}^\star_\ell \equiv \sup_{m_\ell \in \mathcal{M}^{\text{id}}_\ell} \beta^\star_\ell(m_\ell).$$

MST show that if $\mathcal{M}$ can be represented as a finite linear basis, then $\underline{\beta}^\star_\ell$ and $\overline{\beta}^\star_\ell$ can be computed using linear programming. As a particular case of this, they also show that exact nonparametric bounds can be computed by using a constant spline with appropriately chosen knot points. As shown in MST, if $\mathcal{S}$ is a sufficiently rich class of functions, then $\mathcal{B}^{\text{id}}_\ell$ is the smallest set of values for the target parameter that are consistent with both the maintained assumptions ($m_\ell \in \mathcal{M}_\ell$) and the conditional means of Y. Extending this argument to the present setting is straightforward.

4.6 Combining Instruments

The observational equivalence condition (16) restricts each of the L pairs of MTR functions in isolation. We connect them by requiring logical consistency in the unobservable quantities they imply. For example, for every ℓ, a choice of $m_{\ell,1}$ implies a value for $\mathbb{E}[Y_i(1)]$ given by

$$\mathbb{E}[Y_i(1)] = \mathbb{E}\left[\int_0^1 m_{\ell,1}(u, Z_{i,-\ell})\, du\right]. \tag{17}$$

We will restrict attention to choices of m for which the right-hand side of (17) is invariant to $\ell = 1, \ldots, L$.[18] This restricts our attention to MTR pairs like those in Figure 4a, while ruling out inconsistent pairs like those in Figure 4b. The result will be tighter inference on each $\beta^\star_\ell$, as well as on the overall target parameter, $\beta^\star$.

We formalize the property of logical consistency in a similar fashion to the observational equivalence condition, (16). Specifically, given a collection $\mathcal{S}$ of IV–like specifications, we say that a collection of MTR functions $m \equiv (m_1, \ldots, m_L)$ is logically

[18] This is similar in spirit to the concept of a "coherent model" (e.g. Heckman, 1978; Tamer, 2003; Lewbel, 2007; Chesher and Rosen, 2012). However, it is different because (17) is an unobservable quantity—not a feature of the observed data—and so one could proceed without requiring (17) to be invariant to ℓ as in the previous section. Note that Maddala (1983, Section 7.5) uses the phrase "logical consistency" to describe a coherency condition in a simultaneous binary response model, so our use of this phrase differs from his.

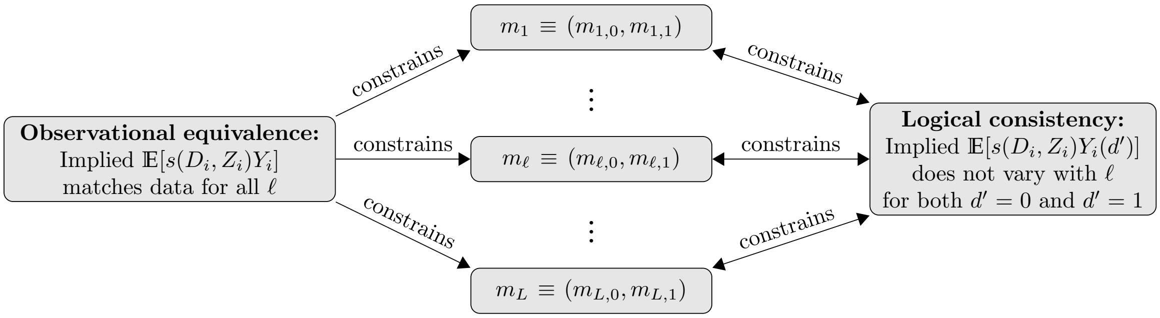

$m_1 \equiv (m_{1,0}, m_{1,1})$
constrains
constrains
Observational equivalence:
Implied $\mathbb{E}[s(D_i, Z_i)Y_i]$
matches data for all ℓ
constrains
$m_\ell \equiv (m_{\ell,0}, m_{\ell,1})$
constrains
Logical consistency:
Implied $\mathbb{E}[s(D_i, Z_i)Y_i(d')]$
does not vary with ℓ
for both $d' = 0$ and $d' = 1$
constrains
constrains
$m_L \equiv (m_{L,0}, m_{L,1})$

Figure 5: *Observational equivalence* (16) *constrains each* $m_\ell \equiv (m_{\ell,0}, m_{\ell,1})$ *in isolation. Logical consistency* (18) *ties the* m_ℓ *together across* $\ell = 1, \ldots, L$*. This allows the information contained in different instruments to flow in the direction of the arrows, and therefore be combined across models that use different instruments.*

consistent under $\mathcal{S}$ if

$$\overbrace{\sum_{d\in\{0,1\}} \mathbb{E}\left[\int m_{\ell,d'}(u, Z_{i,-\ell})\omega_{d,s}(u, Z_i)\, du\right]}^{\mathbb{E}[s(D_i,Z_i)Y_i(d')] \text{ implied by } m_\ell} = \overbrace{\sum_{d\in\{0,1\}} \mathbb{E}\left[\int m_{\ell',d'}(u, Z_{i,-\ell'})\omega_{d,s}(u, Z_i)\, du\right]}^{\mathbb{E}[s(D_i,Z_i)Y_i(d')] \text{ implied by } m_{\ell'}}$$
$$\text{for all } s \in \mathcal{S}, \text{ all } d' \in \{0,1\}, \text{ and all } \ell, \ell' \in \{1, \ldots, L\} \text{ with } \ell \neq \ell', \tag{18}$$

where $\omega_{d,s}$ is still defined as in (16). Unlike observational equivalence, which involves the observed outcome, Y_i, logical consistency involves the implied conditional means of the unobserved potential outcomes, $Y_i(0)$, and $Y_i(1)$. This is why (18) is stated for each $d' \in \{0, 1\}$, with the weights changing in the summation over $d \in \{0, 1\}$. This is also the reason that (18) is directly stated as an equality across different MTR pairs (ℓ and ℓ'), whereas in (16) this equality is implied by matching each MTR pair to the same observable quantity on the left-hand side.

Given a set of IV–like specifications, $\mathcal{S}$, the set of logically consistent MTR pairs is

$$\mathcal{M}^{\text{lc}} \equiv \{m \equiv (m_1, \ldots, m_L) : m \text{ satisfies } (18)\}.$$

To combine multiple instruments together, we focus on the identified set

$$\mathcal{M}^{\text{id}} \equiv \mathcal{M} \cap \mathcal{M}^{\text{obs}} \cap \mathcal{M}^{\text{lc}},$$

where $\mathcal{M}^{\text{obs}} \equiv \mathcal{M}_1^{\text{obs}} \times \cdots \times \mathcal{M}_L^{\text{obs}}$. The identified set for the target parameter is

$$\mathcal{B}^{\text{id}} \equiv \left\{ \beta^\star(m) : m \in \mathcal{M}^{\text{id}} \right\}.$$

Figure 5 illustrates how the logical consistency condition allows information to flow between different selection models. Intuitively, given a set of assumptions, the observational equivalence condition (16) places restrictions on m_ℓ for each ℓ, while the logical consistency condition propagates these restrictions from m_ℓ to $m_{\ell'}$. The result is a sort of equilibrium in which none of the MTR functions contradict each other on their implications for the instrument-invariant quantities $\mathbb{E}[s(D_i, Z_i)Y_i(d')]$ equated in (18). Limiting attention to this smaller set of MTR pairs that are consistent with this equilibrium mechanically shrinks the identified set for the target parameter as well.

The logical consistency condition is a collection of linear equality constraints, so adding it does not fundamentally alter the procedure in MST. In particular, if $\mathcal{M}$ is a convex set, then $\mathcal{B}^{\text{id}}$ is an interval, $[\underline{\beta}^\star, \overline{\beta}^\star]$, with endpoints given by

$$\underline{\beta}^\star \equiv \inf_{m \in \mathcal{M}^{\text{id}}} \Gamma^\star(m) \qquad \text{and} \qquad \overline{\beta}^\star \equiv \sup_{m \in \mathcal{M}^{\text{id}}} \Gamma^\star(m).$$

As a result, if $\mathcal{M}$ can be represented as a finite linear basis, then a straightforward modification of the linear programming procedure developed by MST can be used to compute and/or estimate the identified set for $\beta^\star$.[19] It is also straightforward to extend Proposition 3 in MST to show that if $\mathcal{S}$ is chosen to contain a sufficiently rich class of functions, then $\mathcal{M}^{\text{id}}$ fully exhausts the information contained in the conditional mean of Y given D and Z.

4.7 An Algebraic Example

The content of the logical consistency condition can be illustrated with an algebraic example. Suppose that $Z_i \equiv (Z_{i,1}, Z_{i,2})$ is binary, so that $\mathcal{Z} = \{0,1\}^2$, and that Assumptions E and PM are satisfied. This setting gives rise to two selection equations like (9) with unobservables $U_{i,1}$ and $U_{i,2}$, and therefore two pairs of marginal treatment response functions, m_1 and m_2. To simplify the discussion, we will focus solely on the MTR functions for the treated state, $d = 1$, so that our objects of concern are $m_{1,1}(u, z_2)$ and $m_{2,1}(u, z_1)$, viewed as functions of $(u, z_2) \in [0,1] \times \{0,1\}$ and $(u, z_1) \in [0,1] \times \{0,1\}$, respectively.

[19] In particular, the procedure just needs to be modified to include all selection models, $\ell = 1, \ldots, L$, and to also incorporate the logical consistency constraints (18), which are linear.

Suppose that we assume $m_{1,1}$ is a linear function of u for each value of z_2, i.e.

$$m_{1,1}(u_1, z_2) = \alpha_0 + \alpha_1 u_1 + \alpha_2 z_2 + \alpha_3 z_2 u_1, \tag{19}$$

for some unknown parameters $\alpha \equiv (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$. Brinch, Mogstad, and Wiswall (2012; 2017) showed that α is point identified as long as $p(1, 0) > p(0, 0)$, and $p(1, 1) > p(0, 1)$. Their argument uses the implications of (19) for the observed mean of the treated group:

$$\begin{aligned}
&\mathbb{E}[Y_i | D_i = 1, Z_{i,1} = z_1, Z_{i,2} = z_2] \\
&\quad = \mathbb{E}[Y_i(1) | U_{i,1} \leq p(z_1, z_2), Z_{i,2} = z_2] \\
&\quad = \frac{1}{p(z_1, z_2)} \int_0^{p(z_1,z_2)} m_{1,1}(u, z_2)\, du \\
&\quad = \alpha_0 + \frac{1}{2} p(z_1, z_2)\alpha_1 + z_2 \left[\alpha_2 + \frac{1}{2} p(z_1, z_2)\alpha_3\right].
\end{aligned} \tag{20}$$

Intuitively, if $p(1, 0) > p(0, 0)$, then α_0 and α_1 are point identified by a linear regression of Y_i on a constant and $\frac{1}{2} p(Z_{i,1}, Z_{i,2})$ in the $Z_{i,2} = 0$ group, while $p(1, 1) > p(0, 1)$ ensures that α_2 and α_3 can then be point identified off of the same linear regression in the $Z_{i,2} = 1$ group.

The logical consistency condition is based on the observation that (19) also has implications for the conditional mean of the treated outcome for the *untreated* group. This quantity is not observed, but it can be expressed in terms of α using an argument similar to (20):

$$\begin{aligned}
&\mathbb{E}[Y_i(1) | D_i = 0, Z_{i,1} = z_1, Z_{i,2} = z_2] \\
&\quad = \alpha_0 + \frac{1}{2}\left(1 + p(z_1, z_2)\right)\alpha_1 + z_2 \left[\alpha_2 + \frac{1}{2}\left(1 + p(z_1, z_2)\right)\alpha_3\right].
\end{aligned} \tag{21}$$

Since α is point identified, these counterfactual mean outcomes are also point identified. They could be used to evaluate treatment parameters for the first selection model with unobservable $U_{i,1}$. The more surprising finding is that these counterfactual means can also be used as additional identifying information for the selection model with unobservable $U_{i,2}$.

One way to see this is to consider a specification for $m_{2,1}$ that would typically not be point identified in the current setting. For example, suppose that

$$m_{2,1}(u_2, z_1) = \gamma_0 + \gamma_1 u_2 + \gamma_2 z_1 + \gamma_3 z_1 u_2 + \gamma_4 u_2^2, \tag{22}$$

so that $m_{2,1}$ is more flexible than $m_{1,1}$ in having an additional quadratic term. While this MTR function now has five unknown parameters, $\gamma \equiv (\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5)$, there are still only four observed conditional means: $\mathbb{E}[Y_i|D_i = 1, Z_{i,1} = z_1, Z_{i,2} = z_2]$ for $(z_1, z_2) \in \{0,1\}^2$. If the selection model for $U_{i,2}$ were viewed in isolation, then γ would not be point identified. However, the logical consistency condition effectively provides four more moments via (21). Since α is point identified, these moments can be treated as known.

With eight moments total, it is possible to point identify (indeed, overidentify) the five parameters in γ. In analogy to (20) and (21), the system of equations is given by:

$$\begin{bmatrix} 1 & \frac{p(0,0)}{2} & 0 & 0 & \frac{p(0,0)^2}{3} \\ 1 & \frac{p(1,0)}{2} & 0 & 0 & \frac{p(1,0)^2}{3} \\ 1 & \frac{p(0,1)}{2} & 1 & \frac{p(0,1)}{2} & \frac{p(0,1)^2}{3} \\ 1 & \frac{p(1,1)}{2} & 1 & \frac{p(1,1)}{2} & \frac{p(1,1)^2}{3} \\ 1 & \frac{1+p(0,0)}{2} & 0 & 0 & \frac{1-p(0,0)^3}{3(1-p(0,0))} \\ 1 & \frac{1+p(1,0)}{2} & 0 & 0 & \frac{1-p(1,0)^3}{3(1-p(1,0))} \\ 1 & \frac{1+p(0,1)}{2} & 1 & \frac{(1+p(0,1))}{2} & \frac{1-p(0,1)^3}{3(1-p(0,1))} \\ 1 & \frac{1+p(1,1)}{2} & 1 & \frac{(1+p(1,1))}{2} & \frac{1-p(1,1)^3}{3(1-p(1,1))} \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \end{bmatrix} = \begin{bmatrix} \mathbb{E}[Y_i|D_i = 1, Z_i = (0,0)] \\ \mathbb{E}[Y_i|D_i = 1, Z_i = (1,0)] \\ \mathbb{E}[Y_i|D_i = 1, Z_i = (0,1)] \\ \mathbb{E}[Y_i|D_i = 1, Z_i = (1,1)] \\ \mathbb{E}[Y_i(1)|D_i = 0, Z_i = (0,0)] \\ \mathbb{E}[Y_i(1)|D_i = 0, Z_i = (1,0)] \\ \mathbb{E}[Y_i(1)|D_i = 0, Z_i = (0,1)] \\ \mathbb{E}[Y_i(1)|D_i = 0, Z_i = (1,1)] \end{bmatrix}.$$

The entire right-hand side is known: The first four quantities are observed in the data, and the second set of four are identified from the first model via (21), since α is point identified. The coefficient matrix on the left-hand side can be full rank, depending on the values of the propensity score.[20] When this is the case, the linear system of equations either has no solution, or a unique solution. If there is no solution, then the model is misspecified, while if there is a unique solution, then γ is point identified.[21] Thus, the quadratic MTR specification (22) can be point identified even though the only source of exogenous variation in the second selection model is the binary instrument, $Z_{i,1}$.

4.8 A Numerical Simulation

The logic of the previous section does not require $m_{1,1}$ to be point identified. When $m_{1,1}$ is partially identified, then its implied values of $\mathbb{E}[Y_i(1)|D_i = 0, Z_i = z]$ will also be partially identified. However, logical consistency will still require $m_{2,1}$ to be

[20] For example, take $p(0,0) = .3$, $p(1,0) = .45$, $p(0,1) = .55$, and $p(1,1) = .7$.

[21] It is common to call γ point identified regardless of which case holds, since the identified set consists of no more than a single element for both cases. The ambiguity comes from whether one is tacitly assuming that the model is correctly specified. We maintain a distinction between the two cases here just for clarity.

$z = (z_1, z_2)$	$\mathbb{P}[Z_i = z]$	$p(z)$
$(0, 0)$	.4	.3
$(0, 1)$	.3	.5
$(1, 0)$	.1	.6
$(1, 1)$	.2	.7

Table 2: *The distributions of Z_i and $D_i | Z_i = z$ in the numerical simulation.*

such that its own implied values of these counterfactual moments lie within the same identified set as those implied by $m_{1,1}$. Unless the two sets are the same, this will yield identifying content for $m_{1,1}$ and/or $m_{2,1}$.

We illustrate this possibility with a numerical simulation. The simulation is like the previous example with two binary instruments $Z_i \in \{0, 1\} \times \{0, 1\}$. The distribution of Z_i and the propensity score are shown in Table 2. The propensity score is increasing in each component of Z_i, so that both instruments can be viewed as incentives that make choosing $D_i = 1$ more attractive. We assume that $Y_i \in \{0, 1\}$ is binary, so that conditional expectations of Y_i are bounded between 0 and 1, and we generate the data using model $\ell = 1$ with MTR functions that are linear in u_1:

$$m_{1,0}(u_1, z_2) = .5 - .1u_1 \quad \text{and} \quad m_{1,1}(u_1, z_2) = .8 - .4u_1.$$

In all results that follow, we use a saturated specification of $\mathcal{S}$, so the reported bounds are sharp.

Figure 6 shows bounds on the average treatment on the treated (ATT). These bounds are derived under specifications of $m_\ell \equiv (m_{\ell,0}, m_{\ell,1})$ that are K_ℓth order polynomials in u_ℓ, and fully interacted in $z_{-\ell}$. We implement these polynomials using the Bernstein basis so that it is easy to impose shape constraints. There are three sets of bounds shown for increasing values of $K_1 = K_2$, as well as exact nonparametric bounds indicated with horizontal lines. The exact nonparametric bounds are computed using the constant spline formula developed in Proposition 4 of MST.

The two wider sets of bounds are derived using the $\ell = 1$ and $\ell = 2$ models in isolation. The bounds are different because in the $\ell = 1$ model the instrument is $Z_{i,2}$, with $Z_{i,1}$ serving as a control variable, while in the $\ell = 2$ model the instrument is $Z_{i,1}$, with $Z_{i,2}$ as a control. The third set of bounds is computed while also imposing logical consistency between the two models. This substantially tightens both the nonparametric bounds and the polynomial bounds at all polynomial degrees. Notice in particular that the logical consistency bounds are tighter than the intersections of the $\ell = 1$ and

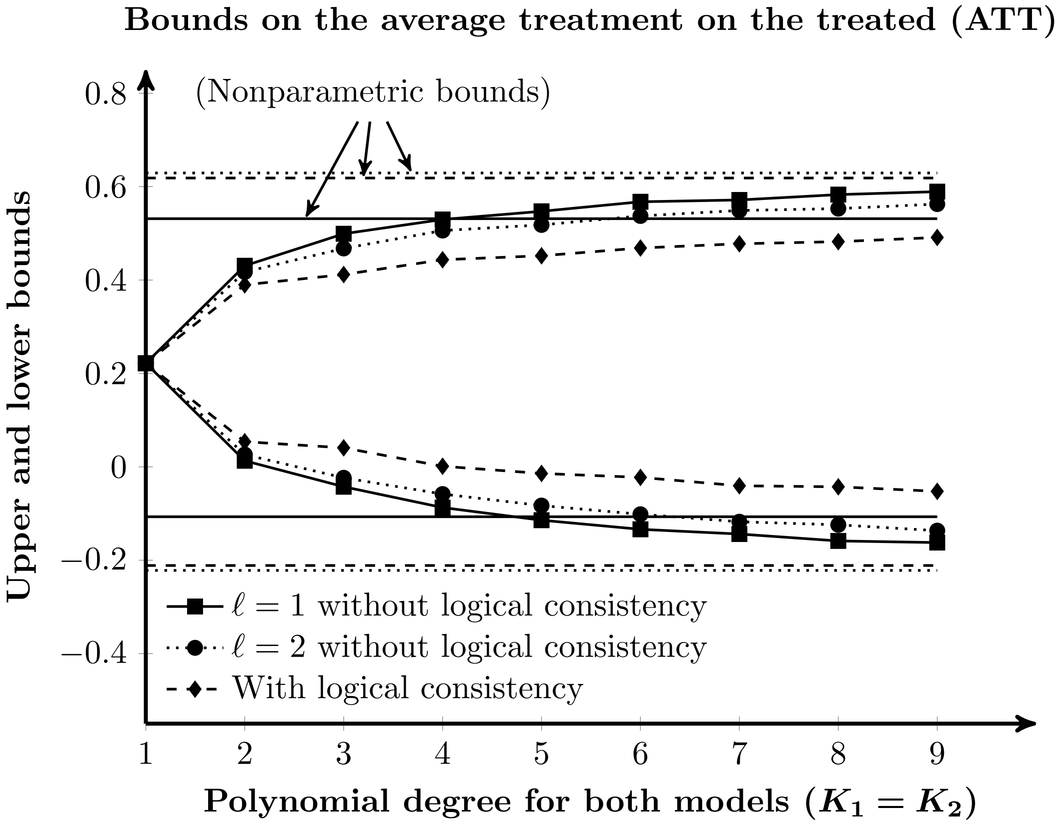


Figure 6: *Imposing logical consistency tightens bounds on the average treatment on the treated (ATT) for both parametric and nonparametric specifications of the MTR functions.*

$\ell = 2$ bounds. Under logical consistency the whole is greater than the sum of its parts.

In Figure 7, we report bounds on $\text{LATE}_{1,\delta\%}$, as defined in (15) for $\delta = 20$. This quantity can only be expressed in terms of the unobservable $U_{i,1}$ for the $\ell = 1$ model. Nevertheless, comparing the four sets of bounds in Figure 7 shows that the $\ell = 2$ model provides information on $\text{LATE}_{1,+20\%}$ through the logical consistency condition. Thus, the logical consistency condition allows information from the $\ell = 2$ model to propagate to the $\ell = 1$ model.

In this data generating process, the additional information is small (but still present) when the $\ell = 2$ model is specified nonparametrically. Adding the nonparametric shape constraints that $m_{2,0}(\cdot, z_1), m_{2,1}(\cdot, z_2)$ and $(m_{2,1} - m_{2,0})(\cdot, z_1)$ are decreasing functions for every z_1 provides substantially more information. An extreme case occurs when the $\ell = 2$ model is specified as linear in u_2 for each value of z_1. Under this assumption, all model-invariant quantities will be point identified by the $\ell = 2$ model. However, a model-specific parameter like $\text{LATE}_{1,+20\%}$ generally remains partially identified. An

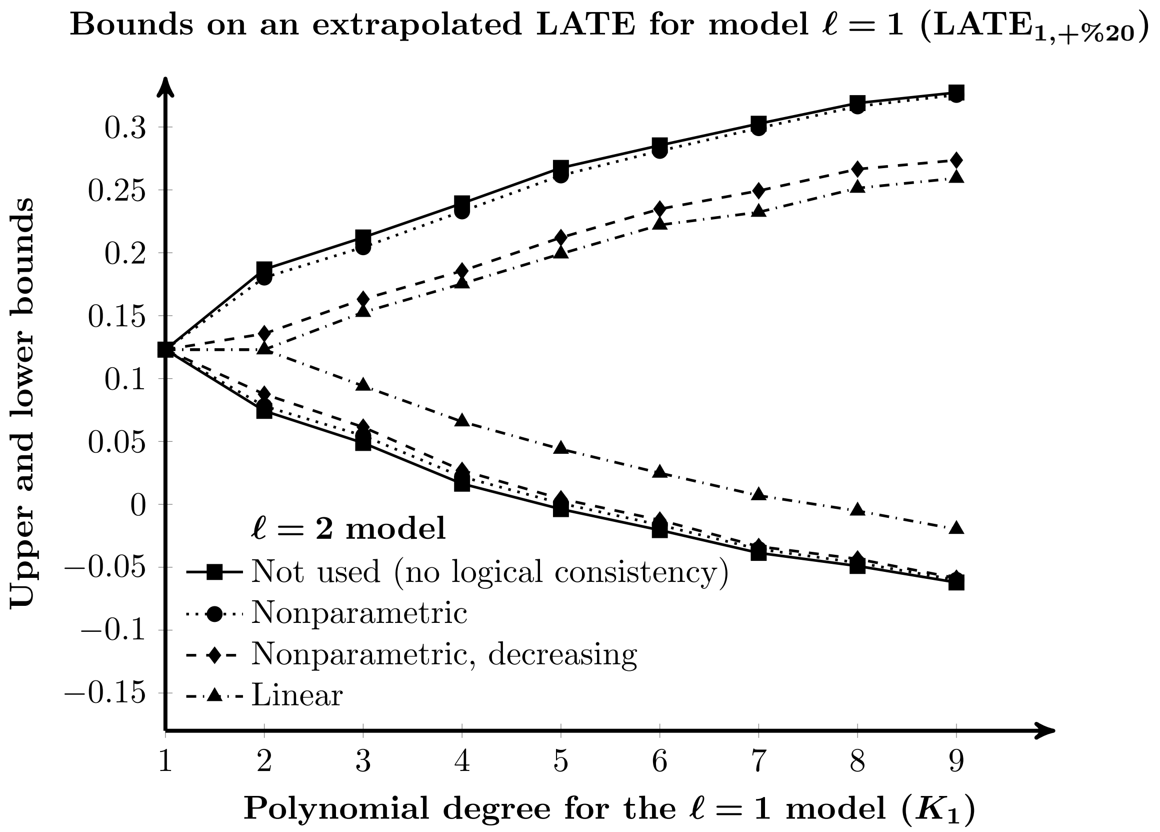


Figure 7: *The $\ell = 2$ model provides identifying content for parameters, such as $LATE_{1,+\%20}$, that can only be defined using the $\ell = 1$ model.*

exception occurs when the $\ell = 1$ model is specified as quadratic. Here, the linearity of the $\ell = 2$ model turns out to be sufficient to achieve point identification of $m_{1,0}$ and $m_{1,1}$ (and therefore $\mathrm{LATE}_{1,+20\%}$) despite the fact that $Z_{i,1}$ is only binary, as suggested by the algebraic example in Section 4.7.

5 Conclusion

The IA monotonicity condition is a cornerstone of modern IV analysis. It is appealed to often, but rarely justified explicitly. As we have shown, it will not hold when there are multiple instruments without severe restrictions on choice heterogeneity. This creates a dilemma for using IV methods to aggregate findings into a larger body of knowledge: Each instrument is associated with a different set of complier groups, but combining multiple instruments together using IA monotonicity requires assuming that these groups are effectively identical in terms of their choice behavior.

In this paper, we have made progress towards resolving this dilemma by considering a weaker, partial version of IA monotonicity. This “partial monotonicity” does not require strong assumptions about choice behavior. It is satisfied under the usual mathematical interpretation of “monotonicity” that each instrument encourages all individuals either towards or away from treatment. We have shown that it still preserves the interpretation of the 2SLS estimand as a positive weighted average of complier groups, except in rare cases. These rare cases can and should be checked for when reporting 2SLS estimates with multiple instruments.

We have also developed a framework for aggregating multiple instruments to conduct inference about specific target parameters. The framework generalizes the approach of Mogstad et al. (2018) to replace IA monotonicity with partial monotonicity. The key idea is that even under partial monotonicity, each instrument still carries identifying content about parameters that are invariant to the instrument. We show that this allows for information aggregation by enforcing this content to be logically consistent across different instruments. This logical consistency condition lets us accumulate the identifying content from multiple instruments; this ensures that the whole of the method is greater than the sum of its parts. The method provides a general blueprint for extracting and aggregating information about treatment effects from multiple controlled or natural experiments while still maintaining plausible conditions on choice behavior.

	(1)	(2)	(3)	(4)
	All papers	Papers that use more than one IV	Papers with more IVs than treatment variables	Papers with more IVs than treatment variables using 2SLS/GMM
American Economic Review	100% 44	39% 17	36% 16	34% 15
Quarterly Journal of Economics	100% 28	54% 15	46% 13	43% 12
Journal of Political Economy	100% 23	65% 15	48% 11	43% 10
Econometrica	100% 15	67% 10	60% 9	53% 8
Review of Economic Studies	100% 12	67% 8	67% 8	58% 7
All	100% 122	53% 65	47% 57	43% 52

Table A.1: *Results of our IV survey by journal.*

A Survey on the Use of Multiple Instruments[22]

We searched the Web of Science Database for articles published between January 2000 and October 2018 containing the words "instrument" or "instrumental variable" in the abstract, title, or topic words. We restricted our search to the following five journals: Journal of Political Economy, American Economic Review, Quarterly Journal of Economics, Review of Economic Studies, and Econometrica. There were 266 articles that matched our search criteria. We restrict our attention to the empirical studies that use at least one IV, which includes 122 papers. The other 144 papers either discussed IV methodology without having an application, or they used the word instrument to describe a policy or financial instrument.

In column (2) of Table A.1, we define a paper as using multiple instruments if at least one specification in the main body of the paper includes more than one IV. The number of IVs is determined by the number of moment conditions used to construct the estimator. If a paper uses multiple IVs, but no more than one instrument in any given specification, then it is not counted as having multiple instruments. If a paper

[22] This survey was carried out with the assistance of Christine Blandhol and John Bonney.

divides a continuous variable into mutually exclusive discrete or binary instruments, then it is counted as having multiple instruments. Furthermore, if $Z_{i,1}$ and $Z_{i,2}$ are IVs and $Z_{i,2}$ is a function of $Z_{i,1}$ (e.g., $Z_{i,2} = Z_{i,1}^2$), then we would still count this specification as having multiple instruments. The bottom row of column (2) reveals that more than half of the IV papers in our sample used more than one instrument.

In column (3) of Table A.1, we define a paper as having more IVs than treatment variables if at least one specification in the main body of the paper includes more IVs than endogenous variables. For example, let $D_{i,1}$ and $D_{i,2}$ denote the endogenous variables used in a specification in the main body of the paper, and $Z_{i,1}$, $Z_{i,2}$, $Z_{i,3}$ denote IVs. If $Z_{i,1}$, $Z_{i,2}$, and $Z_{i,3}$ are included in the instrument set, then the paper has more IVs than endogenous variables. However, if only $Z_{i,1}$ and $Z_{i,2}$ are included in the instrument set, then the paper does not have more IVs than endogenous variables. Comparing column (3) to column (2), we see that most papers that used more than one instrument had fewer treatment variables than IVs.

A few papers that used more IVs than endogenous variables did so in a way that was either nonstandard or unclear. In column (4) of Table A.1 we remove these papers and focus on only those that combined multiple instruments using 2SLS or GMM. This leaves 43% of papers across the five journals. This shows that combining more instruments than treatments through 2SLS or GMM is widespread empirical practice.

B Proofs

Proof of Proposition 1. ($\Rightarrow$) Suppose that nesting statement is not true. Then there exist $j, k \in \mathcal{I}$ such that $\mathcal{Z}_j \not\subseteq \mathcal{Z}_k$ and $\mathcal{Z}_k \not\subseteq \mathcal{Z}_j$. Since the empty set $\emptyset$ is a subset of every set (including itself), this implies that both $\mathcal{Z}_j$ and $\mathcal{Z}_k$ are not empty. Thus, there exists a $z_j \in \mathcal{Z}_j$ such that $z_j \notin \mathcal{Z}_k$, and there exists a $z_k \in \mathcal{Z}_k$ such that $z_k \notin \mathcal{Z}_j$. By the definition of these sets, this means that

$$\begin{aligned} & D_j(z_j) = 1 > 0 = D_j(z_k) \\ \text{and} \quad & D_k(z_j) = 0 < 1 = D_k(z_k). \end{aligned} \tag{23}$$

Thus, $D_i(z_j) \geq D_i(z_k)$ for some $i = j$, but $D_i(z_j) < D_i(z_k)$ for $i = k$. This violates Assumption IAM. By contraposition, it follows that Assumption IAM implies the nesting statement.

($\Leftarrow$) Conversely, if Assumption IAM is not true, then there exist $j, k \in \mathcal{I}$ and $z_j, z_k \in \mathcal{Z}$ such that (23) holds. By definition, (23) implies that $z_j \in \mathcal{Z}_j$, but $z_j \notin \mathcal{Z}_k$, and that $z_k \in \mathcal{Z}_k$, but $z_k \notin \mathcal{Z}_j$. That is, $\mathcal{Z}_k \not\subseteq \mathcal{Z}_j$, and $\mathcal{Z}_j \not\subseteq \mathcal{Z}_k$. It follows that the nesting statement also implies Assumption IAM. *Q.E.D.*

Proof of Proposition 2. Suppose to the contrary that there exist $j, k \in \mathcal{I}(z^\star)$ for which (2) does not hold. Then the matrix

$$\partial V_{jk}(z^\star) \equiv \begin{bmatrix} \partial_1 V_j(z^\star) & \partial_2 V_j(z^\star) \\ \partial_1 V_k(z^\star) & \partial_2 V_k(z^\star) \end{bmatrix} \equiv \begin{bmatrix} \partial V_j(z^\star) \\ \partial V_k(z^\star) \end{bmatrix}$$

is invertible. Thus, the span of $\partial V_{jk}(z^\star)$ is $\mathbb{R}^2$, so there exists a unit vector $v^\star \in \mathbb{R}^2$ such that $\partial V_j(z^\star)v^\star > 0$, while $\partial V_k(z^\star)v^\star < 0$. Taking a Taylor series expansion at $z^\star + \epsilon v^\star$ for sufficiently small $\epsilon > 0$, we have that

$$\begin{aligned} & V_j(z^\star + \epsilon v^\star) \approx V_j(z^\star) + \epsilon \left[\partial V_j(z^\star)\right] v^\star > 0, \\ \text{while} \quad & V_k(z^\star + \epsilon v^\star) \approx V_k(z^\star) + \epsilon \left[\partial V_k(z^\star)\right] v^\star < 0 \end{aligned}$$

since $V_j(z^\star) = V_k(z^\star) = 0$. On the other hand, an ϵ step in the direction $-v^\star$ yields

$$V_j(z^\star - \epsilon v^\star) < 0 \quad \text{while} \quad V_k(z^\star - \epsilon v^\star) > 0.$$

Using (1), we have that

$$\begin{aligned} & D_j(z^\star + \epsilon v^\star) = 1 > D_j(z^\star - \epsilon v^\star) = 0 \\ \text{and} \quad & D_k(z^\star + \epsilon v^\star) = 0 < D_k(z^\star - \epsilon v^\star) = 1, \end{aligned}$$

which shows that Assumption IAM is violated. This establishes the result through contraposition. *Q.E.D.*

Proof of Proposition 3. Take any $(z_\ell, z_{-\ell})$ and $(z'_\ell, z_{-\ell})$ in $\mathcal{Z}$. Since $z_\ell, z'_\ell \in \mathbb{R}$, either $z_\ell \geq z'_\ell$ or $z'_\ell \geq z_\ell$. Suppose that the first case holds. Then $(z_\ell, z_{-\ell}) \geq (z'_\ell, z_{-\ell})$, so $D_i(z_\ell, z_{-\ell}) \geq D_i(z'_\ell, z_{-\ell})$ for all i, as required by Assumption PM. *Q.E.D.*

Proof of Proposition 4. That an individual i cannot belong to more than one of the six groups can be verified by inspection. To see that i must belong to at least one of these groups, note that Assumption AM implies that i must satisfy either

$$D_i(0,0) \leq D_i(0,1) \leq D_i(1,0) \leq D_i(1,1) \tag{24}$$

$$\text{or} \quad D_i(0,0) \leq D_i(1,0) \leq D_i(0,1) \leq D_i(1,1). \tag{25}$$

If i satisfies (24) then by Table 1, their group is $G_i \in \{\text{at}, \text{nt}, 1\text{c}, \text{ec}, \text{rc}\}$; that is, something other than a Z_2 complier. If i satisfies (25), then their group is something other than a Z_1 complier. In either case, they must belong to one of the six groups listed in Table 1. *Q.E.D.*

Proof of Proposition 5. Label the instrument pairs as $z^1 = (0,0)$, $z^2 = (0,1)$, $z^3 = (1,0)$, and $z^4 = (1,1)$, and denote their associated probabilities as $q^k \equiv \mathbb{P}[Z_i = z^k]$ for $k = 1,2,3,4$. We will prove the result for the case with $\pi_{1\text{c}} \geq \pi_{2\text{c}}$, so that the propensity score $p(z^k) \equiv \mathbb{P}[D_i = 1|Z_i = z^k]$ is increasing in k, due to Assumption AM. A symmetric proof applies to the case with $\pi_{2\text{c}} \geq \pi_{1\text{c}}$.

Theorem 2 in Imbens and Angrist (1994) shows that the 2SLS estimand is given by a convex weighted average of three Wald (1940) estimands, which we write as

$$\beta_{2\text{sls}} = \lambda_2 w_2 + \lambda_3 w_3 + \lambda_4 w_4, \tag{26}$$

where the Wald estimands are given by

$$w_k \equiv \frac{\mathbb{E}[Y_i|Z_i = z^k] - \mathbb{E}[Y_i|Z_i = z^{k-1}]}{p(z^k) - p(z^{k-1})},$$

and the weights are defined by

$$\lambda_k \equiv \frac{\left(p(z^k) - p(z^{k-1})\right)\sum_{\ell=k}^4 q^\ell \left(p(z^\ell) - \mathbb{E}[p(Z_i)]\right)}{\sum_{j=2}^4 \left[\left(p(z^j) - p(z^{j-1})\right)\sum_{\ell=j}^4 q^\ell \left(p(z^\ell) - \mathbb{E}[p(Z_i)]\right)\right]}.$$

Theorem 1 of Imbens and Angrist (1994) shows that each Wald estimand, w_k, gives the average treatment effect for individuals who change treatment status in response to a change in the instrument from z^{k-1} to z^k. Using the group definitions in Proposition 4, this implies that w_2 represents the average treatment effect for both the Z_2 compliers and eager compliers. Similarly, w_4 reflects the average treatment effect for the Z_2 compliers and the reluctant compliers. So,

$$\begin{aligned} w_2 &= \left(\frac{\pi_{2\text{c}}}{\pi_{2\text{c}} + \pi_{\text{ec}}}\right)\Delta_{2\text{c}} + \left(\frac{\pi_{\text{ec}}}{\pi_{2\text{c}} + \pi_{\text{ec}}}\right)\Delta_{\text{ec}}, \\ \text{and} \quad w_4 &= \left(\frac{\pi_{2\text{c}}}{\pi_{2\text{c}} + \pi_{\text{rc}}}\right)\Delta_{2\text{c}} + \left(\frac{\pi_{\text{rc}}}{\pi_{2\text{c}} + \pi_{\text{rc}}}\right)\Delta_{\text{rc}}. \end{aligned}$$

However, w_3 is different, since a shift from $z^2 \equiv (0,1)$ to $z^3 \equiv (1,0)$ creates two-way flows. In particular, such a shift induces Z_1 compliers to take treatment, but Z_2 compliers to exit treatment. Using a minor modification of the argument in Imbens and Angrist (1994), it follows that

$$\begin{aligned}
w_3 &= \frac{\mathbb{E}\left[(Y_i(1)-Y_i(0))(D_i(1,0)-D_i(0,1)\right]}{p(z^3)-p(z^2)} \\
&= \left(\frac{\pi_{1\mathrm{c}}}{p(z^3)-p(z^2)}\right)\Delta_{1\mathrm{c}} - \left(\frac{\pi_{2\mathrm{c}}}{p(z^3)-p(z^2)}\right)\Delta_{2\mathrm{c}} \\
&= \left(\frac{\pi_{1\mathrm{c}}}{\pi_{1\mathrm{c}}-\pi_{2\mathrm{c}}}\right)\Delta_{1\mathrm{c}} - \left(\frac{\pi_{2\mathrm{c}}}{\pi_{1\mathrm{c}}-\pi_{2\mathrm{c}}}\right)\Delta_{2\mathrm{c}},
\end{aligned}$$

where the last equality used $p(z^3)-p(z^2) = (\pi_{1\mathrm{c}}+\pi_{\mathrm{ec}}) - (\pi_{2\mathrm{c}}+\pi_{\mathrm{ec}}) = \pi_{1\mathrm{c}}-\pi_{2\mathrm{c}}$.

Substituting the expressions for the Wald estimands into (26), we have

$$\beta_{2\mathrm{sls}} = \sum_{g\in\{1\mathrm{c},2\mathrm{c},\mathrm{ec},\mathrm{rc}\}} \tau_g \Delta_g,$$

where

$$\begin{aligned}
\tau_{\mathrm{ec}} &\equiv \frac{\lambda_2\pi_{\mathrm{ec}}}{\pi_{\mathrm{ec}}+\pi_{2\mathrm{c}}}, & \tau_{1\mathrm{c}} &\equiv \frac{\lambda_3\pi_{1\mathrm{c}}}{\pi_{1\mathrm{c}}-\pi_{2\mathrm{c}}}, \\
\tau_{\mathrm{rc}} &\equiv \frac{\lambda_4\pi_{\mathrm{rc}}}{\pi_{\mathrm{rc}}+\pi_{2\mathrm{c}}}, & \text{and} \quad \tau_{2\mathrm{c}} &\equiv \left(\frac{\lambda_2\pi_{2\mathrm{c}}}{\pi_{\mathrm{ec}}+\pi_{2\mathrm{c}}}\right) - \left(\frac{\lambda_3\pi_{2\mathrm{c}}}{\pi_{1\mathrm{c}}-\pi_{2\mathrm{c}}}\right) + \left(\frac{\lambda_4\pi_{2\mathrm{c}}}{\pi_{\mathrm{rc}}+\pi_{2\mathrm{c}}}\right).
\end{aligned}$$

It is straightforward to verify that $\tau_{\mathrm{ec}}, \tau_{1\mathrm{c}}$, and τ_{rc} are each non-negative, and that

$$\sum_{g\in\{1\mathrm{c},2\mathrm{c},\mathrm{ec},\mathrm{rc}\}} \tau_g = \lambda_2+\lambda_3+\lambda_4 = 1.$$

For $\tau_{2\mathrm{c}}$, note first that

$$\pi_{\mathrm{ec}}+\pi_{2\mathrm{c}} = p(z^2)-p(z^1),$$

and that, similarly,

$$\pi_{1\mathrm{c}}-\pi_{2\mathrm{c}} = p(z^3)-p(z^2) \quad \text{and} \quad \pi_{\mathrm{rc}}-\pi_{2\mathrm{c}} = p(z^4)-p(z^3).$$

Substituting this observation and the definition of λ_k into the expression for $\tau_{2\mathrm{c}}$ and simplifying, we have

$$\tau_{2\mathrm{c}} = \pi_{2\mathrm{c}} \times \frac{q^2\left(p(z^2)-\mathbb{E}[p(Z_i)]\right) + q^4\left(p(z^4)-\mathbb{E}[p(Z_i)]\right)}{\sum_{j=2}^{4}\left[\left(p(z^j)-p(z^{j-1})\right)\sum_{\ell=j}^{4} q^\ell\left(p(z^\ell)-\mathbb{E}[p(Z_i)]\right)\right]}.$$

The denominator of this expression is always positive and $\pi_{2\mathrm{c}}$ is always non-negative.

For the numerator, notice that since $Z_{i,2} = 1$ if and only if $Z_i \in \{z^2, z^4\}$,

$$\begin{aligned}
&q^2 \left(p(z^2) - \mathbb{E}[p(Z_i)]\right) + q^4 \left(p(z^4) - \mathbb{E}[p(Z_i)]\right) \\
&\qquad = \mathbb{E}\left[Z_{i,2}\left(p(Z_i) - \mathbb{E}[p(Z_i)]\right)\right] \\
&\qquad = \mathbb{E}\left[Z_{i,2}\left(D_i - \mathbb{E}[D_i]\right)\right] \equiv \mathrm{Cov}(D_i, Z_{i,2}),
\end{aligned}$$

where the second equality follows by iterating expectations. Thus, the sign of $\tau_{2\mathrm{c}}$ is the same as that of $\mathrm{Cov}(D_i, Z_{i,2})$, which is in turn the same as the sign of $\mathbb{E}[D_i|Z_{i,2} = 1] - \mathbb{E}[D_i|Z_{i,2} = 0]$, since

$$\mathrm{Cov}(D_i, Z_{i,2}) = \left(\mathbb{E}[D_i|Z_{i,2} = 1] - \mathbb{E}[D_i|Z_{i,2} = 0]\right)\mathbb{P}[Z_{i,2} = 1]\,\mathbb{P}[Z_{i,2} = 0].$$

Q.E.D.

Proof of Proposition 6. Since $Z_{i,1}$ and $Z_{i,2}$ are binary,

$$\begin{aligned}
&\mathbb{P}[D_i = 1|Z_{i,2} = 1] - \mathbb{P}[D_i = 1|Z_{i,2} = 0] \\
&\quad = \quad p(1,1)\,\mathbb{P}[Z_{i,1} = 1|Z_{i,2} = 1] + p(0,1)\,\mathbb{P}[Z_{i,1} = 0|Z_{i,2} = 1] \\
&\qquad - p(1,0)\,\mathbb{P}[Z_{i,1} = 1|Z_{i,2} = 0] - p(0,0)\,\mathbb{P}[Z_{i,1} = 0|Z_{i,2} = 0].
\end{aligned}$$

Assumptions AM and E imply that

$$p(1,1) \equiv \mathbb{P}[D_i = 1|Z_{i,1} = 1, Z_{i,2} = 1] = \mathbb{P}[D_i(1,1) = 1] \geq \mathbb{P}[D_i(0,1) = 1] = p(0,1),$$

and similarly that $p(1,0) \geq p(0,0)$ and $p(0,1) \geq p(0,0)$. If $\mathrm{Cov}(Z_{i,1}, Z_{i,2}) \geq 0$, then also $\mathbb{P}[Z_{i,1} = 1|Z_{i,2} = 1] \geq \mathbb{P}[Z_{i,1} = 1] \geq \mathbb{P}[Z_{i,1} = 1|Z_{i,2} = 0]$. Thus,

$$\begin{aligned}
&\Pr[D_i = 1|Z_{i,2} = 1] - \Pr[D_i = 1|Z_{i,2} = 0] \\
&\quad \geq \quad p(1,1)\,\mathbb{P}[Z_{i,1} = 1] + p(0,1)\left(1 - \mathbb{P}[Z_{i,1} = 1]\right) \\
&\qquad - p(1,0)\,\mathbb{P}[Z_{i,1} = 1] - p(0,0)(1 - \mathbb{P}[Z_{i,1} = 1]) \\
&\quad = [p(1,1) - p(1,0)]\,\mathbb{P}[Z_{i,1} = 1] + [p(0,1) - p(0,0)]\left(1 - \Pr[Z_{i,2} = 1]\right) \geq 0.
\end{aligned}$$

By Proposition 5, this implies that $\tau_{2\mathrm{c}} \geq 0$. A symmetric argument shows that $\mathbb{P}[D_i = 1|Z_{i,1} = 1] - \mathbb{P}[D_i = 1|Z_{i,1} = 0] \geq 0$, so that $\tau_{1\mathrm{c}} \geq 0$ as well. *Q.E.D.*

Proof of Proposition 7. By Theorem 2 in Imbens and Angrist (1994),

$$\beta_{2\mathrm{sls}} = \sum_{k=2}^{K} \lambda_k w_k, \tag{27}$$

where the Wald estimands are

$$w_k \equiv \frac{\mathbb{E}[Y_i|Z_i = z^k] - \mathbb{E}[Y_i|Z_i = z^{k-1}]}{p(z^k) - p(z^{k-1})},$$

and the weights are

$$\lambda_k \equiv \frac{\left(p(z^k) - p(z^{k-1})\right) \sum_{\ell=k}^{K} q^\ell \left(p(z^\ell) - \mathbb{E}[p(Z_i)]\right)}{\sum_{j=2}^{K} \left[\left(p(z^j) - p(z^{j-1})\right) \sum_{\ell=j}^{K} q^\ell \left(p(z^\ell) - \mathbb{E}[p(Z_i)]\right)\right]}. \tag{28}$$

By Assumption E,

$$\begin{aligned} w_k &= \frac{\mathbb{E}[Y_i(D_i(z^k)) - Y_i(D_i(z^{k-1}))]}{p(z^k) - p(z^{k-1})} \\ &= \frac{\sum_{g \in \mathcal{G}} \mathbb{E}[Y_i(D_i(z^k)) - Y_i(D_i(z^{k-1}))|G_i = g]\pi_g}{p(z^k) - p(z^{k-1})} \\ &= \frac{\sum_{g:k \in \mathcal{C}_g} \Delta_g \pi_g - \sum_{g:k \in \mathcal{D}_k} \Delta_g \pi_g}{p(z^k) - p(z^{k-1})}, \end{aligned} \tag{29}$$

since $Y_i(D_i(z^k)) - Y_i(D_i(z^{k-1})) = 0$ except when $k \in \mathcal{C}_{G_i}$ or $k \in \mathcal{D}_{G_i}$. Substituting (29) into (27),

$$\begin{aligned} \beta_{\text{2sls}} &= \sum_{k=2}^{K} \lambda_k \left(\frac{\sum_{g:k \in \mathcal{C}_g} \Delta_g \pi_g - \sum_{g:k \in \mathcal{D}_g} \Delta_g \pi_g}{p(z^k) - p(z^{k-1})} \right) \\ &= \sum_{k=2}^{K} \lambda_k \left(\frac{\sum_{g \in \mathcal{G}} \left(\mathbb{1}[k \in \mathcal{C}_g] - \mathbb{1}[k \in \mathcal{D}_g]\right) \Delta_g \pi_g}{p(z^k) - p(z^{k-1})} \right) \\ &= \sum_{g \in \mathcal{G}} \left(\pi_g \sum_{k=2}^{K} \left(\mathbb{1}[k \in \mathcal{C}_g] - \mathbb{1}[k \in \mathcal{D}_g]\right) \left(\frac{\lambda_k}{p(z^k) - p(z^{k-1})} \right) \right) \Delta_g \equiv \sum_{g \in \mathcal{G}} \tau_g \Delta_g. \end{aligned} \tag{30}$$

Substituting the definition of λ_k from (28) and simplifying,

$$\begin{aligned} \tau_g &= \pi_g \sum_{k=2}^{K} \left(\mathbb{1}[k \in \mathcal{C}_g] - \mathbb{1}[k \in \mathcal{D}_g]\right) \left(\frac{\sum_{\ell=k}^{K} q^\ell \left(p(z^\ell) - \mathbb{E}[p(Z_i)]\right)}{\sum_{j=2}^{K} \left[\left(p(z^j) - p(z^{j-1})\right) \sum_{\ell=j}^{K} q^\ell \left(p(z^\ell) - \mathbb{E}[p(Z_i)]\right)\right]} \right) \\ &= \pi_g \sum_{k=2}^{K} \left(\mathbb{1}[k \in \mathcal{C}_g] - \mathbb{1}[k \in \mathcal{D}_g]\right) \left(\frac{\text{Cov}\left(D_i, \mathbb{1}[p(Z_i) \geq p(z^k)]\right)}{\text{Var}(p(Z_i))} \right), \end{aligned} \tag{31}$$

where in the numerator we used

$$\begin{aligned}
\operatorname{Cov}\left(D_i, \mathbb{1}[p(Z_i) \geq p(z^k)]\right) &= \mathbb{E}\left[\mathbb{1}[p(Z_i) \geq p(z^k)]\,(D_i - \mathbb{E}[D_i])\right] \\
&= \mathbb{E}\left[\mathbb{1}[p(Z_i) \geq p(z^k)]\,(p(Z_i) - \mathbb{E}[p(Z_i)])\right] \\
&= \sum_{\ell=k}^{K} q^\ell \left(p(z^\ell) - \mathbb{E}[p(Z_i)]\right),
\end{aligned}$$

and in the denominator we used

$$\begin{aligned}
\operatorname{Var}(p(Z_i)) &= \mathbb{E}\left[D_i\,(p(Z_i) - \mathbb{E}[p(Z_i)])\right] \\
&= \sum_{\ell=1}^{K} p(z^\ell) \left(p(z^\ell) - \mathbb{E}[p(Z_i)]\right) q^\ell \\
&= \sum_{\ell=1}^{K} \left(\sum_{j=2}^{K} \mathbb{1}[j \leq \ell]\left(p(z^j) - p(z^{j-1})\right)\right) \left(p(z^\ell) - \mathbb{E}[p(Z_i)]\right) q^\ell \\
&= \sum_{j=2}^{K} \left[\left(p(z^j) - p(z^{j-1})\right) \sum_{\ell=j}^{K} \left(p(z^\ell) - \mathbb{E}[p(Z_i)]\right) q^\ell\right],
\end{aligned}$$

where the third equality follows from a telescoping sum identity,[23] together with the fact that $\sum_{\ell=1}^{K}\left(p(z^\ell) - \mathbb{E}[p(Z_i)]\right) q^\ell = 0$. Examining the expression for the weights in (31), we have that

$$\operatorname{sgn}(\tau_g) = \mathbb{1}[\pi_g > 0] \times \operatorname{sgn}\left(\sum_{k=2}^{K} \left(\mathbb{1}[k \in \mathcal{C}_g] - \mathbb{1}[k \in \mathcal{D}_g]\right) \operatorname{Cov}\left(D_i, \mathbb{1}[p(Z_i) \geq p(z^k)]\right)\right).$$

It remains to show that $\tau_g = 0$ when $\mathcal{C}_g = \emptyset$, so that only groups that comply with at least one instrument contrast receive weight in the 2SLS estimand. To see that this is so, suppose to the contrary that there is a group g with $\pi_g > 0$ for which $\mathcal{C}_g = \emptyset$, while $\tau_g \neq 0$. Given the structure of τ_g, such a group must have $\mathcal{D}_g \neq \emptyset$. That is, this group must defy at some instrument contrast, even though they do not comply at any other instrument contrasts. We will prove that such a "pure defier" group cannot exist under Assumption PM by establishing a contradiction.

Let $j_0 \in \mathcal{D}_g$ be the instrument contrast at which the "pure defier" group g defies. By definition, $D_i(z^{j_0}) = 0$, while $D_i(z^{j_0-1}) = 1$. Since $\mathcal{C}_g = \emptyset$, it follows that for any i

[23] In particular, that $a^\ell = a^1 + \sum_{j=2}^{K} \mathbb{1}[j \leq \ell](a^j - a^{j-1})$ for any scalars $\{a^\ell\}_{\ell=1}^{K}$.

with $G_i = g$,

$$D_i(z^j) = \mathbb{1}[j < j_0]. \tag{32}$$

In particular, $D_i(z^1) = 1$, while $D_i(z^K) = 0$.

To proceed, it will be helpful to use the following terminology. We call two vectors z^j and z^k *pm-comparable* if they differ in only one component. That is, z^j and z^k are pm-comparable if there exists an $\ell' \in \{1, \ldots, L\}$ such that $z^j_\ell = z^k_\ell$ for all $\ell \neq \ell'$. If and z^j and z^k are pm-comparable, then Assumption PM requires that either $D_i(z^j) \leq D_i(z^k)$ for all $i \in \mathcal{I}$ or that $D_i(z^j) \geq D_i(z^k)$ for all $i \in \mathcal{I}$. Moreover, as we show in Lemma 1, if $j \leq k$, then there cannot exist a group $g^\star$ with $\pi_{g^\star} > 0$ for which individuals i in group $g^\star$ have $D_i(z^j) > D_i(z^k)$. We will now use this result to show that the existence of the "pure defier" group g defined above creates a contradiction.

Let z^{j_1} be the vector whose first component is the same as that of the largest propensity-score instrument value, z^K, while all other components are the same as the smallest, z^1. That is,

$$z^{j_1} \equiv (z^K_1, z^1_{-1}).$$

Then z^{j_1} and z^1 are pm-comparable, and $z^{j_1} \in \operatorname{supp}(Z_i)$, which we have assumed is rectangular. Since $D_i(z^1) = 1$ for any i with $G_i = g$ and $p(z^1)$ is the smallest propensity score value, it follows from Lemma 1 that $D_i(z^{j_1}) = 1$ for these individuals as well. Thus by (32), it must be that $j_1 < j_0$.

Now let z^{j_2} be the same as z^{j_1} except with its second component replaced by z^K_2. That is,

$$z^{j_2} \equiv (z^K_2, z^{j_1}_{-2}) \equiv (z^K_1, z^K_2, z^1_3, \ldots, z^1_L).$$

Then z^{j_2} is pm-comparable to z^{j_1}, and $z^{j_2} \in \operatorname{supp}(Z_i)$. If it were the case that $j_2 \geq j_0$, then $p(z^{j_2}) \geq p(z^{j_0}) \geq p(z^{j_1})$, so that Lemma 1 would imply that $D_i(z^{j_2}) = 1$ for individuals with $G_i = g$. At the same time, (32) would imply that $D_i(z^{j_2}) = 0$ for these individuals, yielding a contradiction. Thus, it must be that $j_2 < j_0$.

Continuing in this way, we find a sequence of vectors $z^{j_1}, z^{j_2}, z^{j_3}, \ldots, z^{j_L}$ that each differ from z^K in one component less than its predecessor, and such that $j_\ell < j_0$ for each ℓ. This process ends once we reach j_L, at which point $z^{j_L} = z^K$ is the instrument value corresponding to the largest propensity score value. However, this implies a contradiction, because $j_L < j_0$ while at the same time $j_L = K \geq j_0$. We conclude that a "pure defier" group cannot exist under Assumption PM, and therefore that the sum in (30) only needs to be indexed over $g \in \mathcal{G}$ for which $\mathcal{C}_g \neq \emptyset$. *Q.E.D.*

Lemma 1. *Suppose that Assumption E holds. Let $\mathcal{G}$ denote the set of all realizations of $\{D_i(z)\}_{z \in \mathcal{Z}}$ that are consistent with Assumption PM. Suppose that z and z' are pm-comparable and that $p(z) \leq p(z')$. Then there does not exist a group $g^\star \in \mathcal{G}$ such that $\pi_{g^\star} > 0$ and $D_i(z) > D_i(z')$ for all i with $G_i = g^\star$.*

Proof of Lemma 1. Since z and z' are pm-comparable, Assumption PM requires that $D_i(z) \leq D_i(z')$ for all $i \in \mathcal{I}$, or $D_i(z) \geq D_i(z')$ for all $i \in \mathcal{I}$. If such a group $g^\star$ did exist, then the latter case would need to hold. However, this would imply that

$$\begin{aligned}
p(z) &\equiv \mathbb{P}[D_i = 1 | Z_i = z] \\
&= \mathbb{P}[D_i(z) = 1] \\
&= \mathbb{P}[D_i(z) = 1 | G_i = g^\star]\pi_{g^\star} + \mathbb{P}[D_i(z) = 1 | G_i \neq g^\star](1 - \pi_{g^\star}) \\
&> \mathbb{P}[D_i(z') = 1 | G_i = g^\star]\pi_{g^\star} + \mathbb{P}[D_i(z') = 1 | G_i \neq g^\star](1 - \pi_{g^\star}) \\
&= \mathbb{P}[D_i = 1 | Z_i = z'] \equiv p(z'),
\end{aligned}$$

which contradicts the assumption that $p(z) \leq p(z')$. *Q.E.D.*

References

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455. 10

Angrist, J. D. and J.-S. Pischke (2009): *Mostly Harmless Econometrics*, Princeton University Press. 1

Bitler, M., H. Hoynes, and T. Domina (2014): "Experimental Evidence on Distributional Effects of Head Start," Tech. rep. 1

Bitler, M. P., J. B. Gelbach, and H. W. Hoynes (2006): "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments," *The American Economic Review*, 96, 988–1012. 1

Brinch, C. N., M. Mogstad, and M. Wiswall (2012): "Beyond LATE with a Discrete Instrument," *Working paper.* 25

——— (2017): "Beyond LATE with a Discrete Instrument," *Journal of Political Economy*, 125, 985–1039. 1, 25

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. 20

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011): "Estimating Marginal Returns to Education," *American Economic Review*, 101, 2754–81. 1

Carneiro, P. and S. Lee (2009): "Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality," *Journal of Econometrics*, 149, 191–208. 1

Carneiro, P., M. Lokshin, and N. Umapathi (2016): "Average and Marginal Returns to Upper Secondary Schooling in Indonesia," *Journal of Applied Econometrics*, 32, 16–36. 1

Chernozhukov, V. and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261. 2

Chesher, A. and A. Rosen (2012): "Simultaneous Equations Models for Discrete Outcomes: Coherence, Completeness and Identification," *cemmap working paper 21/12*. 22

Cornelissen, T., C. Dustmann, A. Raute, and U. Schönberg (forthcoming): "Who benefits from universal childcare? Estimating marginal returns to early child-care attendance," *Journal of Political Economy*. 1

Doyle Jr., J. J. (2007): "Child Protection and Child Outcomes: Measuring the Effects of Foster Care," *The American Economic Review*, 97, 1583–1610. 1

Felfe, C. and R. Lalive (2014): "Does Early Child Care Help or Hurt Children's Development?" Tech. Rep. 8484. 1

Firpo, S., N. M. Fortin, and T. Lemieux (2009): "Unconditional Quantile Regressions," *Econometrica*, 77, 953–973. 1

French, E. and J. Song (2014): "The Effect of Disability Insurance Receipt on Labor Supply," *American Economic Journal: Economic Policy*, 6, 291–337. 1

Havnes, T. and M. Mogstad (2015): "Is universal child care leveling the playing field?" *Journal of Public Economics*, 127, 100–114. 1

Heckman, J. J. (1978): "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica*, 46, 931–959. 17, 22

——— (2001): "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *The Journal of Political Economy*, 109, 673–748. 1

Heckman, J. J. and R. Pinto (2018): "Unordered Monotonicity," *Econometrica*, 86, 1–35. 17

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, 5, 7

——— (2008): "Instrumental Variables in Models with Multiple Outcomes: the General Unordered Case," *Annales d'Economie et de Statistique*, 91/92, 151–174. 17

Heckman, J. J. and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, 669–738. 2, 4, 18, 20

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, 18, 20

——— (2001): "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. 2, 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. 2, 18

——— (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, 18, 20

HULL, P. (2016): "Estimating Hospital Quality with Quasi-Experimental Data," *Working paper*. 1

IMBENS, G. W. AND J. D. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475. 1, 3, 9, 10, 11, 14, 34, 36

KAMAT, V. (2018): "Identification with Latent Choice Sets: The Case of the Head Start Impact Study," *Working paper*. 17

KIRKEBOEN, L. J., E. LEUVEN, AND M. MOGSTAD (2016): "Field of Study, Earnings, and Self-Selection," *The Quarterly Journal of Economics*, 131, 1057–1111. 1, 17

KLINE, P. AND C. R. WALTERS (2016): "Evaluating Public Programs with Close Substitutes: The Case of Head Start*," *The Quarterly Journal of Economics*, 131, 1795–1848. 1, 17

LEE, S. AND B. SALANIÉ (2018): "Identifying Effects of Multivalued Treatments," *Econometrica*, Forthcoming. 17

Lewbel, A. (2007): "Coherency and Completeness of Structural Models Containing a Dummy Endogenous Variable," *International Economic Review*, 48, 1379–1392. 22

Maddala, G. S. (1983): *Limited-dependent and qualitative variables in econometrics*, 3, Cambridge university press. 22

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. 1

Manski, C. (1994): "The selection problem," in *Advances in Econometrics, Sixth World Congress*, vol. 1, 143–70. 2

Manski, C. F. (1990): "Nonparametric Bounds on Treatment Effects," *The American Economic Review*, 80, 319–323. 2

Manski, C. F. and J. V. Pepper (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68, 997–1010. 2

——— (2009): "More on monotone instrumental variables," *Econometrics Journal*, 12, S200–S216. 2

Milgrom, P. and C. Shannon (1994): "Monotone comparative statics," *Econometrica*, 157–180. 8

Moffitt, r. (2008): "Estimating Marginal Treatment Effects in Heterogeneous Populations," *Annales d'Economie et de Statistique*, 239–261. 1

Mogstad, M., A. Santos, and A. Torgovitsky (2018): "Using Instrumental Variables for Inference about Policy Relevant Treatment Parameters," *Econometrica (forthcoming)*. 3, 20, 30

Mogstad, M. and A. Torgovitsky (2018): "Identification and Extrapolation of Causal Effects with Instrumental Variables," *Annual Review of Economics*, 10. 3, 20

Mountjoy, J. (2018): "Community Colleges and Upward Mobility," *Working paper*. 8, 17

Nybom, M. (2017): "The Distribution of Lifetime Earnings Returns to College," *Journal of Labor Economics*, 000–000. 1

Poirier, D. J. (1980): "Partial observability in bivariate probit models," *Journal of Econometrics*, 12, 209–217. 17

Tamer, E. (2003): "Incomplete Simultaneous Discrete Response Model with Multiple Equilibria," *The Review of Economic Studies*, 70, 147–165. 22

Vytlacil, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331–341. 4, 16

Wald, A. (1940): "The Fitting of Straight Lines if Both Variables are Subject to Error," *The Annals of Mathematical Statistics*, 11, 284–300. 34

Walters, C. (2014): "The Demand for Effective Charter Schools," Tech. rep. 1