Supplemental Material for "A Vector Monotonicity Assumption for Multiple Instruments"

Leonard Goff

Last updated: September 1, 2020

Contents

A Results about linear 2SLS

The standard method of combining multiple instruments in applied work is to use some variant of the two-stage least squares (2SLS) estimator. In the language of Proposition 5, this corresponds to either letting $h(z)$ be a linear projection of a treatment indicator on the instruments (what I'll call linear 2SLS), and letting $h(z)$ equal the propensity score (what we call "regression on the propensity score, also referred to as fully-saturated" 2SLS).

Imbens and Angrist (1994) show that under conventional IAM monotonicity, regressing Y on the propensity score function recovers a convex combination of group-specific average treatment effects. In Section B.5, I provide a novel demonstration of this, starting from Lemma 5. However, this result does not extend to vector monotonicity, a property that arises from the fat that VM allows two-way flows between some pairs of points in $\mathcal{Z}$.

In this section I demonstrate two special cases of vector monotonicity in which *linear* 2SLS with binary instruments recovers a convex combination of causal effects, focusing on a case with binary instruments and $\mathcal{Z} = \{0,1\} \times \{0,1\} \cdots \times \{0,1\}$. The first special case is one in which each unit is responsive to at most *one* of the instruments:

Assumption 4 (separable compliance). *For each unit i (i.e. with probability one), there exists a $j \in \{1 \dots J\}$ such that $D_i(z, z_{-j}) = D_i(z, z'_{-j})$ for all $z_{-j}, z'_{-j} \in \mathcal{Z}_{-j}$ and $z \in \{0,1\}$, i.e. treatment assignment only depends on the value of Z_j.*

In the language of Section 3, this corresponds to a case where all units are Z_j "compliers" for some j (equivalently, all compliance groups correspond to Sperner families $\{j\}$ for some j).

In the following theorem, we will also use a slight strengthening of the notion of vector monotonicity:

Assumption 2* (vector monotonicity with aligned covariances). *With the Z_j normalized such that $Cov(D_i, Z_{ij}) \geq 0$, for each $j \in \{1 \dots J\}$ and $z_{-j} \in \{0,1\}^{J-1}$, we have that $D_i(1, z_{-j}) \geq D_i(0, z_{-j})$*

Assumption 2* is stronger than Assumption 2, because when some of the instruments are negatively correlated it is possible that the unconditional covariances are negative, even with $Z_j i = 1$ corresponding to the "pro-treatment" state for instrument j.[1]

Let $\rho_{2sls,lin}$ be the linear 2SLS estimand $\rho_{2sls,lin} := Cov\left(Y_i, \sum_j \pi_j Z_{ji}\right) / Var\left(\sum_j \pi_j Z_{ji}\right)$, where π_j is the population regression coefficient on Z_j from a linear regression of D on the $Z_1 \dots Z_J$. Our result is then:

Theorem SM1. *Under Assumptions 1, 2*, 4:*

$$\rho_{2sls,lin} = \sum_{j=1}^{J} w_j \cdot E[Y_i(1) - Y_i(0) | D_i(1, Z_{-ji}) > D_i^j(0, Z_{-ji})]$$

where the weights w_j are positive and sum to one: $w_j = \frac{P(D_i^j(1) > D_i^j(0)) Cov(D_i, Z_{ji})}{\sum_{j=1}^{J} P(D_i^j(1) > D_i^j(0)) Cov(D_i, Z_{ji})}$.

[1] Note that under the construction in Section 3.3 from discrete to binary instruments, the resulting vector of binary instruments will satisfy Assumption 2* so long as if the CEF $E[D|Z_1 = z_m]$ is monotonic in m.

Proof. See Appendix E. □

Discussion: Linear 2SLS always identifies a sum of single instrument IV estimators $\rho_j := \frac{Cov(Y_i, Z_{ji})}{Cov(D_i, Z_{ji})}$ with weights that add to one (but may be negative), since we can write

$$\rho_{2sls,lin} = \frac{\sum_j \pi_j Cov(Y_i, Z_{ji})}{\sum_j \pi_j Cov(D_i Z_{ji})} = \sum_j \left\{ \frac{\pi_j Cov(D_i, Z_{ji})}{\sum_j \pi_j Cov(D_i Z_{ji})} \right\} \cdot \frac{Cov(Y_i, Z_{ji})}{Cov(D_i, Z_{ji})}$$

where we've used that $D_i - \sum_j \pi_j Z_{ji}$ is uncorrelated with each Z_{ji}.

Separable monotonicity allows linear 2SLS to identify a convex combination of LATEs despite the fact that even under separable compliance each ρ_j need not put positive weight on all compliance groups when the instruments are correlated. Nevertheless the 2SLS weights are such that the overall weight for each compliance group ends up being positive, despite the fact that each ρ_j is a linear combination of SLATE's (defined in Section 4.2) that each place negative weight on some groups.

Theorem SM1 also extends to the estimator defined by regression on the propensity score, because under VM and separable compliance it turns out that the propensity score function must be linear, and hence consistently estimated even with a linear first stage:

Corollary to Theorem SM1. *Under Assumptions 1, 2*, and 4:*

$$\frac{Cov(Y_i, P(Z_i))}{Var(P(Z_i))} = \rho_{2sls,lin}$$

where $P(Z_i) := E[D_i|Z_i]$.

Proof. By Assumption 1:

$$E[D_i|Z_i] = \sum_g P(G_i = g)\mathcal{D}_g(Z_i) = p_{a.t.} + \sum_{j=1}^{J} p_{Z_j} Z_{ji}$$

Since the propensity score is linear in the Z_{ki}, it coincides with the linear projection function used by 2SLS. Now Apply Theorem SM1. □

A second special case in which linear 2SLS can be justified in a context with VM is when the instruments are independent of one another, or slightly more generally, are what I call "unentangled" in selection:

Assumption 5 (instruments *unentangled* in selection). *For* $j \in \{1 \dots J\}$*:*

$$(D_i(0, Z_{-j,i}), D_i(1, Z_{-j,i})) \perp Z_{ji}$$

With these assumptions:

Lemma 1. *Under Assumptions 1, 2 and 5:*

$$\rho_j = E[Y_i(1) - Y_i(0)|D_i^j(1) > D_i^j(0)]$$

Proof. See Appendix E (note that the proof makes use of Assumption 2*, but this is implied by Assumption 2 when Assumption 5 holds). □

Now we can state our result:

Theorem SM2 (2SLS with unentangled binary instruments). *Under Assumptions 1, 2*, and 5, the linear two stage least squares estimand is*

$$\rho_{2sls,lin} = \sum_{j=1}^{J} w_j E[Y_i(1) - Y_i(0)|D_i(1, Z_{-j,i}) > D_i(0, Z_{-j,i})]$$

where the coefficients w_j are positive and sum to unity.

Proof. See Appendix E. □

The assumption that the instruments are unentangled is very strong. While it is weaker than full independence of the instruments, it is hard to articulate concretely a case in which it holds without independence of the instrments.

B Various examples and special cases from the main text

B.1 PM without VM, with two binary instruments

Suppose there are two binary instruments, and that PM holds but not VM. For VM to be violated, there must be a "flip" in which value of one of the instruments –say Z_2– is the "pro-treatment" state, depending on the value of the other instrument. In other words, for some choice of which instrument is called Z_2, and some choice of labeling for the "0" and "1" values of each instrument, we have that:

$$P(D_i(0,1) \geq D_i(0,0)) = 1 \text{ and } P(D_i(1,1) \leq D_i(1,0)) = 1$$

with both

$$P(D_i(0,1) > D_i(0,0) > 0 \text{ and } P(D_i(1,1) < D_i(1,0)) > 0$$

This is without loss of generality, given the choice to arbitrarily assign the labels $0, 1$.

Now consider the set of possible compliance groups that satisfy PM but not VM, denoted as $\mathcal{G}^{PM-VM}$. Any compliance group $g \in \mathcal{G}^{PM-VM}$ must then be either a complier, always-taker, or never-taker with respect to Z_2, when $Z_1 = 0$. Similarly, any compliance group g must then be either a "defier", always-taker, or never-taker with respect to Z_2 when $Z_1 = 1$. The set of pairs (g_0, g_1), where $g_0 \in \{c, a, n\}$ and $g_1 \in \{d, a, n\}$ exhausts the possible compliance groups, since knowing g_0 and g_1 pins down $D_i(z_1, z_2)$ for all four values of z_1, z_2. This generates an exhaustive set of 9 possible compliance groups, shown in Table 1.

However, all nine of the compliance groups cannot coexist at the same time. For example, if both odd compliers and Z_1 defiers both exist in the population, there will be two-way flows when varying Z_1 with Z_2 fixed at zero. This is a consequence of what

name	$\mathbf{Z_1 = 0, Z_2 = 0}$	$\mathbf{Z_1 = 0, Z_2 = 1}$	$\mathbf{Z_1 = 1, Z_2 = 0}$	$\mathbf{Z_1 = 1, Z_2 = 1}$
odd compliers	N	T	T	N
eager compliers	N	T	T	T
1-only	N	T	N	N
reluctant defiers	T	T	T	N
always takers	T	T	T	T
Z_1 defiers	T	T	N	N
2-only	N	N	T	N
Z_1 compliers	N	N	T	T
never takers	N	N	N	N

Table 1: Rows are possible compliance groups in the set $\mathcal{G}^{PM-VM}$. T and N indicate treatment, or non-treatment, respectively. Not all of these groups can coexist in the population without violating PM.

Mogstad, Torgovitsky and Walters (2020) call *logical consistency*, applied to selection. The possibilities separate into two cases, depending on whether there are "odd compliers" in the population. If there are odd compliers, then there can be no Z_1 compliers or Z_1 defiers in the population. This leaves seven groups, depicted in Table 2.

If, on the other hand, $P(G_i = \text{odd complier}) = 0$, then there can be either Z_1 compliers, or Z_1 defiers, but not both. This creates a second type of case. Supposing that $P(G_i = Z_1 \text{ complier}) > 0$, there can be no Z_1 defiers, 1-only units, or reluctant defiers. This leaves five possible groups, depicted in Table 3.

group name	$\mathbf{Z_1 = 0, Z_2 = 0}$	$\mathbf{Z_1 = 0, Z_2 = 1}$	$\mathbf{Z_1 = 1, Z_2 = 0}$	$\mathbf{Z_1 = 1, Z_2 = 1}$
odd compliers	N	T	T	N
eager compliers	N	T	T	T
reluctant defiers	T	T	T	N
1-only	N	T	N	N
2-only	N	N	T	N
always takers	T	T	T	T
never takers	N	N	N	N

Table 2: Case 1, when $P(G_i = \text{odd complier}) > 0$.

group name	$\mathbf{Z_1 = 0, Z_2 = 0}$	$\mathbf{Z_1 = 0, Z_2 = 1}$	$\mathbf{Z_1 = 1, Z_2 = 0}$	$\mathbf{Z_1 = 1, Z_2 = 1}$
eager compliers	N	T	T	T
Z_1 compliers	N	N	T	T
2-only	N	N	T	N
always takers	T	T	T	T
never takers	N	N	N	N

Table 3: Case 2, when $P(G_i = \text{odd complier}) = 0$ and $P(G_i = Z_1 \text{ complier}) > 0$.

The remaining $P(G_i = Z_1 \text{ defier}) > 0$ case is symmetric with respect to Table 3, up to a relabeling of "0" and "1" for Z_1: in addition to Z_1 defiers, there can be reluctant defiers, 1-only units, always takers and never takers.

Case 1 and Case 2 have very different implications for identification. In Case 2, the group-specific average treatment effects Δ_g are identified for all groups aside from always takers and never takers. However, it can be readily verified that Assumption IAM holds in Case 2.

However, in Case 1, the Δ_g for $g \notin \{a.t., n.t.\}$ are not identified. With three linearly independent Wald ratios, we only have three equations for five unknowns. This is also generally true under VM, that the Δ_g are not separately identified. However, here we can also show that the Wald estimands do not even identify the all compliers LATE (ACL). This is shown explicitly in the proof of Proposition 9, but can be seen in a special case by assuming that the 1-only and 2-only groups are not present. Even then we cannot identify the ACL. With this restriction, we would still have $E[Y_i|Z_i = (0,1)] - E[Y_i|Z_i = (0,0)] = p_{odd}\Delta_{odd} + p_{eager}\Delta_{eager}$, $E[Y_i|Z_i = (1,0)] - E[Y_i|Z_i = (1,1)] = p_{odd}\Delta_{odd} + p_{reluct.}\Delta_{reluct.}$, and $E[Y_i|Z_i = (1,0)] - E[Y_i|Z_i = (0,0)] = E[Y_i|Z_i = (0,1)] - E[Y_i|Z_i = (0,0)]$. Thus the third equation gives no further information beyond the first. The observable conditional means of Y_i are compatible with any numerical value of the ACL, which is equal to $p_{odd}\Delta_{odd} + p_{eager}\Delta_{eager} + p_{reluct.}\Delta_{reluct.}$.

B.2 The identifying power of Wald ratios with two binary instruments

For points $z, w \in \mathcal{Z}$ that are ordered component-wise, the standard LATE argument goes through to identify a local average treatment effect. For example, with two binary instruments, there are five unique pairs (z, w) that are ordered in a vector sense. Table 4 describes these in terms of the compliance groups introduced in Section 3.1. Correspond-

z	**w**	$\mathbf{D_i(z) > D_i(w) \iff G_i = \dots}$
(1,0)	(0,0)	Z_1 complier or eager complier
(0,1)	(0,0)	Z_2 complier or eager complier
(1,1)	(0,1)	Z_1 complier or reluctant complier
(1,1)	(1,0)	Z_2 complier or reluctant complier
(1,1)	(0,0)	any $g \in \mathcal{G}^c$

Table 4: Third column indicates which compliance groups G_i lead to $D_i(z) > D_i(w)$ with the indicated z,w. For each row in this Table, a Wald ratio ρ_{zw} identifies a LATE under Assumption VM. In the two binary instrument vase under VM: $\mathcal{G}^c = \{Z_1, Z_2, or, and\}$.

ing to each row in 4 is a LATE that can be identified via a Wald ratio ρ_{zw}, in a case of two binary instruments under VM. For instance, from the first row, $\rho_{(1,0),(0,0)}$ is:

$$E[Y_i(1) - Y_i(0)|G_i \in \{Z_1 \text{ comp.}, \text{reluctant}\}] = \frac{p_{Z_1}}{p_{Z_1} + p_{reluctant}}\Delta_{Z_1} + \frac{p_{reluctant}}{p_{Z_1} + p_{reluctant}}\Delta_{reluctant}$$

Recall that the number of compliance groups under VM scales with the so-called Dedekind numbers Ded_J, which grow much faster than $(2^J \times (2^J - 1))/2$, the number of unique Wald estimands. Furthermore, the ρ_{zw} are not even all linearly-independent: in the example of two binary instruments only three of the five are.

Thus, we can see that it will in general be hopeless to identify Δ_g for each compliance group $g \in \mathcal{G}^c$ separately, because we will lack an order condition on the number of pairs (z, w) in which there is variation in treatment take-up.[2] The exception is the familiar case of $J = 1$, where $(2^J \times (2^J - 1))/2 = \text{Ded}_J - 2 = 1$ and IAM and VM are equivalent.[3] Furthermore, as discussed in Section 3 under VM, we also can't separately identify the occupancy of the compliance groups $p_g = P(G_i = g)$ for $J > 1$, aside from never-takers and always-takers, without further assumptions. An interesting problem is whether for such identification it is sufficient to assume a linear single index model underlying selection.[4]

Note from the final row of Table 4 that the ACL is identified as a Wald ratio: $\rho_{(11),(00)}$. A natural question is whether the other parameters Δ_c falling under the purview of Theorem 1 can also be identified from Wald ratios for the various (z, w) that are ordered as vectors. That this conjecture is true follows from Corollary 1 (just subtract off $E[Y_i|Z_i = 0, \ldots 0]$ from each term and apply that by $E[h(Z_i)] = 0$ it follows that the total coefficient $\sum_{S,z} \lambda'_S A_{S,z}$ on $E[Y_i|Z_i = 0, \ldots 0]$ is equal to zero). Below I give an explicit example.

In a case with two binary instruments, suppose we are interested in $SLATE_1$, the average treatment effect among units for whom $D_i(1, Z_{2i}) > D_i(0, Z_{2i})$. This event is equivalent to the event that i is a Z_1 complier, or i is an *and*-complier and $Z_{2i} = 1$, or i is an *or*-complier and $Z_{2i} = 0$. If one forms the linear combination:

$$
\begin{aligned}
&P(Z_{2i} = 1)\left(E[Y_i|Z_i = (1,1)] - E[Y_i|Z_i = (0,1)]\right) \\
&\qquad\qquad\qquad + P(Z_{2i} = 0)\left(E[Y_i|Z_i = (1,0)] - E[Y_i|Z_i = (0,0)]\right) \\
&\qquad = P(Z_{2i} = 1)\left(p_{Z_1} + p_{reluctant}\right)\left[\frac{p_{Z_1}}{p_{Z_1} + p_{reluctant}}\Delta_{Z_1} + \frac{p_{reluctant}}{p_{Z_1} + p_{reluctant}}\Delta_{reluctant}\right] \\
&\qquad + P(Z_{2i} = 0)\left(p_{Z_1} + p_{eager}\right)\left[\frac{p_{Z_1}}{p_{Z_1} + p_{eager}}\Delta_{Z_1} + \frac{p_{eager}}{p_{Z_1} + p_{eager}}\Delta_{eager}\right] \\
&\qquad = P(Z_{2i} = 1)(p_{Z_1}\Delta_{Z_1} + p_{reluctant}\Delta_{reluctant}) + P(Z_{2i} = 0)(p_{Z_1}\Delta_{Z_1} + p_{eager}\Delta_{eager}) \\
&\qquad = p_{Z_1} \cdot \Delta_{Z_1} + P(Z_{2i} = 1)p_{reluctant} \cdot \Delta_{reluctant} + P(Z_{2i} = 0)p_{eager} \cdot \Delta_{eager}
\end{aligned}
$$

[2] Mountjoy (2019) alludes to this point in the context of a closely related monotonicty assumption.

[3] With IAM, we can identify Δ_g for all $g \in \mathcal{G}^c$ that occur with positive probability when the instruments have full support; the number of linearly-independent Walds is and $|\mathcal{G}^c|$ are both equal to $2^J - 1$.

[4] With continuous instruments and an assumption of no never-takers, Theorem 1 of Ichimura and Thompson (1998) would imply identification of F_β up to a scale normalization, in a model where $D_i = \mathbb{1}(Z_i'\beta_i \geq \beta_{0i})$, with $\beta_i \perp Z_i$ (the VM positivity restriction: $P(\beta_{ji} \geq 0) = 1$ for $j > 0$, is not necessary here for identification). Given the marginal distributions of β_i and Z_i, one could compute p_g for all $g \in \mathcal{G}$.

Similarly

$$\begin{aligned}
P(Z_{2i}=1)&\left(E[D_i|Z_i=(1,1)]-E[D_i|Z_i=(0,1)]\right)\\
&\qquad\qquad\qquad + P(Z_{2i}=0)\left(E[D_i|Z_i=(1,0)]-E[D_i|Z_i=(0,0)]\right)\\
&= p_{Z_1}+P(Z_{2i}=1)p_{reluctant}+P(Z_{2i}=0)p_{eager}\\
&= P\left(D_i(1,Z_{2i})>D_i(0,Z_{2i})\right)
\end{aligned}$$

Thus

$$\begin{aligned}
&\frac{P(Z_{2i}=1)\left(E[Y_i|Z_i=(1,1)]-E[Y_i|Z_i=(0,1)]\right)+P(Z_{2i}=0)\left(E[Y_i|Z_i=(1,0)]-E[Y_i|Z_i=(0,0)]\right)}{P(Z_{2i}=1)\left(E[D_i|Z_i=(1,1)]-E[D_i|Z_i=(0,1)]\right)+P(Z_{2i}=0)\left(E[D_i|Z_i=(1,0)]-E[D_i|Z_i=(0,0)]\right)}\\
&=\frac{p_{Z_1}}{D_i(1,Z_{2i})>D_i(0,Z_{2i})}\Delta_{Z_1}+\frac{P(Z_{2i}=1)p_{reluctant}}{D_i(1,Z_{2i})>D_i(0,Z_{2i})}\Delta_{reluctant}+\frac{P(Z_{2i}=0)p_{eager}}{D_i(1,Z_{2i})>D_i(0,Z_{2i})}\Delta_{eager}\\
&=\frac{P(G_i=Z_1)}{D_i(1,Z_{2i})>D_i(0,Z_{2i})}\Delta_{Z_1}+\frac{P(Z_{2i}=1\&G_i=and)}{D_i(1,Z_{2i})>D_i(0,Z_{2i})}\Delta_{reluctant}+\frac{P(Z_{2i}=0\&G_i=or)}{D_i(1,Z_{2i})>D_i(0,Z_{2i})}\Delta_{eager}\\
&=\frac{P(D_i(1,Z_{2i})>D_i(0,Z_{2i})\&G_i=Z_1)}{D_i(1,Z_{2i})>D_i(0,Z_{2i})}\Delta_{Z_1}+\frac{P(D_i(1,Z_{2i})>D_i(0,Z_{2i})\&G_i=and)}{D_i(1,Z_{2i})>D_i(0,Z_{2i})}\Delta_{reluctant}\\
&\qquad\qquad\qquad+\frac{P(D_i(1,Z_{2i})>D_i(0,Z_{2i})\&G_i=or)}{D_i(1,Z_{2i})>D_i(0,Z_{2i})}\Delta_{eager}\\
&=P(G_i=Z_1|D_i(1,Z_{2i})>D_i(0,Z_{2i}))\Delta_{Z_1}+P(G_i=and|D_i(1,Z_{2i})>D_i(0,Z_{2i}))\Delta_{reluctant}\\
&\qquad\qquad\qquad+P(G_i=or|D_i(1,Z_{2i})>D_i(0,Z_{2i}))\Delta_{eager}\\
&=E[Y_i(1)-Y_i(0)|D_i(1,Z_{2i})>D_i(0,Z_{2i})]=SLATE_1
\end{aligned}$$

To see that this same particular combination of causal effects is operationalized by the 2SLS-like estimator ρ_h suppose we choose $h(z)$ such that $Cov(Z_{1i},H_i)=1$, $Cov(Z_{2i},H_i)=0$, and $Cov(Z_{1i}Z_{2i},H_i)=P(Z_{2i}=1)$. Then, by Lemma 5:

$$\begin{aligned}
\rho_h &= \frac{p_{Z_1}\Delta_{Z_1}+p_{reluctant}P(Z_{2i}=1)\Delta_{reluctant}+p_{eager}(1-P(Z_{2i}=1))\Delta_{eager}}{p_{Z_1}+p_{reluctant}P(Z_{2i}=1)+p_{eager}(1-P(Z_{2i}=1))}\\
&= \frac{{}_{Z_1}\cdot\Delta_{Z_1}+P(Z_{2i}=1)p_{reluctant}\cdot\Delta_{reluctant}+P(Z_{2i}=0)p_{eager}\cdot\Delta_{eager}}{P\left(D_i(1,Z_{2i})>D_i(0,Z_{2i})\right)}
\end{aligned}$$

B.3 Identifying the Δ_g with $J=2$ when one is known

Consider differences in the propensity score $P(z):=E[D_i=z]$ between three of the pairs (z,w) of instrument values listed in Table 4 above. By the information in Table 4 and the law of iterated expectations, these yield sums of the group occupancies $p_g:=P(G_i=g)$, e.g:

$$\begin{pmatrix}P(1,0)-P(0,0)\\P(0,1)-P(0,0)\\P(1,1)-P(0,0)\end{pmatrix}=\begin{pmatrix}1&0&1&0\\0&1&1&0\\1&1&1&1\end{pmatrix}\begin{pmatrix}p_{Z_1}\\p_{Z_2}\\p_{eager}\\p_{reluctant}\end{pmatrix}$$

The choice of the three pairs listed above is arbitrary, but only three such differences can be linearly independent. If a single p_g were known, say p_{Z_1}, then the above equation can be appended and written as

$$\begin{pmatrix} p_{Z_1} \\ P(1,0)-P(0,0) \\ P(0,1)-P(0,0) \\ P(1,1)-P(0,0) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} p_{Z_1} \\ p_{Z_2} \\ p_{eager} \\ p_{reluctant} \end{pmatrix}$$

where the vector on the LHS is identified. The matrix on the RHS is invertible, which leads to identification of the four p_g for $g \in \mathcal{G}^c$.

Now consider the analogous Wald estimands $\rho_{z,w} := \frac{E[Y_i|Z_i=z]-E[Y_i|Z_i=w]}{E[D_i|Z_i=z]-E[D_i|Z_i=w]}$ between the same three pairs (z,w). By the law of iterated expectations, each will provide a weighted average over group specific treatment effects. Stacking these equations with Δ_{tution}, assumed to be known, we have:

$$\begin{pmatrix} \Delta_{Z_1} \\ \rho_{(1,0),(0,0)} \\ \rho_{(0,1),(0,0)} \\ \rho_{(1,1),(0,0)} \end{pmatrix} = \begin{pmatrix} p_{Z_1} & 0 & p_{eager} & 0 \\ 0 & p_{Z_2} & p_{eager} & 0 \\ p_{Z_1} & p_{Z_2} & p_{eager} & p_{reluctant} \end{pmatrix} \begin{pmatrix} \Delta_{Z_1} \\ \Delta_{Z_2} \\ \Delta_{eager} \\ \Delta_{reluctant} \end{pmatrix}$$
$$= \begin{pmatrix} p_{Z_1} & 0 & 0 & 0 \\ 0 & p_{Z_2} & 0 & 0 \\ 0 & 0 & p_{eager} & 0 \\ 0 & 0 & 0 & p_{reluctant} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \Delta_{Z_1} \\ \Delta_{Z_2} \\ \Delta_{eager} \\ \Delta_{reluctant} \end{pmatrix}$$

The diagonal matrix is known from the propensity score calculation above, and is invertible so long as all groups have non-zero size. The binary matrix is again invertible (it is the same as before), and thus the vector of all four Δ_g is identified as:

$$\begin{pmatrix} \Delta_{Z_1} \\ \Delta_{Z_2} \\ \Delta_{eager} \\ \Delta_{reluctant} \end{pmatrix} = = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1/p_{Z_1} & 0 & 0 & 0 \\ 0 & 1/p_{Z_2} & 0 & 0 \\ 0 & 0 & 1/p_{eager} & 0 \\ 0 & 0 & 0 & 1/p_{reluctant} \end{pmatrix} \begin{pmatrix} \Delta_{Z_1} \\ \rho_{(1,0),(0,0)} \\ \rho_{(0,1),(0,0)} \\ \rho_{(1,1),(0,0)} \end{pmatrix}$$

B.4 The matrix M_J for $J = 3$

B.5 Special cases of Lemma 5 under IAM

When we have a single binary instrument, consider the function $h(Z_i) = Z_i$. Under monotonicity, the compliance groups are $\mathcal{G} = \{never-taker, always-taker, complier\}$. The functions $\mathcal{D}_{a.t.}(Z_i)$ and $\mathcal{D}_{n.t.}(Z_i)$ are constants, so only the complier term contributes. Since $\mathcal{D}_{complier}(Z_i) = Z_i$, we have that $Cov(Y_i, h(Z_i)) = P_{complier}\Delta_{complier}$ and $Cov(D_i, h(Z_i)) = P_{complier}$, justifying the traditional Wald IV estimator for $\Delta_{complier}$.

In the case of a vector instrument with finite support and with IAM monotonicity, there is a well-defined ordering $z_1 \dots z_{\mathcal{M}}$ of points in $\mathcal{Z}$ (where $\mathcal{M} = |\mathcal{Z}|$) such that $P(D_i(z_m) \geq D_i(z_{m-1})) = 1$ (this ordering may be non-unique if there are no "compliers" between some

	{1}	{2}	{3}	{1,2}	{1,3}	{2,3}	{1,2,3}
{1}	1						
{2}		1					
{3}			1				
{1,2}				1			
{1,3}					1		
{2,3}						1	
{1,2,3}							1
{1},{2}	1	1		-1			
{2},{3}		1	1			-1	
{1},{3}	1		1		-1		
{1},{2},{3}	1	1	1	-1	-1	-1	1
{1,2},{3}			1	1			-1
{1,3},{2}		1			1		-1
{2,3},{1}	1					1	-1
{1,2},{1,3}				1	1		-1
{1,2},{2,3}				1		1	-1
{1,3},{2,3}					1	1	-1
{1,2},{1,3},{2,3}				1	1	1	-2

Table 5: The matrix M_3 defined in Section 4. Empty cells indicate a zero.

pairs of points in $\mathcal{Z}$). If we choose $h(Z_i) = P(Z_i)$, the propensity score function, ρ_h is equal to the "regression on the propensity score" estimator $Cov(Y_i, P(Z_i))/Var(P(Z_i))$ (since $Cov(D_i, P(Z_i)) = Var(P(Z_i)))$. Note that under IAM G_i maps one-to-one with as the smallest m for which $D_i(z_m) = 1$ (provided i is not a never-taker). Thus we can use the notation $D_m(z) := \mathbb{1}(z \succ z_m)$, indicating that z succeeds z_m in the sequence $z_1 \dots z_{\mathcal{M}}$. The weights are positive so long as for all m: $Cov(D_m(Z_i), P(Z_i)) \geq 0$, which occurs iff:

$$\begin{aligned} &E[P(Z_i)|D_m(Z_i) = 1] - E[P(Z_i)|D_m(Z_i) = 0] \\ &\qquad = E[P(Z_i)|Z_i \succ z_m] - E[P(Z_i)|Z_i \preceq z_m] > 0 \end{aligned}$$

This inequality will hold for all m, since (using independence) $P(z_m)$ is monotonically increasing in m. This yields a novel demonstration of fact that the"regression on the propensity score" estimator identifies a convex combination of LATEs, as shown in Imbens and Angrist (1994).

B.6 Expressing identified parameters in terms of partial treatment effects

Starting with Equation (4), and applying the law of iterated expectations (three times):

$$\begin{aligned}
\Delta_c &= \sum_{z \in \mathcal{Z}} P\left(Z_i = z \left| \bigcup_{k=1}^{K} \{D_i(u_k(Z_i)) > D_i(l_k(Z_i))\}\right.\right) \cdot E\left[Y_i(1) - Y_i(0) \left| \bigcup_{k=1}^{K} \{D_i(u_k(z)) > D_i(l_k(z))\}\right.\right] \\
&= \sum_{z \in \mathcal{Z}} \frac{P(Z_i = z) P\left(\bigcup_{k=1}^{K} \{D_i(u_k(z)) > D_i(l_k(z))\}\right)}{P\left(\bigcup_{k=1}^{K} \{D_i(u_k(Z_i)) > D_i(l_k(Z_i))\}\right)} \cdot E\left[Y_i(1) - Y_i(0) \left| \bigcup_{k=1}^{K} \{D_i(u_k(z)) > D_i(l_k(z))\}\right.\right] \\
&= \sum_{z \in \mathcal{Z}} \sum_{k=1}^{K} \frac{P(Z_i = z) P\left(D_i(u_k(z)) > D_i(l_k(z))\right)}{P\left(\bigcup_{k'=1}^{K} \{D_i(u_{k'}(Z_i)) > D_i(l_{k'}(Z_i))\}\right)} \cdot E\left[Y_i(1) - Y_i(0) \left| D_i(u_k(z)) > D_i(l_k(z))\right.\right] \\
&= \sum_{z \in \mathcal{Z}} \sum_{k=1}^{K} \sum_{m=1}^{M_{kz}} \frac{P(Z_i = z) P\left(D_i(u_{kz}^m) > D_i(l_{kz}^m)\right)}{\sum_{z' \in \mathcal{Z}} \sum_{k'=1}^{K} \sum_{m'=1}^{M_{k'z'}} P(Z_i = z') P\left(D_i(u_{k'z'}^{m'}) > D_i(l_{k'z'}^{m'})\right)} \cdot PTE_{j(z,k,m)}(l_{kz}^m)
\end{aligned}$$

where for each z and k: $\{u_{kz}^m, l_{kz}^m\}_{m=1}^{M_{kz}}$ is a sequence of single-instrument swaps where u_{kz}^m is obtained from l_{kz}^m by flipping instrument $j(z,k,m)$ from 0 to 1, while $l_{kz}^m = u_{kz}^{m+1}$, $l_{kz}^1 = l_k(z)$, and $u_{kz}^{M_{kz}} = u_k(z)$. Note that each $P\left(D_i(u_k^m(z)) > D_i(l_k^m(z))\right)$ can be identified as $E[D_i|Z_i = u_k^m(z)] - E[D_i|Z_i = l_k^m(z)]$.

B.7 Vector monotonicity in Bloom scenarios

In some empirical settings, particular combinations of instrument values and treatment status may be impossible. For example, in a randomized trial of an experimental drug, it may not be feasible to obtain the drug ($D_i = 1$) without being assigned to treatment in the trial ($Z_i = 1$). In the Minneapolis Domestic Violence Experiment analyzed by Angrist (2006), all police officers who were assigned to respond to domestic violence complaint with arrest ($Z_i = 1$) indeed arrested offenders ($D_i = 1$). Such instances are what Angrist and Pischke (2008) refer to as “Bloom scenarios”. In the drug trial example, there are no always-takers. In the domestic violence experiment, there are no never-takers.[5]

In a case with multiple instruments $Z_1 \dots Z_J$, a Bloom scenario holding for one of the instruments implies restrictions on the compliance groups that can occur. In this section, I detail the implications of a single-instrument Bloom scenario with multiple binary instruments satisfying VM.

Consider first a case in which there are no always-takers with respect to Z_1, that is $D_i(0, z_{-j}) = 0$ with probability one for any $z_{-j} \in \mathcal{Z}_{-j}$. Recall from Section 3 that every compliance group under VM maps to a Sperner family on the set of instrument labels, aside from the group of never-takers. For any such family F, it must be the case that each set in F contains 1, since otherwise the compliance group F would be able to

[5] In both cases, however there would still typically be imperfect compliance overall. In the drug trial example, if some individuals assigned to treatment do not take the drug, they must be never-takers (or defiers). In the domestic violence experiment, sometimes officers arrested particularly dangerous individuals even when they were assigned to alternative responses.

take treatment with $Z_1 = 0$. For instance, with two binary instruments, eager compliers ($F = \{1\}, \{2\}$) cannot exist, since they take treatment when $Z_1 = 0, Z_2 = 1$. The possible groups in this $J = 2$ example are: never-takers, Z_1 compliers, and reluctant compliers.

That three compliance groups remain, equal to the number of compliance groups that exist under VM with a single binary instrument, when $J = 2$ fits a general pattern. With J instruments and the no-always-takers Bloom condition on Z_1, the compliance groups can be constructed as follows: take the Sperner family F associated with any VM compliance group on the instrument labels $2 \ldots J$, and include "1" in each $S \in F$. In addition, there will be never-takers, and thus the total number of compliance groups will be equal to the $J - 1^{th}$ number $\mathcal{D}_{J-1}$ in the Dedekind sequence (see Section 3). For instance, with $J = 3$, we have:

F on {2,3}	**corresponding F on {1,2,3}**
$\emptyset$	$\{1\}$
$\{2\}$	$\{1,2\}$
$\{3\}$	$\{1,3\}$
$\{2\}, \{3\}$	$\{1,2\}, \{1,3\}$
$\{2,3\}$	$\{1,2,3\}$

Together with never-takers, there are $\mathcal{D}_2 = 6$ compliance groups in this case.

Now consider the opposite case, where there are no never-takers with respect to Z_1, that is $D_i(1, z_{-j}) = 1$ with probability one for any $z_{-j} \in \mathcal{Z}_{-j}$. In this case any Sperner family corresponding to a compliance group must include a set containing only the label 1, since $Z_1 = 1$ is sufficient for any unit to take treatment. Similar to the previous case, the number of compliance groups satisfying VM will be $\mathcal{D}_{J-1}$, and they can be obtained from the Sperner families corresponding to VM compliance groups for the $J - 1$ instruments $Z_2 \ldots Z_J$. In this case, we simply append the set $S = \{1\}$ to any Sperner family on the set $\{2 \ldots J\}$:

F on {2,3}	**corresponding F on {1,2,3}**
$\emptyset$	$\{1\}$
$\{2\}$	$\{1\}, \{2\}$
$\{3\}$	$\{1\}, \{3\}$
$\{2\}, \{3\}$	$\{1\}, \{2\}, \{3\}$
$\{2,3\}$	$\{1\}, \{2,3\}$

Together with always-takers, there are again $\mathcal{D}_2 = 6$ compliance groups in total.

In these Bloom scenarios, the identification result of Theorem 1 will still be valid, and estimation can proceed as it does in general. Although the Bloom conditions imply restrictions on the propensity score function (e.g. that $E[D_i|Z_i = (0, z_2, z_3)] = 0$ in the first example, for all $z_2, z_3 \in \{0, 1\}$), they do not imply any restrictions that threaten

Assumption 3. In fact, identification is strengthened in these scenarios since they introduce overidentification restrictions. For example, in the second case considered with three binary instruments and no never takers with respect to Z_1, the model implies that

$$
\begin{aligned}
ACL &= \frac{E[Y_i|Z_i=(1,1,1)] - E[Y_i|Z_i=(0,0,0)]}{1-E[D_i|Z_i=(0,0,0)]} = \frac{E[Y_i|Z_i=(1,0,1)] - E[Y_i|Z_i=(0,0,0)]}{1-E[D_i|Z_i=(0,0,0)]} \\
&= \frac{E[Y_i|Z_i=(1,1,0)] - E[Y_i|Z_i=(0,0,0)]}{1-E[D_i|Z_i=(1,0,0)]} = \frac{E[Y_i|Z_i=(1,0,0)] - E[Y_i|Z_i=(0,0,0)]}{1-E[D_i|Z_i=(0,0,0)]}
\end{aligned}
$$

which implies the overidentification restriction that $E[Y_i|Z_i = (1, z_2, z_3)] = 0$ does not depend on $z_2, z_3 \in \{0, 1\}$. In all cases, it is equal to the unconditional average of the treated potential outcome: $E[Y_i(1)]$.

C Additional figures

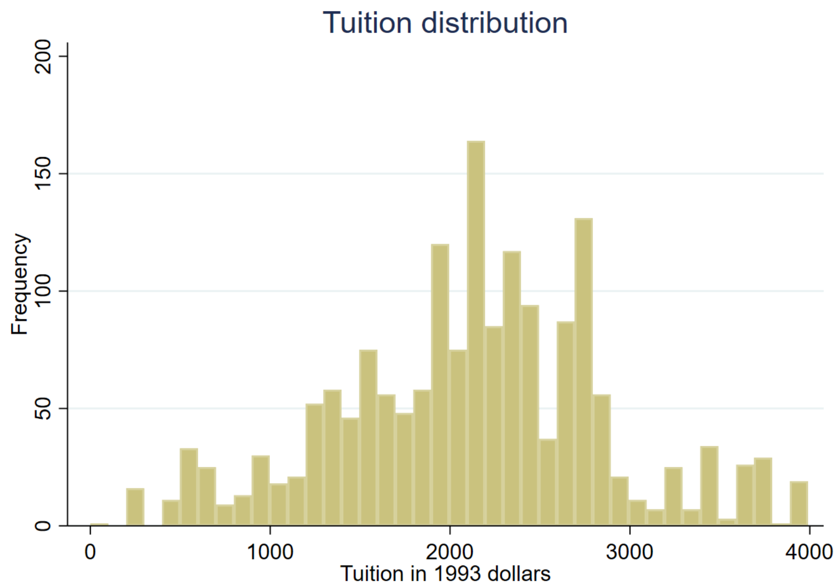

Figure 1: Empirical distribution of the tuition variable from Section 6.

D Additional Empirical Results

D.1 Additional empirical example: the effect of children on labor supply

In this section I revisit the analysis of Angrist and Evans (1998), who study the effect of family size on parental labor supply. Angrist and Evans consider two types of instruments that induce exogenous variation in family size among families that have at least two children: i) whether the first two children have the same sex; and ii) whether the second birth was a multiple birth (i.e. twins, triplets, etc.). Since twins account for the overwhelming majority of multiple births, I refer to multiple births as simply “twins”.

If the first two children in a family have the same sex, this may cause some parents to have a third child in an effort to have children of both sexes. If some parents' furthermore have a preference for having at least one boy, they may respond only to this same sex instrument if the first two are girls, and vice versa if they have a preference for girls. These various sex-preferences can be modeled by two binary instruments for whether mother i has a third or more children (indicated by $D_i = 1$): $Z_{1i} = \mathbb{1}$(i's first two children are girls) and $Z_{2i} = \mathbb{1}$(i's first two children are boys). Vector monotonicity is a reasonable assumption here, saying that $D_i(1,0) > D_i(0,0)$ with probability one and $D_i(0,1) > D_i(0,0)$ with probability one – no mother would have a third child only because her first two kids were of the opposite sex. These two instruments may have distinct "complier" populations, since some parents' may seek a girl, some may seek a boy, and some may seek at least one of each.

Note that VM places no restrictions on $D_i(1,1)$, since the point $Z_1 = Z_2 = 1$ can be ruled out of the set $\mathcal{Z}$ of possible instrument values – this would mean having the first two children be both girls and both boys. As a result, VM and PM are equivalent for these two instruments. Note that IAM can only hold for these two instruments if all mothers who would have a third child with two boys would also have a third child with two girls, or vice versa. This is a strong restriction, which rules out there being some parents who seek at least one girl, and other parents who seek at least one boy.

Twinning (or triplets, etc.) can be thought of as introducing a third binary instrument for family size: $Z_{3i} = \mathbb{1}$(i's second birth was a multiple birth). As the data as coded to record actual births (not inclusive of pregnancies not carried to term), we have a so-called "Bloom scenario" with respect to Z_3: all mothers with a multiple birth after their first child have at least three children. The implications of such Bloom scenarios under VM with multiple instruments is considered in Section B.7. All together, there are five possible compliance groups in the population:

group name	$\mathbf{Z_i = (0,0,0)}$	$\mathbf{Z_i = (1,0,0)}$	$\mathbf{Z_i = (0,1,0)}$	$\mathbf{Z_i = (0,0,1)}$	$\mathbf{Z_i = (1,0,1)}$	$\mathbf{Z_i = (0,1,1)}$
girl compliers	N	T	N	T	T	T
boy compliers	N	N	T	T	T	T
same-sex compliers	N	T	T	T	T	T
twin compliers	N	N	N	T	T	T
always-takers	T	T	T	T	T	T

Table 6: Compliance groups under VM for three instruments drawn from Angrist and Evans (1998), where $Z = (twogirls, twoboys, twins)$.

There are only six columns in Table 6, rather than eight (2^3), because the points $Z_i = (1,1,0)$ and $Z_i = (1,1,1)$ are excluded from $\mathcal{Z}$.[6] Since the support of Z_i is not rectangular (Assumption 3), we will need to appeal to the more general identification result from Appendix A. For this result, the rectangular support condition is replaced by Assumption 3*, which states that there exists a family $\mathcal{F}$ of products of the instruments that are

[6]In addition reducing the support of the instruments by two points, this also reduces the number of compliance groups: there is no "reluctant complier" groups for instruments Z_1 and Z_2.

linearly independent and span the space of compliance functions $\mathcal{D}_g(Z_i)$ for $g \in \mathcal{G}^c$ (here $\mathcal{G}^c$ consists of the four groups that are not never-takers). From Table 6 a spanning family can be seen to be $\mathcal{F} = \{\{1\}, \{2\}, \{3\}, \{1,3\}, \{2,3\}\}$, since we have $\mathcal{D}_{girl}(z) = z_1 + z_3 - z_1 z_3$, $\mathcal{D}_{girl}(z) = z_2 + z_3 - z_2 z_3$, $\mathcal{D}_{twin}(z) = z_3$, and $\mathcal{D}_{samesex}(z) = \mathcal{D}_{girl}(z) + \mathcal{D}_{boy}(z)$. In estimation then, I use the vector $\Gamma_i = (Z_{1i}, Z_{2i}, Z_{3i}, Z_{1i}Z_{3i}, Z_{2i}Z_{3i})'$.

I use the dataset considered by Angrist and Evans, 1998 drawn from the 1980 U.S. census, creating a sample of 394,840 mothers between the ages of 21 and 35 with multiple children (this nearly replicates the sample in Angrist and Evans 1998, which has 5 fewer observations). Table 7 reports the distribution of the instruments and the (unconditional) propensity score function. Angrist and Evans note that twin births may not be unconditionally exogenous, as they are known to be more likely for older mothers and African-American mothers. For simplicity, I ignore this issue and do not condition on any demographic variables X.

$(Z_1, Z_2) =$	Z_3=twins **0**	Z_3=twins **1**	
(0,0)	193,567	1,725	*195,292*
(1,0)	94,618	803	*95,421*
(0,1)	103,275	852	*104,127*
	391,460	*3,380*	

Z_1=two girls	Z_2=two boys **0**	Z_2=two boys **1**
0	.336	.418
1	.436	n/a

Table 7: Cross-tabulation (left) and propensity scores (right) for the three instruments. Propensity scores reported are $\hat{E}[D_i|Z_i = (z_1, z_2, 0)]$ for $z_1, z_2 \in \{0, 1\}$. For all z_1, z_2: $\hat{E}[D_i|Z_i = (z_1, z_2, 1)] = 1$ so these values are not reported. Total $N = 394,840$.

From the left panel of Table 7, we can see that having two boys as the first two children is about 10% more likely then having two girls, whether or not the second birth was a multiple birth. On the right panel ,we see that the data is consistent with vector monotonicity as described above. The proportion of always-takers is identified as 33.6%, which given that there are no never-takers implies that the remaining 66.4% of mothers in this population respond to the three instruments in some way. The propensity score estimates imply that the total of girl compliers and same-sex compliers is 10% of the population, and that the total of boy compliers and same-sex compliers is 15.7%. This indicates that the are nearly 6 percentage points more boy compliers than there are girl compliers among U.S. mothers.

I consider four choices of the outcome variable Y_i drawn from Angrist and Evans, 1998: i) the mother's labor income in the year prior to the census (1979 dollars); ii) weeks worked in the year prior to the census; ii) average hours woekd per week, and iv) an indicator for whether mother i worked for pay (any of i)-iii) are positive). Treatment effect estimates are reported in Table 8. Across all four outcome measures, having a third child is estimated to cause a significant reduction in mothers' labor force participation. As a result of having more than two children, mothers who respond to any of the instruments (the ACL) are 11% less likely to work for pay, work 4 hours less per week and nearly

	(1) 2SLS	(2) ACL	(3) SLATE(girls)	(4) SLATE(boys)	(5) SLATE(twins)
Worked for pay	-0.0889***	-0.111***	-0.102***	-0.182***	-0.106***
	(-7.13)	(-6.59)	(-3.67)	(-4.94)	(-5.55)
Hours worked for pay	-3.573***	-3.970***	-3.478**	-7.459***	-3.740***
	(-7.68)	(-6.29)	(-3.27)	(-5.29)	(-5.23)
Weeks worked for pay	-3.85***	-4.677***	-4.122**	-9.486***	-4.347***
	(-6.99)	(-6.37)	(-3.29)	(-5.74)	(-5.23)
Labor income (1979 USD)	-512.61***	-692.8***	-755.1**	-1539.1***	-617.1**
	(-4.07)	(-3.99)	(-2.67)	(-4.15)	(-3.12)
Size of compliant pop.	n/a	0.755***	0.0700***	0.0515***	0.634***
		(190.30)	(35.93)	(27.34)	(578.59)

t statistics in parentheses, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8: Treatment effect estimates. Each row indicates a different choice of the outcome variable Y_i, while columns (1)-(5) correspond to alternative treatment effect estimators. 2SLS is fully saturated, including interactions between the gender instruments and the twin instruments. Estimates of various Δ_c (columns 2-5) use the unregualrized estimator $\hat{\rho}(\hat{\lambda}, 0)$. Size of compliant population reports estimates of $P(C_i = 1)$. $N = 394,840$.

5 weeks less per year, earning about $700 less for the year. These estimates are all somewhat larger in magnitude than the corresponding 2SLS estimates.

Columns 3-5 of Table 8 report Set LATEs for each of the three instruments individually. The set of compliers to the twins instrument is much larger than those for the boys or girls instrument (63% vs. 5% or 7% of the population); however the estimated value of SLATE(twins) is similar to that of SLATE(girls) across the outcome variables. By contrast, estimates of the LATE among mothers who respond to the two-boys instrument are much larger in magnitude than all other VM treatment effect parameters (columns 2, 3, and 5). This suggests that the mothers most likely to reduce their labor supply–and by much more–are those who had the third kid as a result of seeking a girl after first having two boys.

Finally, I note that Assumption 1 along with the "Bloom condition" that all mothers with a multiple second birth take treatment implies the overidentification restriction that

$$E[Y_i|Z_i = (0, 0, 1)] = E[Y_i|Z_i = (1, 0, 1)] = E[Y_i|Z_i = (0, 1, 1)]$$

(see Section B.7 for details). I test this restriction via an F-test by regressing Y_i on indicators for the cells $(1, 0, 1)$ and $(0, 1, 1)$ and a constant, restricted to the twins subsample. The model implies that both regression coefficients for the cell indicators should be zero. The p-value for the regression F-statistic is about .05 when Y_i indicates worked for pay, .22 for hours worked for pay, .07 for weeks worked for pay, and .19 for mother's income. While we only fail to reject the null-hypothesis implied by the model at the 5% level for the worked for pay outcome, these results could be interpreted as providing mild evidence against the validity of the model. One possible explanation is through number of children: mothers who have gender parity between their first two children as well as twins might be more likely to have *four* or more children, as compared with mothers with mixed genders among their first two kids that also have twins. Such logic challenges the exclusion restriction when this setting is considered with a binary definition of treatment and both twinning and sex-mix instruments. The estimates here should thus be interpreted with caution, as the methods in this paper have not yet been extended to cases with multi-valued treatment.

E Proofs

E.1 Proof of Appendix Proposition 11

We restate the result here:

Proposition 1. *Let the support $\mathcal{Z}$ of the instruments be discrete and finite. Fix a function $c(g, z)$. Let $\mathcal{P}_{DZ}$ denote the joint distribution of D_i and Z_i. Then the following are equivalent:*

1. *Δ_c is (point) identified by $\mathcal{P}_{DZ}$ and $\{\beta_s\}_{s\in\mathcal{S}}$, for some finite set $\mathcal{S}$ of known or identified (from $\mathcal{P}_{DZ}$) measurable functions $s(d, z)$, and $\beta_s := E[s(D_i, Z_i)Y_i]$*
2. *$\Delta_c = \beta_s$ for a single such $s(d, z)$*
3. *$\Delta_c = E[t(D_i, Z_i, Y_i)]$ with $t(d, z, y)$ a known or identified (from $\mathcal{P}_{DZ}$) measurable function*
4. *Δ_c is identified from the set of CEFs $\{E[Y_i|D_i = d, Z_i = z]\}$ for $d \in \{0, 1\}$, $z \in \mathcal{Z}$ along with the joint distribution $\mathcal{P}_{DZ}$*

Where the meaning of "identified" here is that the set of values of a parameter that are compatible with a set of empirical estimands is a singleton, regardless of the distribution of the latent variables $(G_i, Y_i(1), Y_i(0))$ – for all $\mathcal{P}_{DZ}$ within some class (note that the marginal distribution of G_i must also be compatible with $\mathcal{P}_{DZ}$).

Proof. We can show each of the following implications:

- **1 → 4** Any β_s can be written: $\beta_s = \sum_{d,z} P(D_i = d, Z_i = z)s(d, z)E[Y_i|D_i = d, Z_i = z]$, and is thus pinned down by the CEFs $E[Y_i|D_i = d, Z_i = z]$, the joint distribution $\mathcal{P}_{DZ}$, and the known function s.
- **4 → 1** Let $\mathcal{S} = \{s(d, z) = \mathbb{1}(D_i = d)\mathbb{1}(Z_i = z)\}_{d\in\{0,1\},z\in\mathcal{Z}}$. Then each β_s is equal to $P(D_i = d, Z_i = z)E[Y_i|D_i = d, Z_i = z]$ for some d, z. The coefficient is known from $\mathcal{P}_{DZ}$, thus **4.** is a case of **1.**
- **2 → 1** Immediate, since 3 is a special case of 2 with $\mathcal{S}$ a singleton
- **4 → 2** Write any $E[Y_i|D_i = d, Z_i = z] = E[Y_i(d)|D_i = d, Z_i = z] = P(D_i = d|Z_i = z)^{-1}E[Y_i(d)\mathbb{1}(D_i = d)|Z_i = z] = P(D_i = d|Z_i = z)^{-1} \sum_g P(G_i = g|Z_i =$

$z)E[Y_i(d)\mathbb{1}(D_i = d)|G_i = g, Z_i = z] = P(D_i = d|Z_i = z)^{-1}\sum_{g:\mathcal{D}_g(z)=d} P(G_i = g)E[Y_i(d)|G_i = g]$, where we have used independence.

To eliminate the coefficient, simply write: $E[Y_i\mathbb{1}(D_i = d)|Z_i = z] = \sum_{g:\mathcal{D}_g(z)=d} P(G_i = g)E[Y_i(d)|G_i = g]$. If we stack the unknown quantities $P(G_i = g)E[Y_i(d)|G_i = g]$ for all $g \in \mathcal{G}, d \in \{0,1\}$ into a vector x, and the identified quantities $E[Y_i\mathbb{1}(D_i = d)|Z_i = z]$ for all $d \in \{0,1\}, z \in \mathcal{Z}$ into a vector b, then we have a system of linear equations $Ax = b$, where A is a fixed matrix of entries of the form $[A]_{dz,g} = \mathbb{1}(\mathcal{D}_g(z) = d)$. Note that the matrix A here is not the same as the matrix A defined in Corollary 1 to Theorem 1.

Similarly, as we have seen Δ_c can be written as a linear combination of the components of the vector z. Specifically, from Equation(3):

$$\Delta_c = \sum_g \frac{E[c(g, Z_i)]}{E[c(G_i, Z_i)]} P(G_i = g)\left\{E[Y_i(1)|G_i = g] - E[Y_i(0)|G_i = g]\right\}$$

We can now write $\Delta_c = \theta_c' x$, where θ_c is the vector of coefficients $\pm\frac{E[c(g,Z_i)]}{E[c(G_i,Z_i)]}$ from the above equation.

The set of vectors x compatible with the set of identifying restrictions $Ax = b$ can be written as $\{A^\dagger b + (I - A^\dagger A)w\}$ for all arbitrary vectors $w \in \mathbb{R}^{2|\mathcal{G}|}$, where $A^\dagger$ is the Moore-Penrose pseudo-inverse of A. The corresponding set of values for Δ_c is $\{\theta_c' A^\dagger b + \theta_c'(I - A^\dagger A)w\}$. For this set to be a singleton for all w, we must either have $A^\dagger A = I$ (i.e. A has full column rank, which is only possible in the $J = 1$ case), or the vector θ_c must lie in the row space of the matrix A, so that in either case $\theta_c'(I - A^\dagger A)$ is equal to the zero vector. If the set were not a singleton, then Δ_c would not be identified absent additional restrictions, since an infinite collection of values of Δ_c would be compatible with the full set of restrictions $Ax = b$ placed by the CEF's $E[Y_i|D_i = d, Z_i = z]$. Thus, by **4.**, we have that $\Delta_c = \theta_c' A^\dagger b$. This then implies **2.**, if we take $s(d, z) = \frac{P(D_i=d|Z_i=z)}{P(D_i=d,Z_i=z)} \cdot [\theta_c' A^\dagger]_{(d,z)}$, where $[\theta_c' A^\dagger]_{(d,z)}$ is the component of the vector $\theta_c' A^\dagger$ corresponding to the pair (d, z). Note that $A^\dagger$ is a known matrix (without looking at the data), and θ_c is a known function of the marginal distribution of Z_i, up to the factor $E[c(G_i, Z_i)]$, for a fixed function c.

It only remains to be shown that $E[c(G_i, Z_i)]$ is also identified under assumption of **1**. For Δ_c to be pinned down for all joint distributions of $(G_i, Y_i(1), Y_i(0))$, it must be pinned down in the special case where in which $Y_i(d) = d$ with probability one. In this case each $E[Y_i|D_i = d, Z_i = z] = d$, and $\Delta_c = 1$. Thus, using our result above we have that $E[c(G_i, Z_i)] = E[\tilde{s}(d, z)D_i]$, where $\tilde{s}(d, z) := \frac{P(D_i=d|Z_i=z)}{P(D_i=d,Z_i=z)} \cdot [\tilde{\theta}_c' A^\dagger]_{(d,z)}$, where $\tilde{\theta}_c := E[c(G_i, Z_i)]\theta_c = \pm E(c(g, Z_i))$.

- **2 → 3** This is immediate, since $s(d,z)y$ is a possible function $t(d,z,y)$.
- **3 → 2** Consider a joint distribution F of potential outcomes, compliance groups, and instruments, and an alternative distribution F', where the potential outcomes are rescaled by a factor $b \in \mathbb{R}$: i.e. if $(Y_i(1), Y_i(0), G_i) \sim F$ then $(bY_i(1), bY_i(0), G_i) \sim F$. Let $\Delta_c(\cdot)$ denote the causal parameter Δ_c as a function of the joint distribution of $(Y_i(1), Y_i(0), G_i)$, fixing $\mathcal{P}_Z$. Clearly $\Delta(F') = b\Delta_c(F)$. Note that if the distribution of Z_i is held fixed, the distribution of (Y_i, D_i, Z_i) under F' is the same as the distribution of (bY_i, D_i, Z_i) under F, since $Y_i = Y_i(0) + D_i(Y_i(1) - Y_i(0))$. Thus, by assumption that $\beta_s = \Delta_c(F')$ when the observables are generated under F', we must have that $E[s(D_i, Z_i, bY_i)] = bE[s(D_i, Z_i, Y_i)]$. For this to be true for any distribution of $(Y_i(1), Y_i(0), G_i)$ compatible with $P_{\mathcal{DZ}}$ in some class, it must be that $s(d,z,by) = bs(d,z,y)$ for all y, b and d, z having support in such $P_{\mathcal{DZ}}$. Defining $s(d,z)$ as $s(d,z,1)$, we can then write $s(d,z,y)$ as $s(d,z)y$.[7]

□

E.2 Proof of Proposition 8

Let $\mathcal{Y} \subseteq \mathbb{R}$ be the support of Y_i. For any $y \in \mathcal{Y}, z \in \mathcal{Z}$ and $d \in \{0,1\}$, let $F_{(Y\mathbb{1}_d)|Z}(y|z) := E[\mathbb{1}(Y_i \le y)\mathbb{1}(D_i = d)|Z_i = z]$. The strategy will be based on the fact that knowing the distribution of (Y_i, D_i, Z_i) is equivalent to knowing $F_{(Y\mathbb{1}_d)|Z}(y|z)$ for all (y,d,z) along with the marginal distribution of Z_i. By the law of iterated expectations and Assumption 1:

$$F_{(Y\mathbb{1}_d)|Z}(y|z) = E[\mathbb{1}(Y_i \le y)\mathbb{1}(D_i = d)|Z_i = z] = \sum_{g:\mathcal{D}_g(z)=d} P(G_i = g)E[\mathbb{1}(Y_i(d) \le y)|G_i = g]$$

$$= \sum_{g:\mathcal{D}_g(z)=d} P(G_i = g)F_{Y(d)|G}(y|g) := \sum_{g\in\mathcal{G}} A^d_{zg} \cdot P(G_i = g)F_{Y(d)|G}(y|g) \qquad (1)$$

where A^d is a $|\mathcal{Z}| \times |\mathcal{G}|$ matrix with typical entry $[A^d]_{zg} = \mathbb{1}(\mathcal{D}_g(z) = d)$. Now create a $2|\mathcal{Z}| \times |\mathcal{G}|$ matrix A by stacking A^0 and A^1 row-wise. Define $F_{(Y\mathbb{1}_D)|Z}(y)$ to be a $2|\mathcal{Z}| \times 1$ vector of $F_{(Y\mathbb{1}_D)|Z}(y|z)$ over z and d and $F^P_{Y(D)|G}(y)$ to be a $2|\mathcal{G}| \times 1$ vector of $P(G_i = g) \cdot F_{Y(d)|G}(y|g)$ over g and d. We can then write Equation (1) for any fixed value of y as $F_{(Y\mathbb{1}_D)|Z}(y) = AF^P_{Y(D)|G}(y)$. Now let $\mathbf{F}^P_{Y(D)|G}$ represent the whole vector-valued function $F^P_{Y(D)|G}(y) : \mathcal{Y} \to \mathbb{R}^{2|\mathcal{G}|}$, and define $\mathbf{F}_{(Y\mathbb{1}_D)|Z}$ similarly as the function $\mathcal{Y} \to \mathbb{R}^{2|\mathcal{Z}|}$ yielding the vector $F_{(Y\mathbb{1}_D)|Z}(y)$. We now write Eq. (1) holding across all $y \in \mathcal{Y}$ as:

$$\mathbf{F}_{(Y\mathbb{1}_D)|Z} = \mathcal{A} \circ \mathbf{F}^P_{Y(D)|G} \qquad (2)$$

where $\mathcal{A}$ denotes the linear map of functions $\mathcal{Y} \to \mathbb{R}^{2|\mathcal{G}|}$ to functions $\mathcal{Y} \to \mathbb{R}^{2|\mathcal{Z}|}$ defined by:

$$[\mathcal{A} \circ \boldsymbol{\mu}(y)]_{dz} = \sum_g A_{dz,g}\boldsymbol{\mu}(y)_{dg}$$

[7] Note that a similar argument with an F' such that $(Y_i(1)+b, Y_i(0)+b, G_i, Z_i) \sim F$ reveals that the random variable $s(D_i, Z_i)$ must be mean zero.

holding pointwise for each y.

Note that Δ_c can also be written as a linear map applied to the function $\mathbf{F}^P_{Y(d)|G}$. In particular, $\Delta_c = \Theta_c \circ \mathbf{F}^P_{Y(d)|G}$, where for any function $\boldsymbol{\mu}(y)$ from $\mathcal{Y}$ to $\mathbb{R}^{2|\mathcal{G}|}$, $\Theta_c \circ \boldsymbol{\mu}$ is the scalar:

$$\sum_g \frac{E[c(g, Z_i)])}{E[c(G_i, Z_i)]} \left(\int_{\mathcal{Y}} y \cdot d\boldsymbol{\mu}_{1g}(y) - \int_{\mathcal{Y}} y \cdot d\boldsymbol{\mu}_{0g}(y) \right) \tag{3}$$

where $d\mu_{dg}$ is either: a) the differential operator applied to the scalar function of y defined by the dg element of $\boldsymbol{\mu}$ over $y \in \mathcal{Y}$, if Y_i is continuously distributed; or b) the counting measure over $\mathcal{Y}$ if Y_i has a discrete distribution.

Note that under the assumptions of the proposition, the map Θ_c must be identified. To begin, the $E[c(g, Z_i)])$ are identified from the distribution $\mathcal{P}_Z$ of the instruments, given that the function c is known. Identification of $E[c(G_i, Z_i)] = P(C_i = 1)$ requires knowledge of $\mathcal{P}_{DZ}$ but follows from identification of Δ_c in the special case in which $Y_i(0) = 1$ and $Y_i(1) = 0$ both with probability one (which is compatible with any $\mathcal{P}_{DZ}$). Such data can always be constructed synthetically given the observed joint distribution of (D_i, Z_i). Then $E[c(G_i, Z_i)]$ is the unique value such that $\Theta_c \circ \mathbf{F}^P_{\mathbb{1}(d=1)|G} = \Delta_c = 1$.

Back to the general case, note that any joint distribution $\mathcal{P}_{latent}$ of the latent variables $(Y_i(1), Y_i(0), G_i)$ implies a $\mathbf{F}^P_{Y(d)|G}$, and if this $\mathbf{F}^P_{Y(d)|G}$ satisfies Eq. (2) for the observed function $\mathbf{F}_{(Y\mathbb{1}_D)|Z}$ then $\mathcal{P}_{latent}$ will be fully compatible with the data, since $\mathbf{F}_{(Y\mathbb{1}_D)|Z}$ exhausts all remaining knowledge of observables given $\mathcal{P}_{DZ}$; more concretely, the joint distribution of (Y_i, D_i, Z_i) is equivalent to the pair $(\mathbf{F}_{(Y\mathbb{1}_D)|Z}, \mathcal{P}_Z)$. Thus, Δ_c can only be identified if $\Theta \circ \mathbf{F}^P_{Y(d)|G}$ returns a singleton when applied to the set of all $\mathbf{F}^P_{Y(d)|G}$ satisfying Eq. (2), such that the various $F_{Y(d)|G}(y|g)$ implied by $\mathbf{F}^P_{Y(d)|G}$ are proper distributions. Let us introduce some notation to formalize this idea. Define:

$$\mathcal{S} := \{\boldsymbol{\mu} : \mathcal{A} \circ \boldsymbol{\mu} = \mathbf{F}_{(Y\mathbb{1}_D)|Z}\}$$

and

$$\mathcal{T} := \{\boldsymbol{\mu} : \boldsymbol{\mu}(y)_{dg}/P(G_i = g) \text{ is a proper CDF for each } d \text{ and } g \text{ for which } P(G_i = g) > 0\}$$

Given that Θ_c is identified, we will thus have point identification of Δ_c only if

$$\{\Theta_c \cdot \boldsymbol{\mu}\}_{\boldsymbol{\mu} \in (\mathcal{S} \cap \mathcal{T})} \text{ is a singleton for all } \mathcal{P}_{latent} \text{ compatible with } \mathcal{P}_{DZ}$$

We first remark that from this we can see that identification implies that $E[c(a.t., Z_i)] = E[c(n.t., Z_i)] = 0$. To see this, consider any $\boldsymbol{\mu}$ of the form:

$$\boldsymbol{\mu}(y)_{dg} = \begin{cases} F_a(y) & \text{if } \mathbb{1}(g = a.t. \text{ and } d = 0) \\ F_n(y) & \text{if } \mathbb{1}(g = n.t. \text{ and } d = 1) \\ F^P_{Y(d)|G}(y|g) & \text{otherwise} \end{cases}$$

where $F^P_{Y(d)|G}(y|g)$ are the true CDFs of conditional potential outcome distributions, and F_a and F_b are any proper CDF functions. Note then that $\boldsymbol{\mu}(y) \in \mathcal{T}$. But we also have that $\boldsymbol{\mu}(y) \in \mathcal{S}$, since

$$\mathcal{A} \circ \boldsymbol{\mu}(y) = \mathcal{A} \circ \mathbf{F}^P_{Y(D)|G} = \mathbf{F}_{(Y\mathbb{1}_D)|Z}$$

since $A_{1z,n.t} = \mathbb{1}(\mathcal{D}_{n.t.}(z) = 1) = 0$ for all z, and similarly $A_{0z,a.t} = \mathbb{1}(\mathcal{D}_{a.t.}(z) = 0) = 0$ for all z. Intuitively, the map $\mathcal{A}$ does not depend on the combinations $(a.t, 0)$ and $(n.t, 1)$ which are never observed. Thus means that $\mathcal{S} \cap \mathcal{T}$ contains a whole family of functions that differ in arbitrary ways within these two components, subject to F_n and F_a being CDFs. On the other hand,the map Θ_c generally can depend on these components, if the function c is non-zero for always- or never-takers. $\{\Theta_c \cdot \boldsymbol{\mu}\}_{\boldsymbol{\mu} \in (\mathcal{S}\cap\mathcal{T})}$ cannot be a singleton unless $\Theta \circ \boldsymbol{\mu}$ does not depend on the scalar functions $\boldsymbol{\mu}_{0,a.t.}(y)$ or $\boldsymbol{\mu}_{1,n.t.}(y)$, which from Equation (3) occurs only if $E[c(a.t., Z_i)] = E[c(n.t., Z_i)] = 0$ given the distribution $\mathcal{P}_Z$.

Now we construct an element of $\mathcal{S}$ that will enable us to show that Δ_c takes a certain useful form. Consider the vector-valued function $\mathbf{F}^{P*}_{Y(D)|G}(y)$, where:

$$\left[\mathbf{F}^{P*}_{Y(D)|G}(y)\right]_{d,g} := \begin{cases} F^P_{Y(0)|G}(y|a.t.) & \text{if } \mathbb{1}(g = a.t. \text{ and } d = 0) \\ F^P_{Y(1)|G}(y|n.t.) & \text{if } \mathbb{1}(g = n.t. \text{ and } d = 1) \\ \sum_z [(A^d)^\dagger]_{gz} F_{(Y\mathbb{1}_D)|Z}(y, d|z) & \text{otherwise} \end{cases}$$

and $(A^d)^\dagger$ indicates the Moore-Penrose pseudoinverse of the matrix A^d. In this case, again only the "otherwise" terms contribute under the map $\mathcal{A}$ and:

$$\begin{aligned} [\mathcal{A} \circ \mathbf{F}^{P*}_{Y(D)|G}(y)]_{dz} &= \sum_{g,z'} A_{dz,g}[(A^d)^\dagger]_{g,z'} F_{(Y\mathbb{1}_D)|Z}(y, d|z') \\ &= \sum_{g,z'} A^d_{z,g}[(A^d)^\dagger]_{g,z'} F_{(Y\mathbb{1}_D)|Z}(y, d|z') \\ &= \sum_{z'} [A^d(A^d)^\dagger]_{z,z'} F_{(Y\mathbb{1}_D)|Z}(y, d|z') \\ &= [F_{(Y\mathbb{1}_D)|Z}(y)]_{dz} \end{aligned}$$

where the final equality follows from $A^d(A^d)^\dagger = \mathbb{1}_{|\mathcal{Z}|}$, which in turn follows from $(A^d)^\dagger = A^{d'}(A^d A^{d'})^{-1}$ since A^d has full row rank. This confirms that $\mathbf{F}^{P*}_{Y(d)|G} \in \mathcal{S}$. We can verify that A^d has full row rank by observing that if it didn't, then no submatrix B of A^d could have rank $|\mathcal{Z}| = 2^J$. Consider A^1 and the $2^J \times 2^J$ submatrix B formed by taking all rows, and the columns corresponding to the simple compliance groups as well as to the always-takers. Up to an ordering of the entries, this yields an upper triangular matrix of ones, which has full rank of 2^J. Similarly, for A^0 take the same submatrix B but replace the column corresponding to the always-takers with the one corresponding to the never-takers. With the same ordering this submatrix of A^0 is lower triangular matrix of ones which also has full rank of 2^J. Thus, both A^d have full row rank.

By contrast, it will generally not be the case that $\mathbf{F}^{P*}_{Y(D)|G} \in \mathcal{T}$, that is that it yields a proper CDF for each d and g such that $P(G_i = g) > 0$. The problem is ensuring monotonicity. Substituting in Eq. (2), we can rewrite $\mathbf{F}^{P*}_{Y(D)|G}$ as:

$$\left[\mathbf{F}^{P*}_{Y(D)|G}(y)\right]_{d,g} := \begin{cases} F^P_{Y(0)|G}(y|a.t.) & \text{if } \mathbb{1}(g = a.t. \text{ and } d = 0) \\ F^P_{Y(1)|G}(y|n.t.) & \text{if } \mathbb{1}(g = n.t. \text{ and } d = 1) \\ \sum_{g'}[(A^d)^\dagger A^d]_{gg'} F^P_{Y(d)|G}(y|g') & \text{otherwise} \end{cases}$$

Note that $(A^d)^\dagger A^d = A^{d\prime}(A^d A^{d\prime})^{-1}A^d$ is the orthogonal projector onto the rows of A_d, which are indexed by the instrument values z. That $\left[\mathbf{F}^{P*}_{Y(D)|G}(y)\right]_{d,g}$ is weakly increasing in y is not obvious, as the matrix $(A^d)^\dagger A^d$ contains some negative entries for $J > 1$. However, the other properties of a CDF hold, which we now verify. Right-continuity of $\left[\mathbf{F}^{P*}_{Y(D)|G}(y)\right]_{d,g}$ is inherited from $\mathbf{F}^{P}_{Y(D)|G}$. Now we turn to normalization. Note that the vector of values $\{\mathbb{1}(g \neq n.t.)\}_{g\in\mathcal{G}}$ is row of A^1 – and hence an eigenvector of $A^{1\prime}(A^1A^{1\prime})^{-1}A^1$ with eigenvalue one – since there exists a value of z (the one in which all instruments take a value of one) such that all other compliance groups take treatment. Similarly $\{\mathbb{1}(g \neq a.t.)\}_{g\in\mathcal{G}}$ is an eigenvector of $A^{0\prime}(A^0A^{0\prime})^{-1}A^0$ with an eigenvalue of one. Combining with the above expression for $\mathbf{F}^{P*}_{Y(D)|G}$, this implies that $\lim_{y\uparrow y_{max}} \left[\mathbf{F}^{P*}_{Y(D)|G}(y)\right]_{d,g} = 1$, as all the CDF's F^P tend to one, for each (d, g), where y_{max} is the supremum of $\mathcal{Y}$. Also, $\lim_{y\downarrow y_{min}} \left[\mathbf{F}^{P*}_{Y(D)|G}(y)\right]_{d,g} = 0$ as all CDF's F^P tend to zero, where y_{min} is the infimum of $\mathcal{Y}$.

We can use the above results to construct a "hybrid" of $\mathbf{F}^{P*}_{Y(D)|G}$ and $\mathbf{F}^{P}_{Y(D)|G}$ that restores monotonicity and is in fact in $\mathcal{T}$. Consider $\mathbf{F}^{P\alpha}_{Y(D)|G} := \alpha\mathbf{F}^{P*}_{Y(D)|G} + (1-\alpha)\mathbf{F}^{P}_{Y(D)|G}$ for some $\alpha \in \mathbb{R}$. By linearity of $\mathcal{A}$, $\mathbf{F}^{P\alpha}_{Y(D)|G} \in \mathcal{S}$. That each $\left[\mathbf{F}^{P\alpha}_{Y(D)|G}(y)\right]_{d,g}$ is right continuous and tends to zero (one) as y approaches y_{min} (y_{max}) follows from the respective properties of both $\alpha\mathbf{F}^{P*}_{Y(D)|G}$ and $\alpha\mathbf{F}^{P}_{Y(D)|G}$. To ensure that $\left[\mathbf{F}^{P\alpha}_{Y(D)|G}(y)\right]_{d,g}$ is weakly increasing, we only need to choose α to be small enough, so as to not destroy the monotonicity of $\mathbf{F}^{P}_{Y(D)|G}$. The regularity condition in ensures that this is possible for a strictly positive $\bar{\alpha}$. In particular, let $\bar{\alpha} = \min_{d\in\{0,1\}} \inf_{y\in\mathcal{Y}} \min_{g\in\mathcal{G}:P(G_i=g)>0} \left\{ \frac{P(G_i=g)\cdot f_{dg}(y)}{\sum_{g'} |[I-(A^d)^\dagger A^d]_{gg'}| P(G_i=g')\cdot f_{dg'}(y)} \right\}$ where $f_{dg}(y) := P(Y(d) = y|G_i = g)$ if Y has a discrete distribution, and $f_{dg}(y) := \frac{d}{dy}\left(F_{Y(d)|G}(y|g)\right)$ if Y is continuously distributed. This choice of $\bar{\alpha}$ guarantees that $\mathbf{F}^{P\bar{\alpha}}_{Y(D)|G}$ is weakly increasing. Let δ be either a) the differential opterator in the case of continuous Y_i; or b) the finite difference between consecutive values in $\mathcal{Y}$ in the case of a discrete distribution. Then we have by linearity that:

$$
\begin{aligned}
[\delta \circ \mathbf{F}^{P\bar{\alpha}}_{Y(D)|G}(y)]_{dg} &= \delta \circ F^{P*}_{Y(d)|G}(y|g) + \bar{\alpha} \cdot \delta \circ \left(F^{P}_{Y(d)|G}(y|g) - F^{P*}_{Y(d)|G}(y|g)\right) \\
&= f_{dg}(y) + \bar{\alpha}\left(\sum_{g'} [I - (A^d)^\dagger A^d]_{gg'} P(G_i = g') \cdot f_{g'}(y)\right) \\
&\geq f_{dg}(y) - \bar{\alpha}\left(\sum_{g'} \left|[I - (A^d)^\dagger A^d]_{gg'}\right| P(G_i = g') \cdot f_{g'}(y)\right) \geq 0
\end{aligned}
$$

where the final inequality follows from the definition of $\bar{\alpha}$. Finally, $\bar{\alpha} > 0$ so long as for each $d \in \{0, 1\}$ the support of $Y_i(d)$ conditional on $G_i = g$ is independent of all $g \in \mathcal{G}^c$ such that $P(G_i = g) > 0$, and that $f_{dg}(y)$ is separated from zero and upper-bounded uniformly over y. Thus, we have verified that $\mathbf{F}^{P\bar{\alpha}}_{Y(D)|G} \in \mathcal{T}$, as are any $\mathbf{F}^{P\alpha}_{Y(D)|G}$ such that $0 \leq \alpha \leq \bar{\alpha}$.

To proof concludes by using $\mathbf{F}^{P\alpha}_{Y(D)|G}$ to show that Δ_c cannot be identified unless it is equal to $\Delta^*_c = \Theta_c \circ \mathbf{F}^{P*}_{Y(D)|G}$. Identification implies that $\{\Theta_c \cdot \boldsymbol{\mu}\}_{\boldsymbol{\mu} \in (\mathcal{S} \cap \mathcal{T})}$ is a singleton. Suppose that $\Delta_c \neq \Delta^*_c$. Then $\Theta \circ \boldsymbol{D} \neq 0$, where $\boldsymbol{D} = \left(\mathbf{F}^{P*}_{Y(D)|G} - \mathbf{F}^{P}_{Y(D)|G}\right)$. But this would also imply that $\Theta_c \circ \mathbf{F}^{P\alpha}_{Y(D)|G} \neq \Delta_c$ by linearity of Θ_c for all $\alpha \neq 0$, since $\mathbf{F}^{P\alpha}_{Y(D)|G} = \mathbf{F}^{P}_{Y(D)|G} + \alpha \boldsymbol{D}$. Since $\mathbf{F}^{P\alpha}_{Y(D)|G} \in (\mathcal{S} \cap \mathcal{T})$ for all $\alpha < \bar{\alpha}$, this contradicts identification of Δ_c. Thus if Δ_c is point identified, it must be equal to $\Delta^*_c = \Theta_c \circ \mathbf{F}^{P*}_{Y(D)|G}$. We can work this out to be, given the definition of $\mathbf{F}^{P*}_{Y(D)|G}$:

$$
\begin{aligned}
\Delta_c = \sum_{g \in \mathcal{G}^c} \frac{E(c(g, Z_i)|G_i = g)}{E[c(G_i, Z_i)]} &\left(\int y d\left\{\sum_z [(A^1)^\dagger]_{g,z} F_{(Y\mathbb{1}_D)|Z}(y, 1|z))\right\}\right. \\
&\left. - \int y d\left\{\sum_z [(A^0)^\dagger]_{g,z} F_{(Y\mathbb{1}_D)|Z}(y, 0|z))\right\}\right) \\
= \sum_g \frac{E(c(g, Z_i)|G_i = g)}{E[c(G_i, Z_i)]} &\sum_z [(A^1)^\dagger]_{g,z} P(D_i = 1|Z_i = z) \int y dF_{Y|Z,D}(y|z, 1) \\
- \sum_g \frac{E(c(g, Z_i)|G_i = g)}{E[c(G_i, Z_i)]} &\sum_z [(A^0)^\dagger]_{g,z} P(D_i = 0|Z_i = z) \int y dF_{Y|Z,D}(y|z, 0) \\
= \sum_z \left(\sum_g [(A^1)^\dagger]_{g,z} \frac{E(c(g, Z_i)|G_i = g)}{E[c(G_i, Z_i)]} P(D_i = 1|Z_i = z)\right) &E[Y_i|Z_i = z, D_i = 1] \\
- \sum_z \left(\sum_g [(A^0)^\dagger]_{g,z} \frac{E(c(g, Z_i)|G_i = g)}{E[c(G_i, Z_i)]} P(D_i = 0|Z_i = z)\right) &E[Y_i|Z_i = z, D_i = 0]
\end{aligned}
$$

i.e., a linear combination of CEFs of the form $E[Y_i|Z_i = z, D_i = d]$, where the coefficients are all known functions of $\mathcal{P}_{DZ}$. This confirms item 4 of Proposition 11, which implies Property M through Proposition 7 when identification holds for all $\mathcal{P}_{DZ}$ satisfying Assumptions 3 and $P(C_i = 1) > 0$, as the proposition assumes.

E.3 Proof of Proposition 5

First, note that for any $z \in \mathcal{Z}$, since $Y_i = Y_i(0) + D_i(Y_i(1) - Y_i(0))$ we have that:

$$\begin{aligned} E[Y_i|Z_i = z, G_i = g] &= E[Y_i(0)|Z_i = z, G_i = g] + E[D_i(z)(Y_i(1) - Y_i(0))|Z_i = z, G_i = g] \\ &= E[Y_i(0)|G_i = g] + \mathcal{D}_g(z) \cdot E[Y_i(1) - Y_i(0)|G_i = g] \end{aligned}$$

using independence, and thus for any $z, w \in \mathcal{Z}$:

$$E[Y_i|Z_i = z, G_i = g] - E[Y_i|Z_i = w, G_i = g] = (\mathcal{D}_g(z) - \mathcal{D}_g(w)) \cdot E[(Y_i(1) - Y_i(0))|G_i = g]$$

Now:

$$\begin{aligned} &Cov(Y_i, h(Z_i)) \\ &= E[Cov(Y_i, h(Z_i)|G_i)] \\ &= \sum_g P(G_i = g) Cov(Y_i, h(Z_i)|G_i = g) \\ &= \sum_g P(G_i = g) \sum_z h(z) Cov(Y_i, \mathbb{1}(Z_i = z)|G_i = g) \\ &= \sum_g P(G_i = g) \sum_z h(z) \left(E[Y_i \mathbb{1}(Z_i = z)|G_i = g] - E[Y_i|G_i = g] P(Z_i = z|G_i = g) \right) \\ &= \sum_g P(G_i = g) \sum_z h(z) \pi_z \sum_w \pi_w \left(E[Y_i|Z_i = z, G_i = g] - E[Y_i|Z_i = w, G_i = g] \right) \\ &= \sum_g P(G_i = g) \left\{ \sum_{z,w} h(z) \pi_z \pi_w \left(\mathcal{D}_g(z) - \mathcal{D}_g(w) \right) \right\} \Delta_g \end{aligned}$$

where we have used $Cov(A, B) = E[Cov(A, B|C)] + Cov(E[A|C], E[B|C])$ in the first step. Furthermore

$$\begin{aligned} \sum_{z,w} h(z) \pi_z \pi_w \left(\mathcal{D}_g(z) - \mathcal{D}_g(w) \right) &= \sum_z h(z) \pi_z \mathcal{D}_g(z) - \left(\sum_z h(z) \pi_z \right) \left(\sum_w \pi_w \mathcal{D}_g(w) \right) \\ &= Cov(\mathcal{D}_g(Z_i), h(Z_i)) \end{aligned}$$

An analagous sequence of steps shows that the denominator

$$Cov(D_i, h(Z_i)) = \sum_g P(G_i = g) Cov(\mathcal{D}_g(Z_i), h(Z_i))$$

E.4 Proof of Theorem SM1

We start with the $J = 2$ case to build the intuition, and present the generalization afterwards. Simple algebra shows that the 2SLS estimand can be written

$$\rho_{2sls,lin} = \frac{\pi_1 Cov(Y_i, Z_{1i}) + \pi_2 Cov(Y_i, Z_{2i})}{\pi_1 Cov(D_i, Z_{1i}) + \pi_2 Cov(D_i, Z_{2i})}$$

where π_1 and π_2 are the population regression coefficients from the first-stage regression of D on Z_1 and Z_2.

As we've already shown:

$$E[Y_i|Z_{1i}=1]-E[Y_i|Z_{1i}=0]=E[D_i(1,Z_{2i})(Y_i(1)-Y_i(0))|Z_{1i}=1]$$
$$-E[D_i(0,Z_{2i})(Y_i(1)-Y_i(0))|Z_{1i}=0]$$

By separable monotonicity, we can divide all units into 4 groups: always-takers (a.t.), never-takers (n.t.), compliers for the first instrument (Z_1), and compliers for the second instrument (Z_2). Applying the law of total probability to the above expression, only the two complier groups contribute, since $D_i(1,Z_{2i})=D_i(0,Z_{2i})=0$ for the never-takers and

$$E[Y_i(1)-Y_i(0)|Z_{1i}=1,a.t.]=E[Y_i(1)-Y_i(0)|Z_{1i}=0,a.t.]=E[Y_i(1)-Y_i(0)|a.t.]$$

by the independence assumption. Thus we have:

$$\begin{aligned}E[Y|Z_1=1]-E[Y_i|Z_1=0]&=p_{Z_1}E[Y(1)-Y(0)|Z_1=1,Z_1]\\&\quad+p_{Z_2}\left(E[Z_2|Z_1=1,G=Z_2]-E[Z_2|Z_1=0,G=Z_2]\right)E[Y(1)-Y(0)|G=Z_2]\\&=p_{Z_1}E[Y(1)-Y(0)|G=Z_1]+p_{Z_2}\frac{Cov(Z_1,Z_2)}{Var(Z_1)}E[Y(1)-Y(0)|G=Z_2]\end{aligned}\tag{4}$$

where we've used the independence assumption. The same steps lead to an analagous expression for Z_2. Now consider the regression coefficient π_1. It is:

$$\begin{aligned}\pi_1&=\tfrac{1}{Var(Z_1)(1-\rho_{12}^2)}\left[Cov(D,Z_1)-\tfrac{Cov(Z_1,Z_2)}{Var(Z_2)}Cov(D,Z_2)\right]\\&=\tfrac{1}{1-\rho_{12}^2}\left[\tfrac{Cov(D,Z_1)}{Var(Z_1)}-\tfrac{Cov(Z_1,Z_2)}{Var(Z_1)}\cdot\tfrac{Cov(D,Z_2)}{Var(Z_2)}\right]\end{aligned}$$

where ρ_{12} is the Pearson correlation coefficient between Z_1 and Z_2, and we've simplified $Cov(Z_1,Z_1)-\frac{Cov(Z_1,Z_2)}{Var(Z_2)}Z_2)$ to $Var(Z_1)(1-\rho_{12}^2)$. By the same steps as those leading to Eq. (4):

$$\frac{Cov(D,Z_1)}{Var(Z_1)}=E[D|Z_1=1]-E[D|Z_1=0]=p_{Z_1}+p_{Z_2}\frac{Cov(Z_{1i},Z_{2i})}{Var(Z_{1i})}$$

and

$$\frac{Cov(D,Z_2)}{Var(Z_2)}=E[D_i|Z_2=1]-E[D|Z_2=0]=p_{Z_2}+p_{Z_1}\frac{Cov(Z_{1i},Z_{2i})}{Var(Z_{2i})}$$

Thus

$$\begin{aligned}\pi_1&=\frac{1}{1-\rho_{12}^2}\left[p_{Z_1}+\cancel{p_{Z_2}\frac{Cov(Z_1,Z_2)}{Var(Z_1)}}-\frac{Cov(Z_1,Z_2)}{Var(Z_1)}\left(\cancel{p_{Z_2}}+p_{Z_1}\frac{Cov(Z_1,Z_2)}{Var(Z_2)}\right)\right]\\&=p_{Z_1}\frac{1-\rho_{12}^2}{1-\rho_{12}^2}=p_{Z_1}\end{aligned}$$

and similarly $\pi_2=p_{Z_2}$. In other words, under separable monotonicity, the linear regression control in 2SLS is sufficient to isolate the compliers for each instrument (we shall see that this property also holds for $J>2$).

The 2SLS estimator can now be written, using Equation (4):

$$
\begin{aligned}
\rho_{2sls,lin} &= \frac{p_{Z_1}Cov(Y,Z_1)+p_{Z_2}Cov(Y,Z_2)}{p_{Z_1}Cov(D,Z_1)+p_{Z_2}Cov(D,Z_2)} \\
&= \frac{p_{Z_1}\left(p_{Z_1}+p_{Z_2}\frac{Cov(Z_1,Z_2)}{Var(Z_1)}\right)}{p_{Z_1}Cov(D,Z_1)+p_{Z_2}Cov(D,Z_2)} \cdot E[Y(1)-Y(0)|G=Z_1] \\
&\quad + \frac{p_{Z_2}\left(p_{Z_2}+p_{Z_1}\frac{Cov(Z_1,Z_2)}{Var(Z_2)}\right)}{p_{Z_1}Cov(D,Z_1)+p_{Z_2}Cov(D,Z_2)} \cdot E[Y(1)-Y(0)|G=Z_2] \\
&= \frac{p_{Z_1}Cov(D,Z_1)}{p_{Z_1}Cov(D,Z_1)+p_{Z_2}Cov(D,Z_2)} \cdot E[Y(1)-Y(0)|G=Z_1] \\
&\quad + \frac{p_{Z_2}Cov(D,Z_2)}{p_{Z_1}Cov(D,Z_1)+p_{Z_2}Cov(D,Z_2)} \cdot E[Y(1)-Y(0)|G=Z_2]
\end{aligned}
$$

Since $Cov(D,Z_j)\geq 0$ by Assumption 2*, the weights are positive.

To show the $J>2$ case, note that we now have $J+2$ disjoint compliance groups: always-takers, never-takers and compliers for each instrument 1 to J. We use the notation Cj_i to indicate the event that $D_i(1,z_{-j})>D_i(0,z_{-j})$ for all z_{-j} and hence $D_i(1,Z_{-ji})>D(0,Z_{-ji})$. Equation (4) now generalizes, by the law of iterated expections, to:

$$
E[Y|Z_j=1]-E[Y|Z_j=0]=\sum_k p_{Ck}\left(E[Z_k|Z_j=1]-E[Z_k|Z_j=0]\right)E[Y(1)-Y(0)|Ck]
$$

where I've suppressed i indices. Similarly, we have that

$$
E[D|Z_j=1]-E[D|Z_j=0]=\sum_k p_{Ck}\left(E[Z_k|Z_j=1]-E[Z_k|Z_j=0]\right) \tag{5}
$$

This latter expression gives us the property that the multiple regression coefficient $\pi_j = p_{Cj}$ for all j. The reason is that the vector of regression coefficients π is the unique vector satisfying $\Sigma\pi=C$, where Σ is the $J\times J$ covariance matrix of the instruments and C is a vector of covariances between the treatment D and each instrument Z_j. This can be rewritten as:

$$
\sum_k Cov(Z_k,Z_j)\pi_k = Cov(D,Z_j)
$$

Substituting in the guess that $\pi_k=p_{Ck}$ yields Equation (5). The 2SLS estimand is:

$$
\begin{aligned}
\rho_{2sls,lin} &= \frac{\sum_j \pi_j Cov(Y,Z_j)}{\sum_j \pi_j Cov(D,Z_j)} = \frac{\sum_j p_{Cj}\sum_k p_{Ck}Cov(Z_j,Z_k)E[Y(1)-Y(0)|Ck]}{\sum_j p_{Cj}Cov(D,Z_j)} \\
&= \frac{\sum_k p_{Ck}\left(\sum_j p_{Cj}Cov(Z_j,Z_k)\right)E[Y(1)-Y(0)|Ck]}{\sum_j p_{Cj}Cov(D,Z_j)} \\
&= \frac{\sum_k p_{Ck}Cov(D,Z_k)E[Y(1)-Y(0)|Ck]}{\sum_j p_{Cj}Cov(D,Z_j)}
\end{aligned}
$$

where we've used that $Cov(Y,Z_j)=\sum_k p_{Ck}Cov(Z_j,Z_k)E[Y(1)-Y(0)|Ck]$ and $Cov(D,Z_j)=\sum_k p_{Ck}Cov(Z_j,Z_k)$. $Cov(D,Z_k)$ is positive for all k by Assumption 2*.

E.5 Proof of Lemma 1

Fix any $j \in \{1 \dots J\}$.

$$
\begin{aligned}
E[Y_i|Z_{ji}=1]-&E[Y_i|Z_{ji}=0] \\
&= E[D_i(1,Z_{ji})(Y_i(1)-Y_i(0))|Z_{ji}=1] - E[D_i(0,Z_{2i})(Y_i(1)-Y_i(0))|Z_{ji}=0] \\
&= E[(D_i(1,Z_{2i}) - D_i(0,Z_{-j,i}))(Y_i(1)-Y_i(0))] \\
&= P(D_i(1,Z_{-j,i}) > D_i(0,Z_{-j,i}))E[Y_i(1)-Y_i(0)|D_i(1,Z_{-j,i}) > D_i(0,Z_{-j,i})]
\end{aligned}
$$

where we've used $Y_i = Y_i(0) + D_i(Y_i(1) - Y_i(0))$ and Assumption 1 in the first step, and Assumption 5 in the second. Similarly $E[D_i|Z_{ji}=1] - E[D_i|Z_{ji}=0] = P(D_i(1,Z_{-j,i}) > D_i(0,Z_{-j,i}))$ and thus

$$\rho_j = E[Y_i(1)-Y_i(0)|D_i(1,Z_{-j,i}) > D_i(0,Z_{-j,i})]$$

E.6 Proof of Theorem SM2

Note that the 2SLS estimand can be written:

$$\rho_{2sls,lin} = \frac{\sum_j \pi_j Cov(Y,Z_j)}{\sum_j \pi_j Cov(D,Z_j)} = \frac{\sum_j \pi_j Cov(D,Z_j)}{\sum_j \pi_j Cov(D,Z_j)}\rho_j$$

Given Lemma 1, it only remains to be shown that $\pi_j \geq 0$ for all j. As a vector:

$$\pi = \Sigma^{-1}C$$

where Σ is the $J \times J$ covariance matrix of the instruments and C is a vector of covariances between the treatment D and each instrument Z_j. A result known as Farkas' Lemma (see e.g. Gale et al. 1951) states the following: for matrix $A \in \mathbb{R}^{m\times n}$ and vector $b \in \mathbb{R}^m$, exactly one of the following is true:

1. There exists an $x \in \mathbb{R}^n$ such that $Ax = b$ and $x \geq 0$
2. There exists a $y \in \mathbb{R}^m$ such that $A'y \geq 0$ and $b'y < 0$

where for any vector, the notation ≥ 0 indicates that each of its components is weakly positive, etc. Since Σ is an invertible, square, symmetric matrix, Farkas' Lemma is in our case equivalent to:

$$\pi \geq 0 \iff \big(\forall y \in \mathbb{R}^J : \Sigma y \geq 0 \implies C'y \geq 0\big)$$

Now note that element j of the vector Σy is equal to $Cov(Z_j, Z'y)$ and $C'y$ is equal to $Cov(Z'y, D)$ where $Z'y = \sum_{k=1}^J y_k Z_k$ is a linear combination of the instruments. Thus, what we wish to show is that for any $y \in \mathbb{R}^m$:

$$E[Z'y|Z_j=1] \geq E[Z'y|Z_j=0] \quad \forall j \implies E[Z'y|D=1] \geq E[Z'y|D=0]$$

In fact, given the strength of Assumption 5, the left-hand inequality holding for any single j will be sufficient. By the law of iterated expectations:

$$
\begin{aligned}
& E[Z'y|D=1] - E[Z'y|D=0] \\
& = \sum_{z \in \{0,1\}} P(Z_j = z|D=1) E[Z'y|D(z, Z_{-j}) = 1, Z_j = z] \\
& \qquad\qquad\qquad\qquad - P(Z_j = z|D=0) E[Z'y|D(z, Z_{-j}) = 0, Z_j = z] \\
& = P(Z_j = 1|D=1) \{E[Z'y|D(1, Z_{-j}) = 1, Z_j = 1] - E[Z'y|D(0, Z_{-j}) = 1, Z_j = 0]\} \\
& - P(Z_j = 1|D=0) \{E[Z'y|D(1, Z_{-j}) = 0, Z_j = 1] - E[Z'y|D(0, Z_{-j}) = 0, Z_j = 0]\} \\
& + E[Z'y|D(0, Z_{-j}) = 1, Z_j = 0] - E[Z'y|D(0, Z_{-j}) = 0, Z_j = 0]
\end{aligned}
$$

By Assumptions 2 and 1, for any $z, d \in \{0, 1\}$:

$$
\begin{aligned}
E[Z'y|D(z, Z_{-j}) = d, Z_j = z] &= \sum_{z_{-j} \in \{0,1\}^{L-1}} (z, z_{-j})'y \cdot P(Z_{-j} = z_{-j}|D(z, z_{-j}) = d, Z_j = z) \\
&= \sum_{z_{-j} \in \{0,1\}^{L-1}} (z, z_{-j})'y \cdot P(Z_{-j} = z_{-j}|Z_j = z) \\
&= E[Z'y|Z_j = z]
\end{aligned}
$$

and thus $E[Z'y|D = 1] - E[Z'y|D = 0]$ can be simplified to

$$
(E[Z_j|D=1] - E[Z_j|D=0]) (E[Z'y|Z_j = 1] - E[Z'y|Z_j = 0])
$$

which is positive whenever $E[Z'y|Z_j = 1] - E[Z'y|Z_j = 0]$ is positive, since $Cov(D, Z_j) \geq 0$ by Assumption 2*. While we did not make Assumption 2* in the statement of this theorem, it is implied by Assumptions 2 and 5, since:

$$
\begin{aligned}
Cov(D, Z_j) &= P(Z_j)(1 - P(Z_j)) (E[D|Z_j = 1] - E[D|Z_j = 0]) \\
&= P(Z_j)(1 - P(Z_j)) (E[D(1, Z_{-j})|Z_j = 1] - E[D(0, Z_{-j})|Z_j = 0]) \\
&= P(Z_j)(1 - P(Z_j)) E[D(1, Z_{-j}) - D(0, Z_{-j})] & =\geq 0
\end{aligned}
$$

where the third equality follows by Assumption 1 (independence) and the final inequality follows by Assumption 2 and the law of iterated expectations (over Z_{-j}).

References

ANGRIST, B. J. D. and EVANS, W. N. (1998). “Children and Their Parents ’ Labor Supply: Evidence from Exogenous Variation in Family Size”. *American Economic Review* 88 (3), pp. 450–477.

ANGRIST, J. D. (2006). “Instrumental variables methods in experimental criminological research: What, why and how”. *Journal of Experimental Criminology* 2 (1), pp. 23–44.

ANGRIST, J. D. and PISCHKE, J.-S. (2008). *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press.

GALE, D., KUHN, H. and TUCKER, A. (1951). "Linear Programming And The Theory Of Games". *Activity Analysis of Production and Allocation*. Cowles Commission for Research in Economics Monograph No. 13, pp. 314–421.

ICHIMURA, H. and THOMPSON, T. S. (1998). "Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution". *Journal of Econometrics* 86 (2), pp. 269–295.

IMBENS, G. W. and ANGRIST, J. D. (1994). "Identification and Estimation of Local Average Treatment Effects". *Econometrica* 62 (2), pp. 467–475.

MOGSTAD, M., TORGOVITSKY, A. and WALTERS, C. (2020). "Policy Evaluation with Multiple Instrumental Variables". *Working Paper*. URL: https://www.nber.org/papers/w27546.pdf.

MOUNTJOY, J. (2019). "Community Colleges and Upward Mobility". *Working Paper*, pp. 1–83. URL: https://ssrn.com/abstract=3373801.