# Online Appendix for "When does IV identification not restrict outcomes?"

Leonard Goff

Last updated: January 6, 2026

## Contents

# A Defining outcome-nonrestrictive IV identification

## A.1 Notation

Let $\mathcal{P}$ denote the joint distribution of the model fundamentals $(G_i, \tilde{Y}_i, Z_i)$. Given Eq. (2), we can decompose $\mathcal{P}$ as

$$\mathcal{P} = \mathcal{P}_{latent} \times \mathcal{P}_Z,$$

where $\mathcal{P}_Z$ denotes the distribution of the instruments $Z_i$ and $\mathcal{P}_{latent}$ denotes the distribution of the latent variables of the model $\tilde{Y}$ and $G$.[1]

A generic causal parameter of interest $\theta$ is a functional $\theta(\mathcal{P})$ of the distribution $\mathcal{P}$ of model variables. Let $\mathcal{P}_{obs}$ denote the distribution of observable variables $(Y_i, T_i, Z_i)$. Note that $\mathcal{P}_Z$ is a marginalization of $\mathcal{P}_{obs}$ over $Y_i$ and $T_i$. I make use of the following notational convention: for a sub-vector $W_0$ of a random vector $W$, let $\mathcal{P}_{W_0}(\mathcal{P}_W)$ be the

[1]By $\mathcal{P} = \mathcal{P}_{latent} \times \mathcal{P}_Z$, I mean that for any Borel set $\mathcal{B}_L$ of values for $(G_i, \tilde{Y}_i)$ and $\mathcal{B}_Z$ of values for $\mathcal{P}_Z$ we have $\mathcal{P}(\mathcal{B}_L \times \mathcal{B}_Z) = \mathcal{P}_{latent}(\mathcal{B}_L) \cdot \mathcal{P}_Z(\mathcal{B}_Z)$, where $\mathcal{B}_L \times \mathcal{B}_Z$ is the Cartesian product of $\mathcal{B}_L$ and $\mathcal{B}_Z$.

distribution of $W_0$ that arises after marginalizing distribution $\mathcal{P}_W$ over the components of $W$ not included in $W_0$. In this notation, for example, $\mathcal{P}_Z = \mathcal{P}_Z(\mathcal{P}_{obs})$.

Define $\mathscr{P}_{latent}(\mathcal{G})$ to be the set of $\mathcal{P}_{latent}$ compatible with a given selection model $\mathcal{G}$ and admitting of finite moments:

$$\mathscr{P}_{latent}(\mathcal{G}) := \{\mathcal{P}_{latent} \in \mathscr{P}_{\tilde{Y}G} : supp(\mathcal{P}_{\mathcal{G}}(\mathcal{P}_{latent})) \subseteq \mathcal{G}\} \tag{1}$$

where we let $\mathscr{P}_{\tilde{Y}G}$ denote the set of all distributions over $(\tilde{Y}_i, G_i)$, such that $\mathbb{E}[Y_i(t)|G_i = g]$ exists and is finite for each $t \in \mathcal{T}$ and $g \in \mathcal{G}$. Employing a similar notation, we let $\mathscr{P}_Z$ be the set of distributions over instrument values that embed any maintained support restrictions (e.g. that $Z_i$ is binary with $P(Z_i = 1) \in (0, 1)$).

Note that for any $\mathcal{P} = \mathcal{P}_{latent} \times \mathcal{P}_Z$, Eq. (1) and $T_i = T_i(Z_i)$ imply a distribution of observables. Let $\phi$ denote this map so that $\mathcal{P}_{obs} = \phi(\mathcal{P})$. The set of possible distributions of observables given a selection model $\mathcal{G}$ is

$$\mathscr{P}_{obs}(\mathcal{G}) := \{\phi(\mathcal{P}_{latent} \times \mathcal{P}_Z) : \mathcal{P}_{latent} \in \mathscr{P}_{latent}(\mathcal{G}), \mathcal{P}_Z \in \mathscr{P}_Z\}$$

All together, we can think of the basic IV model as the set of distributions $M = \{\mathcal{P}_{latent} \times \mathcal{P}_Z : \mathcal{P}_{latent} \in \mathscr{P}_{latent}(\mathcal{G}), \mathcal{P}_Z \in \mathscr{P}_Z\}$. In this notation note that $\mathscr{P}_{obs}(\mathcal{G}) = \phi(M)$.

### A.2 Observable restrictions implied by the model

In general, $\mathscr{P}_{obs}(\mathcal{G})$ is a strict subset of the set of all joint distributions of $(Y_i, T_i, Z_i)$, i.e. restrictions on $\mathcal{G}$ coupled with Eq. (2) imply testable implications on $\mathcal{P}_{obs}$. These testable implications have been studied in the case of the classic LATE model (see e.g. Kitagawa 2015; Mourifié and Wan 2017; Kédagni and Mourifié 2020, see also Jiang and Sun 2023). Such restrictions are discussed further in Section A.5.

### A.3 Outcome nonrestrictive IV identification

Given a function $c(\cdot)$ introduced in Section 2, denote the subset of $\mathscr{P}_{latent}(\mathcal{G})$ for which $P(c(G_i) = 1) > 0$ given the distribution $\mathcal{P}_G$ of $G_i$ as:

$$\mathscr{P}_{latent,c}(\mathcal{G}) := \{\mathcal{P}_{latent} \in \mathscr{P}_{latent}(\mathcal{G}) \text{ and } P(c(G_i) = 1) > 0 \text{ according to } \mathcal{P}_G(\mathcal{P}_{latent})\} \tag{2}$$

Similarly, let $\mathscr{P}_{obs,c}(\mathcal{G}) := \{\phi(\mathcal{P}_{latent} \times \mathcal{P}_Z) : \mathcal{P}_{latent} \in \mathscr{P}_{latent,c}(\mathcal{G}), \mathcal{P}_Z \in \mathscr{P}_Z\}$. $\mathscr{P}_{obs,c}(\mathcal{G})$ consist of the distributions of observables that respect selection model $\mathcal{G}$ and put positive probability on the groups $g \in \mathcal{G}$ such that $c(g) = 1$. These sets are used in defining outcome-nonrestrictive identification as a simple guarantee that the target parameters $\mu_c^t$ and $\Delta_c^{t,t'}$ are well-defined. But causal parameters that condition on a probability-zero event—such as the marginal treatment effect—can also be accommodated in this framework, as limiting cases of a sequence of parameters for $c_j$ satisfying $P(c_j(G_i) =$

1) > 0 (see Appendix G).

We are now ready to give a definition of outcome-nonrestrictive identification, where the target parameter $\theta$ is expressed as a function $\theta = \theta(\mathcal{P})$ of the data generating process $\mathcal{P}$:

**Definition 1.** *Given a choice model $\mathcal{G}$, we say that parameter $\theta$ with conditioning function $c$ is* ***outcome-nonrestrictive*** *identified under $\mathcal{G}$ if the set*

$$\{\theta(\mathcal{P}) : \phi(\mathcal{P}) = \mathcal{P}_{obs} \text{ and } \mathcal{P} = (\mathcal{P}_{latent} \times \mathcal{P}_Z) \text{ for some } \mathcal{P}_{latent} \in \mathscr{P}_{latent,c}(\mathcal{G}) \text{ and } \mathcal{P}_Z \in \mathscr{P}_Z\}$$

*is a singleton for all $\mathcal{P}_{obs} \in \mathscr{P}_{obs,c}(\mathcal{G})$.*

Point identification in general says there is a unique value $\theta(\mathcal{P})$ compatible with Eq. (2) and $\phi(\mathcal{P})$, for all $\mathcal{P}$ in some set defined by the model. The key requirement that identification be *outcome-nonrestrictive* is that this model is broad enough to include *all* of $\mathscr{P}_{latent,c}(\mathcal{G})$.[2] The set $\mathscr{P}_{latent,c}(\mathcal{G})$ allows what Heckman et al. (2006) call *essential heterogeneity*. The only restrictions on outcomes amount to IV independence (imposed by taking the product measure $\mathcal{P} = \mathcal{P}_{latent} \times \mathcal{P}_Z$), exclusion (implicit in the notation $Y_i(t)$), and finite group-specific means of $\tilde{Y}_i$ (imposed through $\mathcal{P}_{\tilde{Y}G}$ in (1)).[3] Thus $\mathscr{P}_{latent,c}(\mathcal{G})$ is compatible with any marginal distribution $\mathcal{P}_{\tilde{Y}}$ of $\tilde{Y} = \{Y_i(t)\}_{t\in\mathcal{T}}$ or selection-type conditioned distributions $\mathcal{P}_{\tilde{Y}|G=g}$ across various $g \in \mathcal{G}$ whatsoever (provided that they have finite means), so there is no assumption that e.g. treatment effects are homogeneous across units, or are unrelated to counterfactual selection behavior $G_i$.

### A.4 Binary combinations and binary collections

We begin by establishing a terminology to refer to situations in which the identification result for counterfactual means in Eq. (4) can be applied.

**Definition.** Given selection model $\mathcal{G}$, a *binary combination* is a treatment value $t \in \mathcal{T}$ and a function $\alpha : \mathcal{Z} \to \mathbb{R}$ of finite support $\mathcal{Z}_K = \{z_k\}_{k=1}^K$ such that $\sum_{k=1}^K \alpha(z_k) \cdot D_i^{[t]}(z_k) \in \{0, 1\}$ for all $i$, according to $\mathcal{G}$.

Now consider a collection of binary combinations that apply to at least two distinct values $t \in \mathcal{T}$. Let us denote set of coefficients $\alpha$ in each binary combination by $\alpha^{[t]}$, indexed by the treatment value $t$ it will be applied to. In this notation, $\alpha_k^{[t]}$ is the coefficient on $z_k$ in the binary combination corresponding to treatment $t$.

[2] Definition 1 represents a case of point identification as defined in Lewbel (2019) (see also Matzkin 2007), where the known information ($\phi$ in Lewbel's notation) is the distribution $\mathcal{P}_{obs}$, the model value ($m \in M$ in Lewbel's notation) is $\mathcal{P} = \mathcal{P}_{latent} \times \mathcal{P}_Z$, and the model $M$ is the Cartesian product of $\mathscr{P}_{latent,c}(\mathcal{G})$ and $\mathscr{P}_Z$.

[3] $\mathscr{P}_{latent,c}(\mathcal{G})$ does restrict the marginal distributions of $G_i$ and $Z_i$: through $\mathcal{G}$, $P(c(G_i) = 1) > 0$, and $\mathscr{P}_Z$.

**Definition.** A *binary collection* is a set of binary combinations $\{(t, \alpha^{[t]})\}_{t\in\psi}$ for treatment values in set $\psi \subseteq \mathcal{T}$ where $|\psi| \geq 2$, with the property that given the selection model $\mathcal{G}$, the functions $c^{[t,\alpha^{[t]}]}$ and $c^{[t',\alpha^{[t']}]}$ are identical, for any $t, t' \in \psi$.

For a given binary collection, let us for brevity denote the common function $c^{[t,\alpha^{[t]}]}$ for all $t \in \psi$ as $c$. It follows immediately from Theorem 1 that treatment effects $\mathbb{E}[Y_i(t') - Y_i(t)|c(G_i) = 1] = \mathbb{E}[Y_i(t')|c(G_i) = 1] - \mathbb{E}[Y_i(t)|c(G_i) = 1]$ are identified for any pair $t, t' \in \psi$.

When treatment is itself binary, we can generate binary collections from any binary combination where the coefficients sum to zero:

**Proposition A.1.** *Let $\mathcal{T} = \{0, 1\}$, and suppose $(t, \alpha)$ is a binary combination such that $\sum_k \alpha_k = 0$. Then there exists a binary collection with $\psi = \mathcal{T}$. In particular, the coefficients for $t = 0$ are simply $-1$ times the corresponding coefficients for $t = 1$.*

*Proof.* See alternative statement of this result in Section 3. □

The restriction that $\sum_k \alpha_k = 0$ is a natural one, in the following sense:

**Proposition A.2.** *Let $\Delta_c^{t,t'} = \mathbb{E}[Y_i(t') - Y_i(t)|c(G_i) = 1]$ be outcome-nonrestrictive identified from a binary collection with $t' \neq t$. Then if $\mathcal{G}$ contains a group $g_0$ that always takes treatment $t$, it must be the the case that $\sum_k \alpha_k^{[t]} = 0$.*

*Proof.* Since $P(T_i = t'|G_i = g_0) = 0$, the data provide no information on $Y(t')|G_i = g_0$, so we must have $c(g_0) = 0$ (see proof of Proposition 1). Thus $c(g_0) = \sum_k \alpha_k^{[t]} \cdot \mathbb{1}(T_{g_0}(z_k) = t) = \sum_k \alpha_k^{[t]} = 0$. □

For example, in the LATE model of Imbens and Angrist (1994), allowing for "always-takers" (who always take treatment $t = 1$, regardless of $Z_i$) implies that $\sum_z \alpha_z^{[1]} = 0$, while allowing for "never-takers" (who always take treatment $t = 0$) implies that $\sum_z \alpha_z^{[0]} = 0$. Consistent with this, identification of the compliers LATE follows from the binary collection in which $\alpha_1^{[1]} = 1$, $\alpha_0^{[1]} = -1$, $\alpha_1^{[0]} = -1$, and $\alpha_0^{[0]} = 1$.

### A.5 Using binary combinations and collections for testing the model

The existence of binary combinations with $K > 1$ generally yields overidentification restrictions that can used to test the IV model (including exclusion, independence, and the choice of selection model $\mathcal{G}$). In particular, suppose that $|\mathcal{G}| < \infty$ and note that for any Borel set $\mathcal{B}$ of $\mathbb{R}$ and binary combination $(t, \alpha)$, we have that:

$$\sum_{k=1}^{K} \alpha_k \cdot P(Y_i \in \mathcal{B}, T_i = t|Z_i = z_k) = P(Y_i(t) \in \mathcal{B}, c(G_i) = 1) \tag{3}$$

using Eq. (2) and that $P(Y_i(t) \in \mathcal{B}, T_i(z_k) = t) = \sum_{g \in \mathcal{G}} P(G_i = g) \cdot P(Y_i(t) \in \mathcal{B} | G_i = g) \cdot A^{[t]}_{z_k, g}$. Since the RHS of Eq. (3) represents a probability, the LHS must be weakly positive. Provided that not all of the $\alpha_k$ are positive, the implication that $\sum_{k=1}^{K} \alpha_k \cdot P(Y_i \in \mathcal{B}, T_i = t | Z_i = z_k) \geq 0$ is not guaranteed and therefore can be used to test the model assumptions.

Furthermore, finding binary *collections* may yield further overidentification restrictions that make use of the "first stage" data alone. Depending on the selection model, the equality $\sum_{k=1}^{K} \alpha_k^{[t]} \cdot \mathbb{E}\left[D_i^{[t]} | Z_i = z_k\right] = \sum_{k=1}^{K} \alpha_k^{[t']} \cdot \mathbb{E}\left[D_i^{[t']} | Z_i = z_k\right]$ may not be trivially satisfied, even in the case of a binary treatment. See Section 5 for an example of such equality restrictions in the context of an empirical application, and Appendix J.6 for further linear inequality constraints that are based upon first stage empirical moments. Still further testable restrictions hold if one has a binary collection and Eq. (2) holds conditional on observed covariates $X_i$. See Appendix J.2 for details.

# B Proofs

## B.1 Proof of Proposition 1

To ease notation, write $\Delta_c^{t,t'}$ as $\Delta$, $\mu_c^t$ as $\mu(t)$, and $\mu_c^{t'}$ as $\mu(t')$, with $c$ fixed. It is apparent that if $\mu(t')$ and $\mu(t)$ are outcome-nonrestrictive identified, then $\Delta = \mu(t') - \mu(t)$ is too.

Now let us consider the other direction. Suppose that $\mu(t)$ is not outcome-nonrestrictive identified (an analogous argument holds if $\mu(t')$ is not outcome-nonrestrictive identified). Then for some $\mathcal{P}_{obs} \in \mathscr{P}_{obs,c}(\mathcal{G})$, the set $\{\theta_{\mu(t)}(\mathcal{P}) : \mathcal{P} \in M \text{ and } \phi(\mathcal{P}) = \mathcal{P}_{obs}\}$ has at least two elements, where $M := \{\mathcal{P}_{latent} \times \mathcal{P}_Z : \mathcal{P}_{latent} \in \mathscr{P}_{latent,c}(\mathcal{G}), \mathcal{P}_Z \in \mathscr{P}_Z\}$ and we let $\theta_{\mu(t)}(\cdot)$ be the map that yields the value of $\mu(t)$ as a function of $\mathcal{P}$.[4] Accordingly, let $\mathcal{P}_1, \mathcal{P}_2 \in M$ where $\theta_{\mu(t)}(\mathcal{P}_1) = a$ and $\theta_{\mu(t)}(\mathcal{P}_2) = b$ where $a \neq b$ despite $\phi(\mathcal{P}_1) = \phi(\mathcal{P}_2) = \mathcal{P}_{obs}$.

Let us decompose $\mathcal{P}_1$ as $\left(\left\{\mathcal{P}_{Y(s)|G=g}\right\}_{\substack{s \in \mathcal{T} \\ g \in \mathcal{G}}}, \mathcal{P}_G, \mathcal{P}_Z\right)$, which is possible because $\mathcal{P}_1$ satisfies independence Eq. (2) between the instruments and the latent variables. Let $\mathcal{P}(0)$ denote a degenerate distribution at zero in $\mathbb{R}$. Now consider the distribution $\tilde{\mathcal{P}}_1 = \left(\{\mathcal{P}(0)\}_{g \in \mathcal{G}}, \left\{\mathcal{P}_{Y(t)|G=g}\right\}_{\substack{s \in \mathcal{T}, s \neq t' \\ g \in \mathcal{G}}}, \mathcal{P}_G, \mathcal{P}_Z\right)$. That is, $Y_i(t') = 0$ with probability one according to $\tilde{\mathcal{P}}_1$, but the joint distribution of $Z_i$, $G_i$ and all of the other potential outcomes $s \neq t'$ are the same under $\tilde{\mathcal{P}}_1$ as they are under $\mathcal{P}_1$. Note that given this construction: $\theta_{\mu(t)}(\tilde{\mathcal{P}}_1) = \theta_{\mu(t)}(\mathcal{P}_1) = a$, since $\mu(t)$ only depends on the distributions $\mathcal{P}_{Y(t)|G=g}$ and $\mathcal{P}_G$, and $t \neq t'$. Note as well that from $\mathcal{P}_1 \in M$ we know that $\mathcal{P}_{latent}(\mathcal{P}_1) \in \mathscr{P}_{latent,c}(\mathcal{G})$. Since $\mathcal{P}_G$ has not been changed in defining $\tilde{\mathcal{P}}_1$ from $\mathcal{P}_1$, and a degenerate random variable at zero has a finite expectation, it follows that $\mathcal{P}_{latent}(\tilde{\mathcal{P}}_1) \in \mathscr{P}_{latent,c}(\mathcal{G})$ as well. Since $\mathcal{P}_Z$ has also not been changed, we further have that $\tilde{\mathcal{P}}_1 \in M$.

[4] Note that the set $M$ will vary with $c$, but since we are considering a fixed $c$ this is left implicit to ease notation.

Define $\tilde{\mathcal{P}}_2$ analogously from $\mathcal{P}_2$, and observe that similarly $\theta_{\mu(t)}(\tilde{\mathcal{P}}_2) = \theta_{\mu(t)}(\mathcal{P}_2) = b$ and again that $\tilde{\mathcal{P}}_2 \in M$.

Observe furthermore that $\theta_\Delta(\tilde{\mathcal{P}}_1) = \theta_{\mu(t')}(\tilde{\mathcal{P}}_1) - \theta_{\mu(t)}(\tilde{\mathcal{P}}_1) = 0 - a = -a$, and similarly $\theta_\Delta(\tilde{\mathcal{P}}_2) = \theta_{\mu(t')}(\tilde{\mathcal{P}}_2) - \theta_{\mu(t)}(\tilde{\mathcal{P}}_2) = 0 - b = -b$. Thus since since $b \neq a$:

$$\theta_\Delta(\tilde{\mathcal{P}}_1) \neq \theta_\Delta(\tilde{\mathcal{P}}_2) \tag{4}$$

I now show that this contradicts $\Delta$ being outcome-nonrestrictive identified.

To see this, decompose $\mathcal{P}_{obs}$ as $\left(\left\{\mathcal{P}_{Y|T=s,Z=z}\right\}_{\substack{s\in\mathcal{T}\\z\in\mathcal{Z}}}, \left\{\mathcal{P}_{T|Z=z}\right\}_{z\in\mathcal{Z}}, \mathcal{P}_Z\right)$ and define $\tilde{\mathcal{P}}_{obs} = \left(\{\mathcal{P}(0)\}_{z\in\mathcal{Z}}, \left\{\mathcal{P}_{Y|T=s,Z=z}\right\}_{\substack{s\in\mathcal{T},s\neq t'\\z\in\mathcal{Z}}}, \left\{\mathcal{P}_{T|Z=z}\right\}_{z\in\mathcal{Z}}, \mathcal{P}_Z\right)$ where the $\{\mathcal{P}(0)\}_{z\in\mathcal{Z}}$ indicate that $P(Y_i = 0|T_i = t', Z_i = z) = 1$ for all $z \in \mathcal{Z}$ according to $\tilde{\mathcal{P}}_{obs}$. That is, the marginal distribution $\mathcal{P}_{TZ}$ and the conditional distributions $\mathcal{P}_{Y|T=s,Z=z}$ for all $s \neq t'$ and $z$ are unchanged from $\mathcal{P}_{obs}$, but $Y_i = 0$ with probability one conditional on $T_i = t'$.

The next step is to observe that $\phi(\tilde{\mathcal{P}}_1) = \tilde{\mathcal{P}}_{obs}$ and $\phi(\tilde{\mathcal{P}}_2) = \tilde{\mathcal{P}}_{obs}$. To see this, note that $Y_i(t') = 0$ with probability one implies that $Y_i = 0$ with probability one conditional on $T_i = t'$ (provided that $P(T_i = t) > 0$). Now since $\mathcal{P}_1$ and $\tilde{\mathcal{P}}_1$ only differ in $\mathcal{P}_{Y(t')|G=g}$ (leaving $\mathcal{P}_{TZ}$ and and $\mathcal{P}_{Y|T_i=s,Z=z}$ for all $s \neq t'$ and $z$ unchanged), it follows from $\phi(\mathcal{P}_1) = \mathcal{P}_{obs}$ that $\phi(\tilde{\mathcal{P}}_1) = \tilde{\mathcal{P}}_{obs}$, and analogously for $\tilde{\mathcal{P}}_2$. This further implies that $\tilde{\mathcal{P}}_{obs} \in \mathscr{P}_{obs,c}(\mathcal{G})$.

Since $\Delta$ is outcome-nonrestrictive identified and $\tilde{\mathcal{P}}_{obs} \in \mathscr{P}_{obs,c}(\mathcal{G})$, the set $\{\theta_\Delta(\mathcal{P}) : \mathcal{P} \in M \text{ and } \phi(\mathcal{P}) = \tilde{\mathcal{P}}_{obs}\}$ must be a singleton. Given that $\phi(\tilde{\mathcal{P}}_1) = \phi(\tilde{\mathcal{P}}_2) = \tilde{\mathcal{P}}_{obs}$ and $\tilde{\mathcal{P}}_1, \tilde{\mathcal{P}}_2 \in M$ we must then have $\theta_\Delta(\tilde{\mathcal{P}}_1) = \theta_\Delta(\tilde{\mathcal{P}}_2)$. This yields a contradiction with (4).

We can generalize Proposition 1 as follows. For any vector of coefficients $\rho_t$ for each $t \in \mathcal{T}$, define $\theta_c^\rho := \sum_t \rho_t \cdot \mu_c^t$. $\mu_c^t$ is a special case of $\theta_\rho^c$ in which $\rho_t$ is equal to one for a single treatment, and zero for all others. Similarly, $\Delta_c^{t,t'}$ is a case of $\theta_c^\rho$ in which $\rho_{t'} = 1$, $\rho_t = -1$, and all other components of $\rho$ are equal to zero. In Section 5, the local average complimentarity parameter $\lambda_c = \mu_c^C - \mu_c^A - \mu_c^B + \mu_c^0$ is an example of $\theta_c^\rho$ where $\rho_C = \rho_0 = 1$ and $\rho_A = \rho_B = -1$.

In general, let $\psi(\rho) \subseteq \mathcal{T}$ be the set of treatments for which $\rho_t \neq 0$. Clearly $\theta_c^\rho$ is outcome-nonrestrictive identified if $\mu_c^t$ is for each $t \in \psi(\rho)$. The above argument articulated for treatment effects extends immediately to show that $\theta_c^\rho$ is also outcome-nonrestrictive identified *only* if $\mu_c^t$ is for each $t \in \psi(\rho)$. To see this, we again begin with a value $t \in \psi(\rho)$ such that $\mu(t)$ is not outcome-nonrestrictive identified, i.e. $\theta_{\mu(t)}(\mathcal{P}_1) = a$ and $\theta_{\mu(t)}(\mathcal{P}_2) = b$ with $a \neq b$, where $\mathcal{P}_1$ and $\mathcal{P}_2$ are the corresponding latent variable distributions in $M$ such that $\phi(\mathcal{P}_1) = \phi(\mathcal{P}_2) = \mathcal{P}_{obs}$. Let $d_1 = \sum_{s\neq t} \rho_s \cdot \theta_{\mu(s)}(\mathcal{P}_1)$ and $d_2 = \sum_{s\neq t} \rho_s \cdot \theta_{\mu(s)}(\mathcal{P}_2)$ such that $\theta_c^\rho = \rho_t \cdot a + d_1$ under $\mathcal{P}_1$ and $\theta_c^\rho = \rho_t \cdot b + d_2$ under $\mathcal{P}_2$.

Suppose that $\theta_c^\rho$ is outcome-nonrestrictive identified. In this case, we must have that $d_2 = d_1 + \rho_t \cdot (a - b)$. Now consider the distributions $\tilde{P}_1$, $\tilde{P}_2$ and $\tilde{P}_{obs}$ defined above, where we take $t' \neq t$ to be any other treatment in $\psi(\rho)$ other than $t$. We have already seen above that $\tilde{\mathcal{P}}_{obs} \in \mathscr{P}_{obs,c}(\mathcal{G})$, $\tilde{\mathcal{P}}_1, \tilde{\mathcal{P}}_2 \in M$ and $\phi(\tilde{\mathcal{P}}_1) = \phi(\tilde{\mathcal{P}}_2) = \tilde{\mathcal{P}}_{obs}$. Thus we must have

that $\theta_c^\rho$ is the same under both $\tilde{\mathcal{P}}_1$ and $\tilde{\mathcal{P}}_2$. Instead, we have that under $\tilde{\mathcal{P}}_1$, $\theta_c^\rho$ is equal to $\rho_t \cdot a + d_1 - \rho_{t'} \cdot \theta_{\mu(t')}(\mathcal{P}_1)$, and under $\tilde{\mathcal{P}}_1$, $\theta_c^\rho$ is equal to

$$\rho_t \cdot b + d_2 - \rho_{t'} \cdot \theta_{\mu(t')}(\mathcal{P}_2) = \rho_t \cdot b + d_1 + \rho_t \cdot (a-b) - \rho_{t'} \cdot \theta_{\mu(t')}(\mathcal{P}_2) = \{\rho_t \cdot a + d_1\} - \rho_{t'} \cdot \theta_{\mu(t')}(\mathcal{P}_2)$$

Thus we must have that $\theta_{\mu(t')}(\mathcal{P}_2) = \theta_{\mu(t')}(\mathcal{P}_1)$. This argument can be repeated for every $t' \in \psi(\rho)$, $t' \neq t$, and we then have that $d_1 = d_2$. This in turn implies that $\rho_t \cdot (a-b) = 0$, which contradicts $a \neq b$ with $\rho_t \neq 0$. We have thus arrived at a contradiction.

### B.2 Proof of Theorem 2

#### Setup and notation

Let $\mathcal{Y} \subseteq \mathbb{R}$ be the support of $Y$. For any $y \in \mathcal{Y}, z \in \mathcal{Z}$ and $t \in \mathcal{T}$, define $F_{(YD)|Z=z}(y,t) := \mathbb{E}[\mathbb{1}(Y_i \leq y)\mathbb{1}(T_i = t)|Z_i = z]$. This function acts like a CDF for $Y_i$ and a probability mass function for $T_i$, conditional on $Z_i = z$. We begin with the observation that knowing the distribution $\mathcal{P}_{obs}$ of $(Y_i, T_i, Z_i)$ is equivalent to knowing the value of $F_{(YD)|Z=z}(y,t)$ for all $(y,t,z)$ along with the observable distribution of the instruments $\mathcal{P}_Z$.

By the law of iterated expectations over $G_i$ and using independence (2):

$$\begin{aligned} F_{(YD)|Z=z}(y,t) &= \mathbb{E}\left\{\mathbb{E}[\mathbb{1}(Y_i(t) \leq y)\mathbb{1}(T_i(z) = t)|Z_i = z, G_i]\right\} \\ &= \sum_{g: A_{zg}^{[t]}=1} P(G_i = g) \cdot \mathbb{E}[\mathbb{1}(Y_i(t) \leq y)|G_i = g] \\ &= \sum_{g: A_{zg}^{[t]}=1} P(G_i = g) \cdot F_{Y(t)|G=g}(y) := \sum_{g \in \mathcal{G}} A_{zg}^{[t]} \cdot P(G_i = g) \cdot F_{Y(t)|G=g}(y) \end{aligned} \tag{5}$$

I use the following Lemma to assume that $A^{[t]}$ has full row rank, without loss of generality:

**Lemma 1.** *If $\mu_g^t$ is outcome-nonrestrictive identified given instrument support $\mathcal{Z}$, it remains outcome-nonrestrictive identified using data from $Z_i \in \mathcal{Z}_0$, where $\mathcal{Z}_0 \subseteq \mathcal{Z}$ is a subset of instrument values for which the rows of $A^{[t]}$ for $z \in \mathcal{Z}_0$ are linearly independent of one another.*

A special case of Lemma 1 is an observation by Heckman and Pinto (2018) that one can remove any rows of $A^{[t]}$ that is an exact copy of another row (i.e. there are two instrument values for which all response types behave the same regarding whether they choose treatment $t$ or not), and there is hence a direct redundancy over instrument values.

#### Outcome-nonrestrictive identification

Now define $\mathbf{F}_{(YD)|Z}(y)$ to be a $|\mathcal{T}| \cdot |\mathcal{Z}| \times 1$ vector of $F_{(YD)|Z=z}(y,t)$ over $z$ and $t$ and $\mathbf{G}^*(y)$ to be the unknown $|\mathcal{T}| \cdot |\mathcal{G}|$-component vector of $P(G_i = g) \cdot F_{Y(t)|G=g}(y)$ over $g$ and $t$, for a fixed $y$. Now let $\mathbf{G}^*$ represent the whole vector-valued function $\mathbf{G}^* : \mathcal{Y} \to \mathbb{R}^{|\mathcal{T}| \cdot |\mathcal{G}|}$, and define $\mathbf{F}_{(YD)|Z}$ similarly as the function $\mathcal{Y} \to \mathbb{R}^{|\mathcal{T}| \cdot |\mathcal{Z}|}$ yielding the vector $\mathbf{F}_{(YD)|Z}(y)$. Note

that $\mathcal{P}_Z$ and $\mathbf{F}_{(YD)|Z}$ encode the entire distribution $\mathcal{P}_{obs}$ of observables $(Y, T, Z)$ while $\mathcal{P}_Z$ and $\mathbf{G}^*$ encode the entire distribution $\mathcal{P}$ of model primitives $(\tilde{Y}, G, Z)$.

The relationship between the two can be characterized by writing Eq. (5) as:

$$\mathbf{F}_{(YD)|Z} = \mathcal{A} \circ \mathbf{G}^* \tag{6}$$

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

$$[\mathcal{A} \circ \boldsymbol{\mu}(y)]_{tz} = \sum_g A^{[t]}_{z,g} \cdot \boldsymbol{\mu}(y)_{tg}$$

holding for each $y$, for any vector-valued function $\boldsymbol{\mu} : \mathcal{Y} \to \mathbb{R}^{|\mathcal{T}|\cdot|\mathcal{G}|}$.

Let $\theta = \mathbb{E}[Y_i(t)|c(G_i) = 1]$ be the parameter of interest. Note that similar to (6), $\theta$ can also be written as a linear map applied to the function $\mathbf{G}^*$. In particular $\theta = \Theta \circ \mathbf{G}^*$, where for any function $\boldsymbol{\mu} : \mathcal{Y}$ to $\mathbb{R}^{|\mathcal{T}|\cdot|\mathcal{G}|}$, $\Theta \circ \boldsymbol{\mu}$ is the scalar:

$$\sum_{g \in \mathcal{G}} \frac{c_g}{P(c(G_i) = 1)} \cdot \int_{\mathcal{Y}} y \cdot d\boldsymbol{\mu}(y)_{t,g} \tag{7}$$

The set of such $\boldsymbol{\mu}$ that recover the distribution of observables can be written as:

$$\mathcal{S} := \{\boldsymbol{\mu} : \mathcal{A} \circ \boldsymbol{\mu} = \mathbf{F}_{(YD)|Z}\}$$

However, some such candidate values $\boldsymbol{\mu} \in \mathcal{S}$ for $\mathbf{G}^*$ may correspond to $F_{Y(t)|G=g}(\cdot)$ that do not represent valid CDFs. Accordingly, let us define

$$\mathcal{R} := \{\boldsymbol{\mu} : [\boldsymbol{\mu}(y)]_{tg}/P(G_i = g) \text{ is a proper CDF for each } t \in \mathcal{T} \text{ and } g \in \mathcal{G} \text{ s.t. } P(G_i = g) > 0\}$$

The remainder of this section establishes that for $\theta$ to be outcome-nonrestrictive identified, the set $\mathcal{S} \cap \mathcal{R}$ must map to a singleton under $\Theta$.

Note that the sets $\mathcal{R}$ and $\mathcal{S}$ as well as the map $\Theta$ depend on the distribution $\mathcal{P}_{latent}$ (through $\mathbf{F}_{(YD)|Z}$ for $\mathcal{S}$ and through the $P(G_i = g)$ for $\mathcal{R}$ and $\Theta$).[5] Let us denote this dependence by $\mathcal{S}(\mathcal{P}_{latent})$, $\mathcal{R}(\mathcal{P}_{latent})$ and $\Theta(\mathcal{P}_{latent})$, though I will later leave this dependence implicit to ease notation.

Definition 1 of outcome-nonrestrictive identification, translated into this notation, says that

$$\{\Theta(\mathcal{P}_{latent}) \circ \boldsymbol{\mu} : \boldsymbol{\mu} \in \mathcal{R}(\mathcal{P}_{latent}) \text{ and } \boldsymbol{\mu} \in \mathcal{S}(\mathcal{P}_{latent})\} \text{ is a singleton } \forall \mathcal{P}_{latent} \in \mathscr{P}_{latent,c}(\mathcal{G}) \tag{8}$$

The following regularity condition will prove to be useful later in the proof:

**Condition REG.** *Fix a $t \in \mathcal{T}$. For some $g^* \in \mathcal{G}$, there exists a $\underline{L} > 0$ and $\bar{L} < \infty$ such*

[5] Note that the map $\Theta$ depends on $t$ and the vector $c$ as well, also left implicit for ease of exposition.

*that for any* $g' \in \mathcal{G}$ *and* $y' > y$*:*

$$\underline{L} \leq \frac{F_{Y(t)|G=g'}(y') - F_{Y(t)|G=g'}(y)}{F_{Y(t)|G=g^*}(y') - F_{Y(t)|G=g^*}(y)} \leq \bar{L}$$

Note that whether or not Condition REG holds is a property of $\mathcal{P}_{latent}$. A sufficient condition is that $Y$ is discrete and finite and the support of $Y(t)|G = g$ is the same for all $g$. Another sufficient condition is that i) $Y$ is continuously distributed with the support of the density $f_{Y(t)|G=g}(y)$ the same for all $g$ and $t$; ii) the density on this set $\mathcal{Y}$ is bounded from below by $\underline{M} > 0$ for all $g$, and iii) similarly $\sup_{y\in\mathcal{Y}} f_{Y(t)|G=g}(y) \leq \bar{M}$ for some $\bar{M} < \infty$, for all $g$.[6] A mixture of distributions satisfying the above will also satisfy REG.

Let $\bar{\mathscr{P}}_{latent,c}(\mathcal{G})$ denote the set of distributions $\mathcal{P}_{latent} \in \mathscr{P}_{latent,c}(\mathcal{G})$ that satisfy Condition REG. $\bar{\mathscr{P}}_{latent,c}(\mathcal{G})$ is never empty (given $\mathcal{G} \neq \emptyset$), since we have seen above that for any $|\mathcal{G}| > 0$ there are always distributions that satisfy REG (with examples for each of discrete, continuous or mixed $Y$). Further, $\mathscr{P}_{latent,c}(\mathcal{G})$ only limits the support of $G$ and places no constraint on the distribution of $\tilde{Y}|G$. Note from (8) that if $\theta$ is outcome-nonrestrictive identified, $\{\Theta \cdot \boldsymbol{\mu}\}_{\boldsymbol{\mu}\in(\mathcal{S}(\mathcal{P}_{latent})\cap\mathcal{R}(\mathcal{P}_{latent})}$ must be a singleton for all $\mathcal{P}_{latent}$ such that $supp\{\mathcal{P}_G(\mathcal{P}_{latent})\} \subseteq \mathcal{G}$, including any $\mathcal{P}_{latent} \in \bar{\mathscr{P}}_{latent,c}(\mathcal{G})$.

The remainder of the proof of Theorem 2 shows that if $c \notin rs(A^{[t]})$, it is always possible to find $\mathcal{P}_{latent} \in \bar{\mathscr{P}}_{latent,c}(\mathcal{G})$ such that $\{\Theta(\mathcal{P}_{latent}) \cdot \boldsymbol{\mu}\}_{\boldsymbol{\mu}\in(\mathcal{S}(\mathcal{P}_{latent})\cap\mathcal{R}(\mathcal{P}_{latent})}$ is not in fact a singleton.

**A candidate for $\mathbf{G}^*$ that recovers observables**

To see this, we will explicitly construct a functional $\mathbf{G}$ of $\mathcal{P}_{latent}$, that generally differs from $\mathbf{G}^*$ and lets us define an "alternative" to $\mathcal{P}_{latent}$ but still recovers observables.

Consider the vector-valued function $\mathbf{G}$, where the $t, g$ component of $\mathbf{G}(y)$ is:

$$[\mathbf{G}(y)]_{t,g} := \begin{cases} P(G_i = g) \cdot F_{Y(t)|G}(y|g) & \text{if } \max_{z\in\mathcal{Z}} \mathbb{1}(T_g(z) = t) = 0 \\ \sum_z [(A^{[t]})^+]_{g,z}] \cdot F_{(YD)|Z}(y, t|z) & \text{if } \max_{z\in\mathcal{Z}} \mathbb{1}(T_g(z) = t) = 1 \end{cases}$$

and $(A^{[t]})^+$ indicates the Moore-Penrose pseudoinverse of the matrix $A^{[t]}$.

The reason for separating out the two cases in the definition of $\mathbf{G}$ is that if there exists a group $g$ that acts as a "never-taker" with respect to treatment $t$ such that $\max_{z\in\mathcal{Z}} \mathbb{1}(T_g(z) = t) = 0$, then this corresponds to a column of all zeros in $A^{[t]}$. A property of the Moore-Penrose inverse is that if column $g$ of $A^{[t]}$ is all zeros, then the corresponding row $g$ of $(A^{[t]})^+$ is also all zeros (see e.g. Hung and Markham 1975) which would leave $[\mathbf{G}(y)]_{t,g} = 0$ for all $y$ if we did not separate out this case. This would make it impossible for $\mathbf{G}$ to represent a possible candidate for $\mathbf{G}^*$ (i.e. $\mathbf{G} \in \mathcal{R}$). The above

[6] In the discrete case, let $\underline{L} = \min_{y\in\mathcal{Y}, g\in\mathcal{G}} P(Y(t) = y|G = g)P(Y(t) = y|G = g^*)$ and $\bar{L} = 1/\min_{y\in\mathcal{Y}} P(Y(t) = y|G = g^*)$. In the continuous case let $\bar{L} = \frac{\max_{g\in\mathcal{G}} \sup_{y\in\mathcal{Y}} f_{Y(t)|G=g}(y)}{\min_{g\in\mathcal{G}} \inf_{y\in\mathcal{Y}} f_{Y(t)|G=g}(y)} \leq \bar{M}/\underline{M}$ and $\underline{L} = \frac{\min_{g\in\mathcal{G}} \inf_{y\in\mathcal{Y}} f_{Y(t)|G=g}(y)}{\max_{g\in\mathcal{G}} \sup_{y\in\mathcal{Y}} f_{Y(t)|G=g}(y)} \geq \underline{M}/\bar{M}$.

construction avoids this problem by simply replacing such problematic combinations of $(g,t)$ by using the actual $[\mathbf{G}^*(y)]_{t,g}$ (which are unknown). Note that if the first case holds for *all* $g \in \mathcal{G}$, then the matrix $A^{[t]}$ is simply the zero matrix, and outcome-nonrestrictive identification cannot hold, by Lemma 1. Thus, we can continue under the assumption that the second case holds for at least some $g \in \mathcal{G}$.

Let use see now that $\mathbf{G}$ "recovers observables", by which I mean that $\mathcal{A} \circ \boldsymbol{\mu} = \mathbf{F}_{(YD)|Z}$ and hence $\mathbf{G} \in \mathcal{S}$. Indeed:

$$\begin{aligned}
[\mathcal{A} \circ \mathbf{G}(y)]_{t,z} &= \sum_g A^{[t]}_{z,g} \left[\mathbf{G}(y)\right]_{t,g} \\
&= \cancel{\sum_{g:\max_{z\in\mathcal{Z}} \mathbb{1}(T_g(z)=t)=0} A^{[t]}_{z,g} \cdot P(G_i = g) \cdot F_{Y(t)|G}(y|g)} \\
&\qquad + \sum_{g:\max_{z\in\mathcal{Z}} \mathbb{1}(T_g(z)=t)=1} \sum_{z'} A^{[t]}_{z,g} [(A^{[t]})^+]_{g,z'} F_{(YD)|Z}(y,t|z') \\
&= \sum_{g,z'} A^{[t]}_{z,g} [(A^{[t]})^+]_{g,z'} F_{(YD)|Z}(y,t|z') \\
&= \sum_{z'} [A^{[t]}(A^{[t]})^+]_{z,z'} F_{(YD)|Z}(y,t|z') = [F_{(YD)|Z}(y)]_{tz}
\end{aligned}$$

where the second and third equalities use that $A^{[t]}_{z,g} = 0$ for all $z$, if $g$ is such that $\max_{z\in\mathcal{Z}} \mathbb{1}(T_g(z) = t) = 0$. The final equality follows from $A^{[t]}(A^{[t]})^+ = I_{|\mathcal{Z}|}$, which in turn follows from $(A^{[t]})^+ = A^{[t]\prime}(A^{[t]}A^{[t]\prime})^{-1}$ since we can by Lemma 1 assume that $A^{[t]}$ has full row rank.

$\mathbf{G}$ may still however not be in $\mathcal{R}$, as its definition above does not ensure that each $F_{Y(t)|G}(y|g)$ is necessarily weakly increasing in $y$ with a limit of unity as $y \uparrow \infty$. Note that $[\mathbf{G}]_{t,g}/P(G_i = g)$ does have the final two properties of a CDF: right-continuity and a left limit of zero. To see this, substitute (6) into the definition of $\mathbf{G}$, to rewrite as:

$$[\mathbf{G}(y)]_{t,g} := \begin{cases} P(G_i = g) \cdot F_{Y(t)|G}(y|g) & \text{if } \max_{z\in\mathcal{Z}} \mathbb{1}(T_g(z) = t) = 0 \\ \sum_{g'} [(A^{[t]})^+ A^{[t]}]_{g,g'} \cdot P(G_i = g') \cdot F_{Y(t)|G}(y|g') & \text{if } \max_{z\in\mathcal{Z}} \mathbb{1}(T_g(z) = t) = 1 \end{cases} \tag{9}$$

Right continuity of each element of $\mathbf{G}(y)$ in $y$ follows from right-continuity of the $F_{Y(t)|G}(y|g')$. Note that $\lim_{y\downarrow-\infty} [\mathbf{G}(y)]_{t,g} = 0$ follows from each of the CDFs $F_{(YD)|Z}$ approaching zero as $y \downarrow -\infty$, given that the components of $A^{[t]}$ and $P(G_i = g)$ are finite.

Let $\beta_{t,g} := \lim_{y\uparrow\infty} [\mathbf{G}(y)]_{t,g}$. For any $t, g$ such that $\max_{z\in\mathcal{Z}} \mathbb{1}(T_g(z) = t) = 0$, it follows from the definition of $\mathbf{G}$ that $\beta_{t,g} = P(G_i = g)$, since each of the $F_{Y(t)|G}(y|g)$ are valid CDFs. For the other $t, g$, use (9) to see that

$$\begin{aligned}
\beta_{t,g} &= \lim_{y\uparrow\infty} \sum_{g'} [(A^{[t]})^+ A^{[t]}]_{g,g'} \cdot P(G_i = g') \cdot F_{Y(t)|G}(y|g') = \sum_{g'} [(A^{[t]})^+ A^{[t]}]_{g,g'} \cdot P(G_i = g') \\
&= [(A^{[t]})^+ A^{[t]} P]_g
\end{aligned}$$

where $P$ is a vector of $P(G_i = g)$ for all $g \in \mathcal{G}$.

Unless $[(A^{[t]})^+ A^{[t]} P]_g = P_g$ for all $g \in \mathcal{G}$, the functions $[\mathbf{G}(y)]_{t,g}$ may thus not represent properly normalized CDFs. In fact, they may not even be monotonic in $y$. However, we can still use $\mathbf{G}$ as a building block to construct another set of functions that satisfy all of the properties of a CDF.

**A broader class of candidates that also recover observables but represent CDFs**

Given some fixed $g^* \in \mathcal{G}$, let us define a vector valued function $\mathbf{D} : \mathcal{Y} \to \mathbb{R}^{|\mathcal{T}| \cdot |\mathcal{G}|}$ with components:

$$[\mathbf{D}(y)]_{t,g} := (P_g - \beta_{t,g}) \cdot F_{Y(t)|G}(y|g^*) = [\left\{I - (A^{[t]})^+ A^{[t]}\right\} P]_g \cdot F_{Y(t)|G}(y|g^*) \tag{10}$$

Now let us define for any $\lambda \in [0, 1]$ the convex combination of $\mathbf{G} + \mathbf{D}$ and $\mathbf{G}^*$:

$$\mathbf{G}^\lambda := \lambda \left(\mathbf{G} + \mathbf{D}\right) + (1 - \lambda)\mathbf{G}^* = \mathbf{G}^* + \lambda \left\{\mathbf{G} - \mathbf{G}^* + \mathbf{D}\right\} \tag{11}$$

Our first observation will be that $\mathcal{A} \circ \mathbf{G}^\lambda = \mathbf{F}_{(YD)|Z}$, i.e. $\mathbf{G}^\lambda$ still recovers observables and thus $\mathbf{G}^\lambda \in \mathcal{S}$. To see this, note that:

$$\begin{aligned}
[\mathcal{A} \circ \mathbf{G}^\lambda(y)]_{t,z} &= [\mathcal{A} \circ \mathbf{G}(y)]_{t,z} + \lambda \cdot [\mathcal{A} \circ \left\{\mathbf{G} - \mathbf{G}^* + \mathbf{D}\right\}(y)]_{t,g} \\
&= [F_{(YD)|Z}(y)]_{t,z} + \lambda \cdot \cancel{[\mathcal{A} \circ \mathbf{G}(y)]_{t,g}} - \cancel{\lambda \cdot [\mathcal{A} \circ \mathbf{G}^*(y)]_{t,g}} + \lambda \cdot [\mathcal{A} \circ \mathbf{D}(y)]_{t,g} \\
&= [F_{(YD)|Z}(y)]_{t,z} + \cancel{\lambda \cdot \sum_{g,g'} A^{[t]}_{z,g} \cdot [(I - (A^{[t]})^+ A^{[t]})]_{g,g'} \cdot P(G_i = g') \cdot F_{Y(t)|G}(y|g^*)} \\
&= [F_{(YD)|Z}(y)]_{t,z} + \lambda \cdot \sum_{g'} \cancel{[A^{[t]}(I - (A^{[t]})^+ A^{[t]})]_{z,g'}} \cdot P(G_i = g') \cdot F_{Y(t)|G}(y|g^*) \\
&= [F_{(YD)|Z}(y)]_{t,z}
\end{aligned}$$

since $\mathcal{A} \circ \mathbf{G}^* = \mathcal{A} \circ \mathbf{G}$ and $A^{[t]}(A^{[t]})^+ A^{[t]} = A^{[t]}$.

Now, we verify that for a small enough $\lambda$, $\mathbf{G}^\lambda$ yields $F_{Y(t)|G}(y|g)$ that satisfy the properties of a CDF and hence $\mathbf{G}^\lambda \in \mathcal{R}$. First, note that $\left[\mathbf{G}^\lambda(y)\right]_{t,g}$ is right-continuous in $y$, since each of $[\mathbf{G}(y)]_{t,g}$, $[\mathbf{G}^*(y)]_{t,g}$, and $[\mathbf{D}(y)]_{t,g}$ are. We also have that $\lim_{y \downarrow -\infty} \left[\mathbf{G}^\lambda(y)\right]_{t,g} = 0$, since

$$\lim_{y \downarrow -\infty} [\mathbf{G}(y)]_{t,g} = \lim_{y \downarrow -\infty} [\mathbf{G}^*(y)]_{t,g} = \lim_{y \downarrow -\infty} [\mathbf{D}(y)]_{t,g} = 0$$

Note as well that

$$\begin{aligned}
\lim_{y \uparrow \infty} \left[\mathbf{G}^\lambda(y)\right]_{t,g} &= \lim_{y \uparrow \infty} [\mathbf{G}^*(y)]_{t,g} + \lambda \cdot \lim_{y \uparrow \infty} [\left\{\mathbf{G} - \mathbf{G}^* + \mathbf{D}\right\}(y)]_{t,g} \\
&= P_g + \lambda \cdot \left\{\lim_{y \uparrow \infty} [\mathbf{G}(y)]_{t,g} - \lim_{y \uparrow \infty} [\mathbf{G}^*(y)]_{t,g} + \lim_{y \uparrow \infty} [\mathbf{D}(y)]_{t,g}\right\} \\
&= P_g + \lambda \cdot \{\beta_{t,g} - P_g + (P_g - \beta_{t,g}) \cdot 1\} = P_g
\end{aligned}$$

matching the correct normalization, i.e. $\lim_{y\uparrow\infty}[\mathbf{G}^*(y)]_{t,g} = P_g \cdot \lim_{y\uparrow\infty} F_{Y(t)|G=g}(y) = P_g$.

It only remains to be seen that for a small enough value of $\lambda$, $\left[\mathbf{G}^\lambda(y)\right]_{t,g}$ is weakly increasing in $y$. This is always possible given that $\mathcal{P}_{latent}$ satisfies Condition REG:

**Proposition B.1.** *Given Condition REG,* $\left[\mathbf{G}^\lambda(y)\right]_{t,g}$ *is non-decreasing in* $y$ *for any* $\lambda \in (0, \bar{\lambda}]$*, where* $\bar{\lambda} = \frac{\underline{L}}{2|\mathcal{G}|\cdot\bar{L}} > 0$.

Given Proposition B.1, we have shown that for $\lambda \leq \bar{\lambda}$, $\mathbf{G}^\lambda \in \mathcal{R}$ and hence $\mathbf{G}^\lambda \in (\mathcal{S} \cap \mathcal{R})$.

**Outcome-nonrestrictive identification implies** $c \in rs(A^{[t]})$

Consider now any $\mathcal{P}_{latent} \in \bar{\mathscr{P}}, c(\mathcal{G})$ and choose the $g^* \in \mathcal{G}$ in the definition of $\mathbf{D}$ so that REG holds for that $g^*$. We know that there exist $\lambda > 0$ small enough that $\mathbf{G}^\lambda \in (\mathcal{S} \cap \mathcal{R})$. For any such $\lambda$, outcome-nonrestrictive identification of $\theta$ now requires that $\Theta \circ \mathbf{G}^\lambda = \Theta \circ \mathbf{G}^*$. This in turn requires, by Eq. (11), that $\Theta \circ \{\mathbf{G} - \mathbf{G}^* + \mathbf{D}\} = 0$. Now:

$$
\begin{aligned}
&\Theta \circ \{\mathbf{G} - \mathbf{G}^* + \mathbf{D}\} \\
&= \frac{1}{P(c(G_i)=1)} \sum_g c_g \cdot \left\{ \int_{\mathcal{Y}} y \cdot d\mathbf{G}(y)_{t,g} - \int_{\mathcal{Y}} y \cdot d\mathbf{G}^*(y)_{t,g} + \int_{\mathcal{Y}} y \cdot d\mathbf{D}(y)_{t,g} \right\} \\
&= \frac{1}{P(c(G_i)=1)} \sum_g c_g \sum_{g'} [I - (A^{[t]})^+ A^{[t]}]_{g,g'} \cdot P(G_i = g') \cdot \mathbb{E}[Y_i(t)|G_i = g'] \\
&\qquad + \frac{1}{P(c(G_i)=1)} \sum_g c_g \sum_{g'} [(I - (A^{[t]})^+ A^{[t]})P]_{g,g'} \cdot P(G_i = g') \cdot \mathbb{E}[Y_i(t)|G_i = g^*] \\
&= \frac{1}{P(c(G_i)=1)} \sum_{g'} [c'(I - (A^{[t]})^+ A^{[t]})]_{g'} \cdot P(G_i = g') \cdot \mathbb{E}[Y_i(t)|G_i = g'] \\
&\qquad\qquad + \frac{1}{P(c(G_i)=1)} \sum_{g'} [c'(I - (A^{[t]})^+ A^{[t]})]_{g'} \cdot P(G_i = g') \cdot \mathbb{E}[Y_i(t)|G_i = g^*] \\
&= \frac{1}{P(c(G_i)=1)} \sum_{g'} [c'(I - (A^{[t]})^+ A^{[t]})]_{g'} \cdot P(G_i = g') \cdot \{\mathbb{E}[Y_i(t)|g'] - \mathbb{E}[Y_i(t)|g^*]\}
\end{aligned}
\tag{12}
$$

Note that although the map $\Theta$ depends on the distribution $\mathcal{P}_G$, the constructions $\mathbf{G}$, $\mathbf{D}$ and $\mathbf{G}^\lambda$ all use the same distribution $\mathcal{P}_G$ from the actual distribution $\mathcal{P}_{latent}$. It is for this reason that $P(c(G_i) = 1)$ factors out in Eq. (12), and the RHS can only be non-zero if the sum over $g'$ appearing in it evaluates to zero.

Suppose that $c \notin rs(A^{[t]})$ so that $c'(I - (A^{[t]})^+ A^{[t]}) = \tilde{c}'$ for some non-zero vector $\tilde{c}$. Provided that $P(G_i = g') \cdot \{\mathbb{E}[Y_i(t)|G_i = g'] - \mathbb{E}[Y_i(t)|G_i = g^*]\}$, thought of as a vector across $g' \in \mathcal{G}$, is not perfectly orthogonal in $\mathbb{R}^{|\mathcal{G}|}$ to $\tilde{c}$, we will have that

$$
\sum_{g'} \tilde{c}'_{g'} \cdot P(G_i = g') \cdot \{\mathbb{E}[Y_i(t)|G_i = g'] - \mathbb{E}[Y_i(t)|G_i = g^*]\} \neq 0
$$

There is always a $\mathcal{P}_{latent} \in \bar{\mathscr{P}}_{latent,c}(\mathcal{G})$ such that this non-orthogonality holds, because the

relative magnitudes of $P(G_i = g)$ and level-differences $\mathbb{E}[Y_i(t)|G_i = g'] - \mathbb{E}[Y_i(t)|G_i = g^*]$ in $Y_i(t)$ can be varied without violating REG or changing the support of $G_i$. Thus if $c \notin rs(A^{[t]})$, we can obtain $\Theta \circ \{\mathbf{G} - \mathbf{G}^* + \mathbf{D}\} \neq 0$ for some $\mathcal{P}_{latent} \in \mathscr{P}_{c,latent}(\mathcal{G})$, and $\theta$ is not outcome-nonrestrictive identified.

### B.2.1 Proof of Proposition B.1

The key to ensuring monotonicity will be to choose $\lambda$ small enough that any decreases with $y$ in the components of $\mathbf{G}^\lambda$ are dominated by increases in the corresponding components of $\mathbf{G}^*$, so that each $\left[\mathbf{G}^\lambda\right]_{t,g}$ is monotonically increasing. For $\left[\mathbf{G}^\lambda(y)\right]_{t,g}$ to be monotonically increasing in $y$ we need that for any $y' > y$: $\left[\mathbf{G}^\lambda(y')\right]_{t,g} - \left[\mathbf{G}^\lambda(y)\right]_{t,g} \geq 0$, i.e. that

$$[\mathbf{G}^*(y')]_{t,g} - [\mathbf{G}^*]_{t,g} \geq \lambda \cdot [(\mathbf{G}^* - \mathbf{G})(y') - (\mathbf{G}^* - \mathbf{G})(y)]_{t,g} - ([\mathbf{D}(y')]_{t,g} - [\mathbf{D}]_{t,g})\} \quad (13)$$

Let us turn first to $[(\mathbf{G}^* - \mathbf{G})(y)]_{t,g}$. Fix a $g$ and $t$, and any $y' > y$. Then, by (9):

$$[\mathbf{G}(y')]_{t,g} - [\mathbf{G}^*(y)]_{t,g} = \begin{cases} P(G_i = g) \cdot \{F_{Y(t)|G}(y'|g) - F_{Y(t)|G}(y|g)\} \\ \sum_{g'} [(A^{[t]})^+ A^{[t]}]_{g,g'} \cdot P(G_i = g') \cdot \left\{F_{Y(t)|G}(y'|g') - F_{Y(t)|G}(y|g')\right\} \end{cases} \quad (14)$$

where the first line indicates the case that $g$ is such that $\max_{z \in \mathcal{Z}} \mathbb{1}(T_g(z) = t) = 0$, and the second that $\max_{z \in \mathcal{Z}} \mathbb{1}(T_g(z) = t) = 1$. Thus $[(\mathbf{G}^* - \mathbf{G})(y')]_{t,g} - [(\mathbf{G}^* - \mathbf{G})(y)]_{t,g}$ is equal to 0 if $\max_{z \in \mathcal{Z}} \mathbb{1}(T_g(z) = t) = 0$, and

$$\sum_{g'} [I - (A^{[t]})^+ A^{[t]}]_{g,g'} \cdot P(G_i = g') \cdot \left\{F_{Y(t)|G}(y'|g') - F_{Y(t)|G}(y|g')\right\}$$

if $\max_{z \in \mathcal{Z}} \mathbb{1}(T_g(z) = t) = 1$.

Thus we have by REG that

$$\begin{aligned}
&\left|[(\mathbf{G}^* - \mathbf{G})(y')]_{t,g} - [(\mathbf{G}^* - \mathbf{G})(y)]_{t,g}\right| \\
&= \left|\sum_{g'} [I - (A^{[t]})^+ A^{[t]}]_{g,g'} \cdot P(G_i = g') \cdot \left\{F_{Y(t)|G}(y'|g') - F_{Y(t)|G}(y|g')\right\}\right| \\
&= \left\{F_{Y(t)|G}(y'|g^*) - F_{Y(t)|G}(y|g^*)\right\} \cdot \left|\sum_{g'} [I - (A^{[t]})^+ A^{[t]}]_{g,g'} \cdot P(G_i = g') \cdot \frac{F_{Y(t)|G}(y'|g') - F_{Y(t)|G}(y|g')}{F_{Y(t)|G}(y'|g^*) - F_{Y(t)|G}(y|g^*)}\right| \\
&\leq \left\{F_{Y(t)|G}(y'|g^*) - F_{Y(t)|G}(y|g^*)\right\} \cdot |\mathcal{G}|^{1/2} \cdot \sqrt{\sum_{g'} P(G_i = g')^2 \cdot \left(\frac{F_{Y(t)|G}(y'|g') - F_{Y(t)|G}(y|g')}{F_{Y(t)|G}(y'|g^*) - F_{Y(t)|G}(y|g^*)}\right)^2} \\
&\leq \left\{F_{Y(t)|G}(y'|g^*) - F_{Y(t)|G}(y|g^*)\right\} \cdot |\mathcal{G}| \cdot \max_{g'} P(G_i = g') \cdot \max_{g'} \left|\frac{F_{Y(t)|G}(y'|g') - F_{Y(t)|G}(y|g')}{F_{Y(t)|G}(y'|g^*) - F_{Y(t)|G}(y|g^*)}\right| \\
&\leq \left\{F_{Y(t)|G}(y'|g^*) - F_{Y(t)|G}(y|g^*)\right\} \cdot |\mathcal{G}| \cdot \max_{g'} \left|\frac{F_{Y(t)|G}(y'|g') - F_{Y(t)|G}(y|g')}{F_{Y(t)|G}(y'|g^*) - F_{Y(t)|G}(y|g^*)}\right|
\end{aligned}$$

using that $[I - (A^{[t]})^+ A^{[t]}]$ is a projection (so that $|[I - (A^{[t]})^+ A^{[t]}]v| \leq |v|$ for any vector $v \in \mathbb{R}^{|\mathcal{G}|}$) and by the Cauchy-Schwarz inequality. Let $\delta_t^*(y', y) := F_{Y(t)|G}(y'|g^*) - F_{Y(t)|G}(y|g^*)$.

Then, by REG:

$$\left|[(\mathbf{G}^* - \mathbf{G})(y')]_{t,g} - [(\mathbf{G}^* - \mathbf{G})(y)]_{t,g}\right| \leq \delta_t^*(y', y) \cdot |\mathcal{G}| \cdot \bar{L}$$

Now consider $[\mathbf{D}(y)]_{t,g}$. Fix a $g$ and $t$, and any $y' > y$. Similarly, we have that

$$\begin{aligned}\left|[\mathbf{D}(y')]_{t,g} - [\mathbf{D}(y)]_{t,g}\right| &= \left|\sum_{g'}[I - (A^{[t]})^+ A^{[t]}]_{g,g'} \cdot P(G_i = g') \cdot \{F_{Y(t)|G}(y'|g^*) - F_{Y(t)|G}(y|g^*)\}\right| \\ &\leq \{F_{Y(t)|G}(y'|g^*) - F_{Y(t)|G}(y|g^*)\} \cdot |P| \\ &\leq \{F_{Y(t)|G}(y'|g^*) - F_{Y(t)|G}(y|g^*)\} \cdot |\mathcal{G}|\end{aligned}$$

So, using Condition REG:

$$\left|[\mathbf{D}(y')]_{t,g} - [\mathbf{D}(y)]_{t,g}\right| \leq \delta_t^*(y', y) \cdot |\mathcal{G}| \cdot \bar{L}$$

We can thus put an upper bound on the RHS of (13)

$$\lambda \cdot \{[(\mathbf{G}^* - \mathbf{G})(y') - (\mathbf{G}^* - \mathbf{G})(y)]_{t,g} - (\lim_{y' \downarrow y}[\mathbf{D}(y')]_{t,g} - [\mathbf{D}]_{t,g})\} \leq 2\lambda \cdot \delta_t^*(y', y) \cdot |\mathcal{G}| \cdot \bar{L}$$

Meanwhile, by REG:

$$\begin{aligned}&\left\{[\mathbf{G}^*(y')]_{t,g} - [\mathbf{G}^*(y)]_{t,g}\right\} \\ &\quad = \{F_{Y(t)|G}(y'|g^*) - F_{Y(t)|G}(y|g^*)\} \cdot \frac{F_{Y(t)|G}(y'|g') - F_{Y(t)|G}(y|g')}{F_{Y(t)|G}(y'|g^*) - F_{Y(t)|G}(y|g^*)} \geq \delta_t^*(y', y) \cdot \underline{\mathrm{L}}\end{aligned}$$

Thus inequality (13) then holds provided that $\delta_t^*(y', y) \cdot \underline{\mathrm{L}} \geq 2\lambda \cdot \delta_t^*(y', y) \cdot |\mathcal{G}| \cdot \bar{L}$, which holds trivially if $\delta_t^*(y', y) = 0$ and if and only if $\lambda \leq \frac{\underline{\mathrm{L}}}{2|\mathcal{G}| \cdot \bar{L}}$ if $\delta_t^*(y', y) > 0$.

A visualization of the intuition behind this result is depicted in Figure 1.

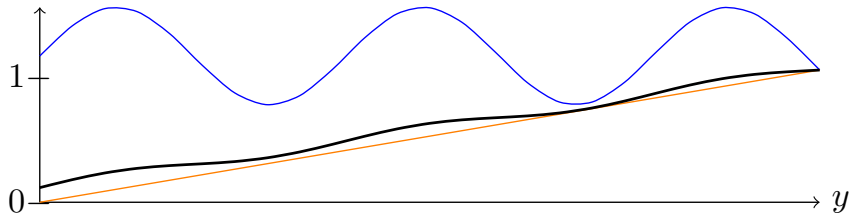


**Figure 1:** Depiction of Proposition B.1. The blue sinusoidal function depicts an example of a $\left[(\mathbf{G} + \mathbf{D})^\lambda(y)\right]_{t,g}$ that is not weakly increasing. The orange curve depicts $[\mathbf{G}^*(y)]_{t,g}$ which is weakly increasing. The black curve depicts $\left[\mathbf{G}^\lambda(y)\right]_{t,g}$, which is a linear combination of the blue and orange functions with weights $\lambda = 0.1$ and $1 - \lambda = 0.9$, respectively. This value of $\lambda$ is small enough that the black curve is weakly increasing everywhere.

### B.3 Proof of Lemma 1

Suppose that $A^{[t]}$ does not have full row rank. This implies that for some $\mathcal{Z}_0 \subset \mathcal{Z}$, each of the remaining rows of $A^{[t]}$ for $z \notin \mathcal{Z}_0$ can be written as a linear combination of the

rows of $A^{[t]}$ for $z \in \mathcal{Z}_0$. Take such a $z^* \notin \mathcal{Z}_0$, and accordingly let

$$A^{[t]}_{z^*,g} = \sum_{z \in \mathcal{Z}_0} \gamma_z \cdot A^{[t]}_{z,g} \quad \text{for all } g \in \mathcal{G}$$

Note then that Eq. (5) implies that

$$\begin{aligned} F_{(YD)|Z=z^*}(y,t) &= \sum_{g \in \mathcal{G}} A^{[t]}_{z^*,g} \cdot P(G_i = g) \cdot F_{Y(t)|G=g}(y) \\ &= \sum_{g \in \mathcal{G}} \left( \sum_{z \in \mathcal{Z}_0} \gamma_z \cdot A^{[t]}_{z,g} \right) \cdot P(G_i = g) \cdot F_{Y(t)|G=g}(y) \\ &= \sum_{z \in \mathcal{Z}_0} \gamma_z \cdot \sum_{g \in \mathcal{G}} A^{[t]}_{z,g} \cdot P(G_i = g) \cdot F_{Y(t)|G=g}(y) = \sum_{z \in \mathcal{Z}_0} \gamma_z \cdot F_{(YD)|Z=z}(y,t) \end{aligned}$$

where the RHS on the last line does not depend on the distribution of observables for $i$ such that $Z_i = z^*$. Thus, $F_{(YD)|Z=z^*}(y,t)$ adds no information that is not contained in $F_{(YD)|Z=z}(y,t)$ for $z \in \mathcal{Z}_0$. If $\mu^t_g$ is outcome-nonrestrictive identified, it must be using the distribution $\mathcal{P}_{YTZ|Z \in \mathcal{Z}_0}$ rather than the full unconditional distribution $\mathcal{P}_{obs} = \mathcal{P}_{YTZ}$.

### B.4 Proof of Proposition 2

Suppose first that $|\mathcal{G}| > |\mathcal{Z}|$ and $A = A^{[t]}$ has full row rank of $|\mathcal{Z}|$. Then since $A$ has full row-rank of $|\mathcal{Z}|$, there exists a subset of $|\mathcal{Z}|$ columns that are linearly independent from one another. Write $A = [\tilde{A}, \tilde{A}^c]$ where $\tilde{A}$ is an invertible $|\mathcal{Z}| \times |\mathcal{Z}|$ matrix of these columns, and $\tilde{A}_c$ are the others. Write the system $A'\alpha = c$ in this notation as

$$\begin{bmatrix} \tilde{A}' \\ \tilde{A}'_c \end{bmatrix} \alpha = \begin{pmatrix} \tilde{c} \\ \tilde{c}_c \end{pmatrix}$$

where $\tilde{c}$ denotes the $|\mathcal{Z}|$ components of $c$ corresponding to the columns of $A$ put into in $\tilde{A}$, and $\tilde{c}_c$ are the remaining entries $c_g$ of $c$. Then $\alpha = \tilde{A}'^{-1}\tilde{c}$, which can be seen by left-multiplying the above equation by the $|\mathcal{Z}| \times |\mathcal{G}|$ matrix $[\tilde{A}'^{-1}, \mathbf{0}^{|\mathcal{Z}| \times |\mathcal{G}| - |\mathcal{Z}|}]$. Intuitively, the system $A'\alpha = c$ is over-determined, so we only only need the components $\tilde{c}$ of $c$ to uniquely determine the vector $\alpha$.

Now consider the case in which $|\mathcal{G}| < |\mathcal{Z}|$, so that the system $A'\alpha = c$ is now undetermined. Suppose for now that the rank of $A$ is $|\mathcal{G}|$ so that it has full column rank. One solution $\alpha$ can then be obtained by writing $A = \begin{bmatrix} \tilde{A} \\ \tilde{A}_c \end{bmatrix}$ where $\tilde{A}$ is an invertible $|\mathcal{G}| \times |\mathcal{G}|$ matrix representing $|\mathcal{G}|$ linearly independent *rows* of $A$. Now consider $\alpha = \begin{pmatrix} \tilde{A}^{-1}c \\ \mathbf{0}^{(|\mathcal{Z}|-|\mathcal{G}|)\times 1} \end{pmatrix}$ where note that $\tilde{A}^{-1}c$ is $|\mathcal{G}|-$component vector. This represents a solution to $A'\alpha = c$ since

$$A' \begin{pmatrix} \tilde{A}'^{-1}c \\ \mathbf{0}^{(|\mathcal{Z}|-|\mathcal{G}|)\times 1} \end{pmatrix} = [\tilde{A}, \tilde{A}_c] \begin{pmatrix} \tilde{A}^{-1}c \\ \mathbf{0}^{(|\mathcal{Z}|-|\mathcal{G}|)\times 1} \end{pmatrix} = c$$

We can combine the constructions in the two special cases considered above to relax any assumptions about the cardinality of $\mathcal{Z}$ and $\mathcal{G}$ or the rank of $A$. Let the rank of $A$ be $k \leq \min\{|\mathcal{Z}|, |\mathcal{G}|\}$. Write $A = A_k[I_k, M]$ where $A_k$ is a $k \times |\mathcal{G}|$ matrix composed of $k$ linearly independent columns of $A$, and $M$ is $(|\mathcal{G}| - k) \times k$ matrix that expresses the remaining $(|\mathcal{G}| - k)$ columns of $A$ as linear combinations of the columns of $A$ represented in $A_k$. Write $c = \begin{pmatrix} \tilde{c}_k \\ \tilde{c}_c \end{pmatrix}$ where $\tilde{c}_k$ collects the corresponding $k$ components of $c$. Note that if $c' = \alpha' A$ has a solution, then $c' = \tilde{c}_k'[I_k, M]$, since $c' = (\alpha_k' A_k)[I, M]$ where the $k$ components of $c'$ corresponding to the columns in $A_k$ are $\alpha_k' A_k$, so $\tilde{c}_k' = \alpha_k' A_k$. Now split the rows of $A_k$ as $A_k = \begin{bmatrix} \tilde{A} \\ \tilde{A}_c \end{bmatrix}$ where $\tilde{A}$ is a square invertible $k \times k$ matrix representing $k$ linearly independent rows of $A_k$ and $\tilde{A}_c$ is $(|\mathcal{Z}| - k) \times k$. Now $\alpha = \begin{pmatrix} \tilde{c}_k' \tilde{A}^{-1} \\ \mathbf{0}^{(|\mathcal{Z}|-k)\times 1} \end{pmatrix}$ represents a solution to $c' = \alpha' A$ because $[\tilde{c}_k' \tilde{A}^{-1}, \mathbf{0}^{1\times(|\mathcal{Z}|-k)}]A = [\tilde{c}_k' \tilde{A}^{-1}, \mathbf{0}^{1\times(|\mathcal{Z}|-k)}] \begin{bmatrix} \tilde{A} \\ \tilde{A}_c \end{bmatrix} [I_k, M] = \tilde{c}_k'[I_k, M] = c'$.

In all of the three cases considered above, we can write any non-zero elements $\alpha_z$ of a $\alpha$ yielding a binary combination as components $x_z$ of $x = M^{-1}b$, where $M$ is an invertible $n \times n$ binary matrix (i.e. having entries of 0 or 1), and $b$ an n-component binary vector. Equivalently, $x$ represents the unique solution to $Mx = b$. Cramer's rule for such a solution establishes that the $x_z$ can be written as $x_z = \frac{det(M_z)}{det(M)}$, where $M_z$ is a matrix that replaces the column $z$ of the matrix $M$ with the vector $b$. Since both $M$ and $b$ are composed of binary entries, the matrix $M_z$ is always binary as well. The result now follows as stated in Proposition 2 since 0 is always a possible value of $det(M_z)$.

### B.5 Proof of Proposition 3

Given $\mathbb{E}[\nu_i | Z_i = 0]$, the parameter $\gamma_3 - \gamma_1 - \gamma_2$ is given by

$$\begin{aligned}
\gamma_3 - \gamma_1 - \gamma_2 &= \mathbb{E}[Y_i|Z_i = C] - \mathbb{E}[Y_i|Z_i = A] - \mathbb{E}[Y_i|Z_i = B] + \mathbb{E}[Y_i|Z_i = 0] \\
&= \mathbb{E}[Y_i(T_i(C)) - Y_i(T_i(A)) - Y_i(T_i(B)) + Y_i(T_i(0))] \\
&= \mathbb{E}[Y_i(C) - Y_i(A) - Y_i(B) + Y_i(0)] + \mathbb{E}[Y_i(T_i(C)) - Y_i(C)] \\
&\quad - \mathbb{E}[Y_i(T_i(A)) - Y_i(A)] - \mathbb{E}[Y_i(T_i(B)) - Y_i(B)] + \mathbb{E}[Y_i(T_i(0)) - Y_i(0)]
\end{aligned}$$

Each of the last three terms in the final line can differ from zero in ways that do not offset one another, provided that imperfect compliance is allowed, i.e. $P(T_i(z) \neq z) > 0$ for some $z \in \mathcal{Z}$.

# C Extended analysis of identification under NSOG

It is known that unconditional means $\mathbb{E}[Y_i(t)]$ of a given potential outcome $Y_i(t)$ can be point-identified, given an order condition on the instruments, under an assumption of "no-selection on gains" (NSOG) (see e.g. Kolesár (2013) and Arora et al. (2021) for versions of this result).[7] Note that identification of $\mathbb{E}[Y_i(t)]$ and $\mathbb{E}[Y_i(t')]$ immediately implies identification of unconditional average treatment effects $\mathbb{E}[Y_i(t') - Y_i(t)]$ as well.

NSOG says that treatment effects are mean independent of actual treatment, given any realization of the instruments:

**Assumption NOSG (no selection on gains).** *For any $t, t', t_1, t_2 \in \mathcal{T}$ and $z \in \mathcal{Z}$:*

$$\mathbb{E}[Y_i(t') - Y_i(t)|T_i = t_1, Z_i = z] = \mathbb{E}[Y_i(t') - Y_i(t)|T_i = t_2, Z_i = z]$$

NSOG implies that if we consider any fixed treatment value $0 \in \mathcal{T}$, then $\mathbb{E}[Y_i(t') - Y_i(0)|T_i = t, Z_i = z] = \mathbb{E}[Y_i(t') - Y_i(0)|Z_i = z]$ for any $t, z$, which coupled with independence (2) in turn implies that $\mathbb{E}[Y_i(t') - Y_i(0)|T_i = t, Z_i = z] = \mathbb{E}[Y_i(t') - Y_i(0)] := \Delta_{t'}$, where note that $\Delta_{t'}$ does not depend on $z$ or $t$. This normalization against an arbitrary treatment $0 \in \mathcal{T}$ allows us to carry around one less index in our expressions.

## C.1 Identification under NSOG

This subsection first shows that $\mathbb{E}[Y_i(t)]$ can be point identified for each $t \in \mathcal{T}$ under NSOG, given rich enough support of the instruments. The proof essentially follows that of Arora et al. (2021), which adapts an argument from Kolesár (2013) to cases in which the treatments $\mathcal{T}$ are not necessarily ordered.

NSOG implies that:

$$\mathbb{E}[Y_i - Y_i(0)|T_i = t, Z_i = z] = \mathbb{E}[Y_i(t) - Y_i(0)|T_i = t, Z_i = z] = \Delta_t$$

Averaging over the conditional distribution of $T_i$ given $Z_i = z$, we have by the law of iterated expectations that

$$\mathbb{E}[Y_i - Y_i(0)|Z_i = z] = \sum_{t \in \mathcal{T}} P(T_i = t|Z_i = z) \cdot \Delta_t \tag{15}$$

To now see that $\mathbb{E}[Y_i(t)]$ can be identified under NSOG given rich enough instrument support, let us assume that $|\mathcal{Z}| \geq |\mathcal{T}|$ and suppose that there exists a set of $|\mathcal{T}|$ instrument values $\tilde{\mathcal{Z}} \subseteq \mathcal{Z}$ such that the $|\mathcal{T}| \times |\mathcal{T}|$ matrix $\Sigma$ with entries $\Sigma_{zt} = P(Z_i = z, T_i = t)$ over all $z \in \tilde{\mathcal{Z}}$ is invertible, with $P(Z_i = z) > 0$ for each $z \in \tilde{\mathcal{Z}}$.

Eq. (15) can be re-written by multiplying both sides by $P(Z_i = z)$ as

$$\mathbb{E}[\{Y_i - Y_i(0)\} \cdot \mathbb{1}(Z_i = z)] = \sum_{t \in \mathcal{T}} \Sigma_{zt} \cdot \Delta_t$$

[7] Kolesár (2013) calls this "constant average treatment effects", and does not use the term NSOG.

for each $z \in \tilde{\mathcal{Z}}$. Equivalently, using independence:

$$\begin{aligned}
\mathbb{E}[Y_i \cdot \mathbb{1}(Z_i = z)] &= P(Z_i = z) \cdot \mathbb{E}[Y_i(0)] + \sum_{t \in \mathcal{T}} \Sigma_{zt} \cdot \Delta_t \\
&= P(Z_i = z) \cdot \mathbb{E}[Y_i(0)] + \sum_{t \in \mathcal{T}, t \neq 0} \Sigma_{zt} \cdot \Delta_t \\
&= \left\{ P(Z_i = z, T_i = 0) + \sum_{t \in \mathcal{T}, t \neq 0} \Sigma_{zt} \right\} \cdot \mathbb{E}[Y_i(0)] + \sum_{t \in \mathcal{T}, t \neq 0} \Sigma_{zt} \cdot \Delta_t \\
&= P(Z_i = z, T_i = 0) \cdot \mathbb{E}[Y_i(0)] + \sum_{t \in \mathcal{T}, t \neq 0} \Sigma_{zt} \cdot \mathbb{E}[Y_i(0)] + \sum_{t \in \mathcal{T}, t \neq 0} \Sigma_{zt} \cdot \Delta_t \\
&= P(Z_i = z, T_i = 0) \cdot \mathbb{E}[Y_i(0)] + \sum_{t \in \mathcal{T}, t \neq 0} \Sigma_{zt} \cdot \mathbb{E}[Y_i(t)] \\
&= \sum_{t \in \mathcal{T}} \Sigma_{zt} \cdot \mathbb{E}[Y_i(t)]
\end{aligned}$$

using that $\Delta_0 = 0$ in the second equality. This yields a system of $|\mathcal{T}|$ equations in the $|\mathcal{T}|$ unknowns $\mathbb{E}[Y_i(t)]$ with identified coefficients $\Sigma_{zt}$. Given that $\Sigma^{-1}$ is invertible, we have then that

$$\mathbb{E}[Y_i(t)] = \sum_{z \in \tilde{\mathcal{Z}}} \Sigma^{-1}_{tz} \cdot \mathbb{E}[Y_i \cdot \mathbb{1}(Z_i = z)] \tag{16}$$

Note that if $|\mathcal{Z}| \geq |\mathcal{T}|$ there may be overidentification restrictions implied by NSOG, that the RHS of (16) is the same for different possible choices of $\tilde{\mathcal{Z}} \subset \mathcal{Z}$ (note that $\Sigma$ also depends on the choice of $\tilde{\mathcal{Z}}$). Furthermore, the RHS of (16) is the estimand of a two-stage least squares regression of $Y_i$ on indicators for the mutually-exclusive treatments in $\mathcal{T}$ (and no constant), instrumented by indicators for the mutually-exclusive instrument values in $\tilde{\mathcal{Z}}$.

### C.2 How Theorem 2 does not cover NSOG

Since the result of the last section makes no assumption about which response types can show up in the population, it is compatible with any selection model $\mathcal{G} \subseteq \{0,1\}^{\mathcal{T}^{\mathcal{Z}}}$, including for example the full powerset $\{0,1\}^{\mathcal{T}^{\mathcal{Z}}}$ of possible response types $\mathcal{T}^{\mathcal{Z}}$.

Whatever $\mathcal{G}$ is, unconditional means like $\mathbb{E}[Y_i(t)]$ correspond to the choice $c = (1, \ldots, 1)'$ in $\mathbb{R}^{|\mathcal{G}|}$. As long as $\mathcal{G}$ allows never-takers with respect to treatment $t$, this choice of $c$ will not lie in the rowspace of $A^{[t]}$. The unrestricted selection model $\mathcal{G} = \{0,1\}^{\mathcal{T}^{\mathcal{Z}}}$, for example, features such never-takers for any $t \in \mathcal{T}$. Thus the result of the last section demonstrates that it is possible to achieve point identification of $\mu^t_c$ without $c \in rs(A^{[t]})$, if we impose NSOG and that $\Sigma^{-1}$ exists.

Note that the imposing of NSOG makes this identification *not* outcome-nonrestrictive. However, it is illustrative to see where the proof of Theorem 2 breaks down in the case of the NSOG identification result. Let $NSOG$ denote the set of distributions $\mathcal{P}$ for which $P_{latent}(\mathcal{P})$ satisfies NSOG. In this notation, the last section establishes that $\{\theta(\mathcal{P}) : \mathcal{P} \in$

$(M \cap NSOG)$ and $\phi(\mathcal{P}) = \mathcal{P}_{obs}\}$ is a singleton for all $\mathcal{P}_{obs}$ that satisfy the rich support condition that $\Sigma^{-1}$ exists, which requires that there be no $\mathcal{P}, \mathcal{P}' \in M \cap NSOG$ such that $\phi(\mathcal{P}) = \phi(\mathcal{P}')$ but $\theta(\mathcal{P}) \neq \theta(\mathcal{P}')$ and such that $\Sigma^{-1}$ exists under $\mathcal{P}$ or $\mathcal{P}'$.

To see that there is no contradiction with Theorem 2, I below show that given a $\mathcal{P} \in (M \cap NSOG)$, the alternative distribution $\mathcal{P}'$ defined from it in the proof of Theorem 2 does not lie within $NSOG$ when $(1, \ldots 1)' \notin rs(A^{[t]})$. In partcular, the remainder of this section shows that if $(1, \ldots 1)' \notin rs(A^{[t]})$ for any given $t \in \mathcal{T}$, the construction $\mathcal{P}'$ utilized in the proof of Theorem 2 cannot lie in $NSOG$. If on the other hand $(1, \ldots 1)' \in rs(A^{[t]})$, then Eq. (12) in the proof of Theorem 2 shows that $\theta(\mathcal{P}) = \theta(\mathcal{P}')$, consistent with $\theta$ being identified.

Recall that the way in which the proof of Theorem 2 builds a candidate $\mathcal{P}'$ from the actual distribution $\mathcal{P}$ is to construct from the set of true potential outcome CDFs $\mathbf{G}^*$ $[\mathbf{G}^*(y)]_{t,g} := P(G_i = g) \cdot F_{Y(t)|G=g}(y)$ a new set of such CDFs $\mathbf{G}^\lambda$. For continuity with the notation used in this discussion so far, let $\mathcal{P}'$ correspond to the collection of CDFs $\mathbf{G}^\lambda$, and let us make explicit whether outcome expectations are with respect to the distribution $\mathcal{P}$ or $\mathcal{P}'$,[8]. Then we have by integrating Eq. (9) that:

$$P(G_i = g) \cdot \mathbb{E}_{\mathcal{P}'}[Y_i(t)|G_i = g] = P(G_i = g) \cdot \mathbb{E}_{\mathcal{P}}[Y_i(t)|G_i = g]$$
$$+ \lambda \cdot \sum_{g'} [I - (A^{[t]})^+ A^{[t]}]_{g,g'} \cdot P(G_i = g') \cdot \{\mathbb{E}_{\mathcal{P}}[Y_i(t)|G_i = g'] - \mathbb{E}_{\mathcal{P}}[Y_i(t)|G_i = g^*]\}$$

Then using independence (2):

$$\mathbb{E}_{\mathcal{P}'}[Y_i(t) - Y_i(0)|G_i = g, Z_i = z] = \mathbb{E}_{\mathcal{P}'}[Y_i(t)|G_i = g] - \mathbb{E}_{\mathcal{P}'}[Y_i(0)|G_i = g]$$
$$= \mathbb{E}_{\mathcal{P}}[Y_i(t) - Y_i(0)|G_i = g]$$
$$+ \lambda \cdot \sum_{g'} [I - (A^{[t]})^+ A^{[t]}]_{g,g'} \cdot \frac{P(G_i = g')}{P(G_i = g)} \cdot \{\mathbb{E}_{\mathcal{P}}[Y_i(t)|G_i = g'] - \mathbb{E}_{\mathcal{P}}[Y_i(t)|G_i = g^*]\}$$
$$- \lambda \cdot \sum_{g'} [I - (A^{[0]})^+ A^{[0]}]_{g,g'} \cdot \frac{P(G_i = g')}{P(G_i = g)} \cdot \{\mathbb{E}_{\mathcal{P}}[Y_i(0)|G_i = g'] - \mathbb{E}_{\mathcal{P}}[Y_i(0)|G_i = g^*]\}$$

[8]The response type probabilities $P(G_i = g)$ are the same for both $\mathcal{P}$ and $\mathcal{P}'$ so I leave this implicit for ease of exposition.

Therefore, for any $t_1 \in \mathcal{T}$:

$$\mathbb{E}_{\mathcal{P}'}[Y_i(t) - Y_i(0)|T_i = t_1, Z_i = z] = \mathbb{E}_{\mathcal{P}'}[Y_i(t) - Y_i(0)|A^{[t_1]}_{z,G_i} = 1, Z_i = z]$$
$$= \sum_g P(G_i = g|A^{[t_1]}_{z,G_i} = 1) \cdot \mathbb{E}_{\mathcal{P}'}[Y_i(t) - Y_i(0)|G_i = g, Z_i = z]$$
$$= \frac{1}{P(A^{[t_1]}_{z,G_i} = 1)} \sum_g P(G_i = g) \cdot A^{[t_1]}_{z,g} \cdot \mathbb{E}_{\mathcal{P}'}[Y_i(t) - Y_i(0)|G_i = g, Z_i = z]$$
$$= \frac{1}{P(A^{[t_1]}_{z,G_i} = 1)} \cdot \sum_g P(G_i = g) \cdot A^{[t_1]}_{z,g} \cdot \Big[\mathbb{E}_{\mathcal{P}}[Y_i(t) - Y_i(0)|G_i = g]$$
$$+\lambda \cdot \sum_{g'} [I - (A^{[t]})^+ A^{[t]}]_{g,g'} \cdot \frac{P(G_i = g')}{P(G_i = g)} \cdot \{\mathbb{E}_{\mathcal{P}}[Y_i(t)|G_i = g'] - \mathbb{E}_{\mathcal{P}}[Y_i(t)|G_i = g^*]\}$$
$$-\lambda \cdot \sum_{g'} [I - (A^{[0]})^+ A^{[0]}]_{g,g'} \cdot \frac{P(G_i = g')}{P(G_i = g)} \cdot \{\mathbb{E}_{\mathcal{P}}[Y_i(0)|G_i = g'] - \mathbb{E}_{\mathcal{P}}[Y_i(0)|G_i = g^*]\}\Big]$$
$$= \Delta_t + \frac{\lambda}{P(A^{[t_1]}_{z,G_i} = 1)} \cdot \Bigg[\sum_{g'} [A^{[t_1]}(I - (A^{[t]})^+ A^{[t]})]_{z,g'} \cdot P(G_i = g') \cdot \{\mathbb{E}_{\mathcal{P}}[Y_i(t)|G_i = g'] - \mathbb{E}_{\mathcal{P}}[Y_i(t)|G_i = g^*]\}$$
$$- \sum_{g'} [A^{[t_1]}(I - (A^{[0]})^+ A^{[0]})]_{z,g'} \cdot P(G_i = g') \cdot \{\mathbb{E}_{\mathcal{P}}[Y_i(0)|G_i = g'] - \mathbb{E}_{\mathcal{P}}[Y_i(0)|G_i = g^*]\}\Bigg] \quad (17)$$

where $\Delta_t := \mathbb{E}_{\mathcal{P}}[Y_i(t) - Y_i(0)]$. Note that we can simplify the denominator as $P(A^{[t_1]}_{z,G_i} = 1) = \sum_g P(G_i = g) \cdot A^{[t_1]}_{z,g} = [A^{[t_1]}P]_z$, where $P$ is a vector of response type probabilities $P_g = P(G_i = g)$. Since $\Sigma_{zt} = P(Z_i = z, T_i = t) = P(Z_i = z) \cdot P(T_i = t|Z_i = z) = P(Z_i = z) \cdot \sum_g P(G_i = g) \cdot A^{[t]}_{z,g} = P(Z_i = z) \cdot [A^{[t]}P]_z$. We can thus rewrite $P(A^{[t_1]}_{z,G_i} = 1)$ as $\Sigma_{zt}/P(Z_i = z)$.

For us to have $\mathcal{P}' \in NSOG$, it must be the case that the RHS of (17) does not depend on $z$ or $t_1$, and equals $\Delta'_t(\lambda) := \mathbb{E}_{\mathcal{P}'}[Y_i(t) - Y_i(0)]$ for any $\mathcal{P} \in (M \cap NSOG)$. In the notation $\Delta'_t(\lambda)$ we make explicit that the value of $\mathbb{E}_{\mathcal{P}'}[Y_i(t) - Y_i(0)]$ could depend on $\lambda$. In the case of $t_1 = t$, expression (17) for $\Delta'_t(\lambda)$ simplifies to

$$\Delta_t - \frac{\lambda}{\Sigma_{zt}} \cdot P(Z_i = z) \cdot \sum_{g'} [A^{[t]} - A^{[t]}(A^{[0]})^+ A^{[0]}]_{z,g'} \cdot P(G_i = g') \cdot \{\mathbb{E}_{\mathcal{P}}[Y_i(0)|G_i = g'] - \mathbb{E}_{\mathcal{P}}[Y_i(0)|G_i = g^*]\}$$

using that $A^{[t]}(A^{[t]})^+ A^{[t]} = A^{[t]}$. Similarly, taking $t_1 = 0$, we have that $\Delta'_t(\lambda)$ is equal to

$$\Delta_t + \frac{\lambda}{\Sigma_{zt}} \cdot P(Z_i = z) \cdot \sum_{g'} [A^{[0]} - A^{[0]}(A^{[t]})^+ A^{[t]}]_{z,g'} \cdot P(G_i = g') \cdot \{\mathbb{E}_{\mathcal{P}}[Y_i(t)|G_i = g'] - \mathbb{E}_{\mathcal{P}}[Y_i(t)|G_i = g^*]\}$$

Note that for any $\mathcal{P}$, there exists a small enough $\lambda > 0$ that $\mathcal{P}' \in M$. For the above equations to simultaneously hold for any such $\lambda > 0$, we must have for any $z$ such that

$P(Z_i = z) > 0$:

$$\sum_{g'}[A^{[0]}(I - (A^{[t]})^+A^{[t]})]_{z,g'} \cdot P(G_i = g') \cdot \{\mathbb{E}_{\mathcal{P}}[Y_i(t)|G_i = g'] - \mathbb{E}_{\mathcal{P}}[Y_i(t)|G_i = g^*]\}$$
$$+ \sum_{g'}[A^{[t]}(I - (A^{[0]})^+A^{[0]})]_{z,g'} \cdot P(G_i = g') \cdot \{\mathbb{E}_{\mathcal{P}}[Y_i(0)|G_i = g'] - \mathbb{E}_{\mathcal{P}}[Y_i(0)|G_i = g^*]\} = 0 \tag{18}$$

for all $\mathcal{P} \in M \cap REG \cap NSOG$. Consider a distribution $\mathcal{P}$ for which $\mathcal{P}_Z$ has full support $\mathcal{Z}$, and for which conditional average treatment effects take the separable form $\mathbb{E}[Y_i(t)|G_i = g] = \lambda_g + \Delta_t$, where $\Delta_0 := 0$. Defining $\tilde{\lambda}_g := \lambda_g - \lambda_{g^*}$, Eq. (18) reads in this case:

$$\sum_{g'}[A^{[0]}(I-(A^{[t]})^+A^{[t]})]_{z,g'} \cdot P(G_i = g') \cdot \tilde{\lambda}_{g'} + \sum_{g'}[A^{[t]}(I - (A^{[0]})^+A^{[0]})]_{z,g'} \cdot P(G_i = g') \cdot \tilde{\lambda}_{g'} = 0$$

Given that $\tilde{\lambda}_{g'}$ can be freely chosen such that $P(G_i = g') \cdot \tilde{\lambda}_{g'} = \mathbb{1}(g' = g)$ for any $g \in \mathcal{G}$ and $\mathcal{P}_G$, this can only be true when $A^{[0]}(I - (A^{[t]})^+A^{[t]}) = A^{[t]}(I - (A^{[0]})^+A^{[0]})$ entry by entry as matrices. We'll now see that this can only be true for all $t \in \mathcal{T}$ if $c = (1, \dots 1)' \in rs(A^{[t]})$ for all $t \in \mathcal{T}$.

Note that the matrix $(A^{[t]})^+A^{[t]}$ is an orthogonal projector onto into $rs(A^{[t]})$, and $(A^{[0]})^+A^{[0]}$ is an orthogonal projector onto into $rs(A^{[0]})$, and the required condition is

$$A_z^{[t]\prime}(I - (A^{[0]})^+A^{[0]}) = -A_z^{[0]\prime}(I - (A^{[t]})^+A^{[t]})$$

for all $z \in \mathcal{Z}$, where the row-vector $A_z^{[t]\prime}$ denotes row $z$ of the matrix $A^{[t]}$, and similarly for $A^{[0]}$. Note that the row-vector $A_z^{[t]\prime}(I - (A^{[0]})^+A^{[0]})$ belongs to the orthogonal complement of $rs(A^{[0]})$ in $\mathbb{R}^{|\mathcal{G}|}$. It is thus orthogonal to any row of $A^{[0]}$, including $A_z^{[0]\prime}$. But $-A_z^{[0]\prime}(I - B)$ cannot be orthogonal to $c_z$ unless $A_z^{[0]\prime}(A^{[t]})^+A^{[t]} = A_z^{[0]\prime}$ so that $-A_z^{[0]\prime}(I - (A^{[t]})^+A^{[t]})$ is the zero vector. In that case, note that $A_z^{[t]\prime}(I - (A^{[0]})^+A^{[0]})$ is the zero vector as well, so we have that $A_z^{[t]\prime} \in rs(A^{[0]})$ and $A_z^{[0]\prime} \in rs(A^{[t]})$. Compiling over all $z \in \mathcal{Z}$, we have that $A^{[0]}$ and $A^{[t]}$ have the same row-space. Repeating this argument over all $t \in \mathcal{T}$, we have that $rs(A^{[t]})$ is the same for all $t \in \mathcal{T}$.

Now let us see that this in turn implies that $(1, \dots 1)' \in rs(A^{[t]})$. Note that $\sum_{t' \in \mathcal{T}} A_z^{[t']\prime} = (1, \dots 1)'$ for any $z$, because all response types take one and only one treatment when $Z_i = z$. But since $A_z^{[t']\prime} \in rs(A^{[t']})$, it must also be in the rowspace of $A^{[t]}$. Since $A_z^{[t']\prime} \in rs(A^{[t]})$ for each $t'$, the linear combination $\sum_{t' \in \mathcal{T}} A_z^{[t']\prime} = (1, \dots 1)'$ is also in $rs(A^{[t]})$. Thus we have shown that $\mathcal{P}' \in NSOG$ implies that $(1, \dots 1)' \in rs(A^{[t]})$ for all $t$.

### C.3 Further examples to which Theorem 2 does *not* apply

Another example of an IV identification result that is not covered by Theorem 2 is the "compliers–defiers" result of de Chaisemartin (2017) that the local average treatment effect among a subset of compliers is identified in a setting with a binary treatment and instrument, if there are more compliers than defiers and a subset of the compliers have the

same average treatment effect as the defiers. Again, this additional assumption places restrictions on the joint distribution of response types $G_i$ and potential outcomes $\tilde{Y}_i$. Further, the identified parameter conditions on an event (a particular subgroup of the compliers) that is less course than the groups $G_i$ that are defined simply by counterfactual selection behavior, so does not fit the form $\Delta_c^{t,t'} = \mu_c^{t'} - \mu_c^t$ that Theorem 2 and Proposition 1 speak to. Similar considerations apply to recent results of (Comey et al., 2023) that show identification of the local average treatment effect among "supercompliers" in a setting in which $\mathcal{Y} = \mathcal{T} = \mathcal{Z} = \{0, 1\}$, where the supercompliers are defined as the subset of compliers that have a strictly positive treatment effect. This model imposes monotonicity in the outcome equation, and the conditioning event for the supercomplier LATE conditions both on selection behavior *and* a property of outcomes, namely that $Y_i(1) > Y_i(0)$.

Another type of identification result that is not covered by Theorem 2 above—although it is outcome-nonrestrictive—is identification of a treatment effect parameter that does not maintain two fixed treatment values $t$ and $t'$ across all units included in the parameter. An example of this kind arises in Kline and Walters (2016), in which the identified causal parameter compares the effect of Head Start to one of two next-best alternatives (either traditional pre-school or no pre-school). This estimand combines two response types for which this next-best alternative is generally different. See Section F.0.2 for details.

When $c \notin rs(A^{[t]})$, Theorem 2 establishes that the parameter $\mu_c^t$ is not *point* identified in an outcome-nonrestrictive manner. However, the data may still provide identifying information about the value of $\mu_c^t$ if auxiliary conditions are maintained, for example that the support of $Y_i$ is bounded with known bounds. Appendix I considers partial identification of $\mu_c^t$ in such settings, and also relates the results of this paper to recent results by Bai et al. (2024), who focus on bounding the ATE and unconditional means in particular.

# D Relationship to recent work

This section discusses how the results of this paper relate to recent results characterizing identification in IV models by Navjeevan, Pinto and Santos (2023) (NPS) as well as Heckman and Pinto (2018).

## D.1 Relationship to Navjeevan, Pinto and Santos (2023)

NPS consider *unconditional* expectations of functions taking the form $\mathbb{E}[\ell(\tilde{Y}_i, G_i)]$, which in general are allowed to mix potential outcomes and potential treatments, as well as covariates. NPS do not define or explore in depth a notion of "outcome-nonrestrictive" identification, as their framework allows the researcher to impose restrictions on outcomes of the types discussed in Section C.3.

NPS do mention conditional average treatment effects as a motivation for specializing

their general result to cases in which $\ell$ takes the separable form $Y_i(t) \cdot c(G_i)$, for some $t \in \mathcal{T}$ (see their Section 4.4). In these separable cases, NPS derive results that are related to but distinct from my Theorems 1 and 2 (which were obtained independently).

In particular, Corollary 4.4 of NPS assumes discrete instruments and supposes that no additional restrictions are placed on the distribution of unobservables aside from the existence of finite first moments. This model is thus essentially the same as the model $M$ (see Footnote 2) I use to define outcome-nonrestrictive identification. From Corollary 4.4, NPS derive two important implications. Firstly, they find that the conditions on function $c(\cdot)$ for identification of $\mathbb{E}[Y_i(t) \cdot c(G_i)]$ are equivalent to those for identification of $\mathbb{E}[f(Y_i(t)) \cdot c(G_i)]$ for any bounded function $f(\cdot)$. This implies that $\mathbb{E}[f(Y_i(t)) \cdot c(G_i)]$ is identified if and only if $P(c(G_i) = 1)$ is (take $f(\cdot) = 1$). Second, NPS find that a moment of the form $\mathbb{E}[f(Y_i(t)) \cdot c(G_i)]$ is identified if and only if the function $c(g)$ can be written as $\mathbb{E}[\kappa(Z_i) \cdot \mathbb{1}(T_i = t) | G_i = g]$ for some function $\kappa$. Though NPS do not characterize it in this way, one can see that this is equivalent to $c \in rs(A^{[t]})$ by applying the law of iterated expectations over $Z_i$.[9]

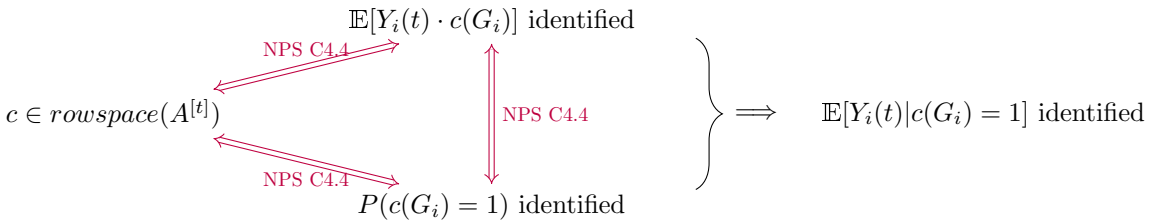


**Figure 2:** On left, $\Longleftrightarrow$ symbols (in purple) depict implications Corollary 4.4 of NPS, for parameters of the form $\mathbb{E}[Y_i(t)|c(G_i) = 1]$. On right, $\Longrightarrow$ symbol (in black) depicts an implication of $\mathbb{E}[Y_i(t)|c(G_i) = 1] = \frac{\mathbb{E}[f(Y_i(t)) \cdot c(G_i)]}{P(c(G_i)=1)}$.

These results of NPS are summarized in Figure 2. Taken together, they imply Theorem 1 but not Theorem 2 of this paper. Since $\mathbb{E}[Y_i(t) \cdot c(G_i)]$ and $P(c(G_i) = 1)$ are both identified when $c$ is in the rowspace of $A^{[t]}$, the results of NPS readily establish that $\mathbb{E}[Y_i(t)|c(G_i) = 1]$ is identified provided that $P(c(G_i) = 1) > 0$, in the case of a binary valued function $c$. However, their results do not establish that the conditional expectation $\mathbb{E}[Y_i(t)|c(G_i) = 1]$ is *only* identified when $c \in rs(A^{[t]})$ holds. Instead, they show that $\mathbb{E}[Y_i(t) \cdot c(G_i)]$ and $P(c(G_i) = 1)$ can only be identified separately if $c \in rs(A^{[t]})$ holds.

By contrast, Theorem 2 establishes the necessary direction of $c \in rs(A^{[t]})$ for when $\mathbb{E}[Y_i(t)|c(G_i) = 1]$ is identified (given discrete instruments), as depicted in Figure 3 below. While Theorem 1 establishes that $\mathbb{E}[Y_i(t) \cdot c(G_i)]$, $P(c(G_i) = 1)$ and $\mathbb{E}[Y_i(t)|c(G_i) = 1]$ are all identified if $c$ belongs to the rowspace of $A^{[t]}$, Theorem 2 establishes that $\mathbb{E}[Y_i(t)|c(G_i) = 1]$ is *only* identified if $c$ belongs to the rowspace of $A^{[t]}$.

Beyond Theorem 2, the present paper also differs from NPS by its exploration of the

[9] The closest way in NPS of stating this condition to $c \in rs(A^{[t]})$ seems to be Eq. (28) from their discussion of the selection model of Kline and Walters (2016). In my notation their Eq. (28) reads as $\min_{\alpha \in \mathbb{R}^{|\mathcal{Z}|}} \left( c(g) - \sum_z \alpha_z A^{[t]}_{zg} \right)^2 = 0$, which is equivalent to $c \in rs(A^{[t]})$.

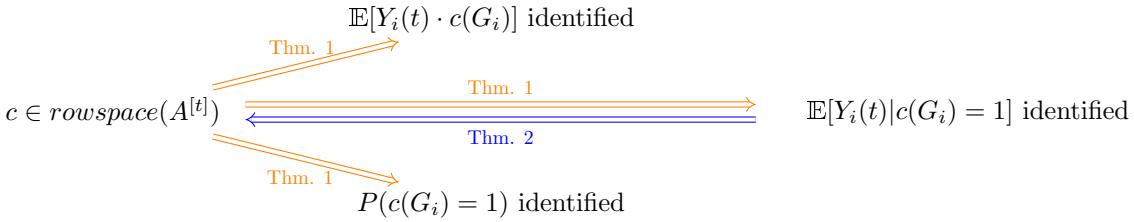


**Figure 3:** Implications of Theorems 2 (in blue) and 1 (in orange) of this paper.

implications of $c \in rs(A^{[t]})$ for the identification of conditional average treatment effects, in the case that $c$ is binary-valued. This requires finding functions $c$ that belong to the *intersection* of rowspaces of $A^{[t]}$ and $A^{[t']}$ for $t' \neq t$ together with the unit cube, as we saw in Section 4.2. This analysis shows, in the positive direction (Theorem 1), how $c \in rs(A^{[t]})$ synthesizes many identification results for treatment effects from the literature (Appendix G). In the other direction (Theorem 2), this allows one to exhaustively catalog identification results for a given support of $T_i$ and $Z_i$, as described in Section 4.

*An illustrative example:* To appreciate the difference between $\mathbb{E}[Y_i(t)|c(G_i) = 1]$ being identified and $\mathbb{E}[Y_i(t) \cdot c(G_i)]$ being identified, consider a setting with a binary treatment and binary instrument in which $\mathcal{G}$ allows all four response types: always-takers, never-takers, compliers and defiers. The choice model $\mathcal{G}$ is represented by the matrix $A = A^{[1]}$:

| | *n.t.* | *comp.* | *def.* | *a.t.* |
|---|---|---|---|---|
| $\mathbf{z = 0}$ | 0 | 0 | 1 | 1 |
| $\mathbf{z = 1}$ | 0 | 1 | 0 | 1 |

Since $c = (0, 1, 0, 0)'$ does not belong to the rowspace of $A$, treatment effects or counterfactual means among compliers are not outcome-nonrestrictive identified. Similarly, the proportion of compliers is not identified.[10] However, it is straightforward to see that if one maintains the assumption that compliers and defiers share the same average treatment effect, then the average treatment effect among compliers becomes identified and is equal to the conventional Wald ratio (Angrist et al., 1996) $(\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0])/(\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0])$. This example demonstrates that a parameter like $\mathbb{E}[Y_i(1) - Y_i(0)|G_i = comp.]$ can in general be identified even when $P(G_i = comp.)$ and $\mathbb{E}[\{Y_i(1) - Y_i(0)\} \cdot \mathbb{1}(G_i = comp.)]$ are not, if restrictions are imposed on outcomes.[11] Theorem 2 shows that this however cannot occur when identification is outcome-nonrestrictive.

[10] The difference $\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0]$ instead identifies $P(G_i = comp.) - P(G_i = def.)$. We can also identify the quantities $\{P(G_i = comp.) + P(G_i = a.t.)\}$, $\{P(G_i = def.) + P(G_i = a.t.)\}$, $\{P(G_i = comp.) + P(G_i = n.t.)\}$ and $\{P(G_i = def.) + P(G_i = n.t.)\}$, but not $P(G_i = comp.) + P(G_i = def.)$.

[11] Note further that although we can also write $\mathbb{E}[Y_i(1) - Y_i(0)|G_i = comp.] = \frac{\mathbb{E}[Y_i(1) \cdot \mathbb{1}(G_i = comp.)]}{\mathbb{E}[\mathbb{1}(G_i = comp.)]} - \frac{\mathbb{E}[Y_i(0) \cdot \mathbb{1}(G_i = comp.)]}{\mathbb{E}[\mathbb{1}(G_i = comp.)]}$ none of the quantities $\mathbb{E}[Y_i(1) \cdot \mathbb{1}(G_i = comp.)]$, $\mathbb{E}[Y_i(0) \cdot \mathbb{1}(G_i = comp.)]$, or $\mathbb{E}[\mathbb{1}(G_i = comp.)]$ are identified in isolation, even with the outcome restriction that $\mathbb{E}[Y_i(1) - Y_i(0)|G_i = comp.] = \mathbb{E}[Y_i(1) - Y_i(0)|G_i = def.]$.

### D.2 Relationship to Heckman and Pinto (2018)

Theorem T-2 of Heckman and Pinto, 2018 (henceforth HP) states, in my notation, that the following hold:

1. If $c'(I-({A^{[t]}}^+A^{[t]})=\mathbf{0}$ for a vector $c \in \mathbb{R}^{|\mathcal{G}|}$, then $\sum_g c_g \cdot \mathbb{E}[Y_i(t)|G_i=g]$ is identified.
2. If $c'K_T=\mathbf{0}$ for a vector $c \in \mathbb{R}^{|\mathcal{G}|}$, then $\sum_g c_g \cdot P(G_i=g)$ is identified.

where ${A^{[t]}}^+$ is the Moore-Penrose pseudo-inverse of the matrix $A^{[t]}$, and $K_T = I-(\mathbf{A}^+\mathbf{A})$ with $\mathbf{A}$ a matrix that stacks $A^{[t]}$ row-wise over the $t \in \mathcal{T}$.

A property of the Moore-Penrose pseudo inverse is that the matrix $({A^{[t]}}^+A^{[t]})$ projects onto the rowspace of $A^{[t]}$, and similarly $(\mathbf{A}^+\mathbf{A})$ projects onto the rowspace of $\mathbf{A}$. The condition $c \in rs(A^{[t]})$ implies that $c \in rs(\mathbf{A})$, and thus combining elements 1. and 2. above we have by Theorem T-2 that $c \in rs(A^{[t]})$ implies that both the numerator and the denominator of the RHS of Eq. (4), $\frac{\sum_g c_g \cdot \mathbb{E}[Y_i(t)|G_i=g]}{\sum_g c_g \cdot P(G_i=g)}$, are identified. Thus my Theorem 1 for a parameter of the form $\mu_c^t$ can be seen as a corollary of Theorem T-2 of HP, in the case that the vector $c$ is binary-valued so that $\frac{\sum_g c_g \mathbb{E}[Y_i(t)|G_i=g]}{\sum_g c_g P(G_i=g)}$ can be interpreted as a *conditional* counterfactual mean $\mu_c^t = \mathbb{E}[Y_i(t)|c(G_i)=1]$ (when $c \in \mathbb{R}^{|\mathcal{G}|}$ generally, $\frac{\sum_g c_g \mathbb{E}[Y_i(t)|G_i=g]}{\sum_g c_g P(G_i=g)}$ may not have an interpretation as a single conditional mean). While HP apply their Theorem T-2 to consider parameters of the form $\mu_c^t$ in the specific setting of the LATE model, they do not appear to highlight the general importance of binary-valued $c$ for their Theorem T-2.

## E Algorithms to enumerate outcome-nonrestrictive identification results

**Algorithm 1:**

Begin with a given instrument support $\mathcal{Z}$ and set of treatments $\mathcal{T}$, and $t' \neq t$ in $\mathcal{T}$:

1. Loop over all possible choice models $\mathcal{G}$ given $\mathcal{Z}$ and $\mathcal{T}$. There are $2^{|\mathcal{G}^m|=2^{|\mathcal{T}|^{|\mathcal{Z}|}}}$ of these, where we let $\mathcal{G}^m$ denote the set of all $|\mathcal{T}|^{|\mathcal{Z}|}$ conceivable response types (mappings from $\mathcal{Z}$ to $\mathcal{T}$)
2. Given the results of Section 4.2, find a basis for the left null-space $ns(A^{[t',t]})$ of $A^{[t',t]} := \begin{bmatrix} A^{[t']} \\ A^{[t]} \end{bmatrix}$ via a QR decomposition of $A^{[t',t]}$. Represent this basis by a $k \times 2|\mathcal{Z}|$ matrix $N^{[t',t]}$, where $k$ is the dimension of $ns(A^{[t',t]})$. For any vector $\alpha \in ns(A^{[t',t]'})$, let $\alpha_1(\alpha) = [I_{|\mathcal{Z}|}, \mathbf{0}_{|\mathcal{Z}|\times|\mathcal{Z}|}]\alpha$ be its first $|\mathcal{Z}|$ components, and let $\mathcal{C}^{[t,t']} = \{A^{[t']'}\alpha_1(\alpha) : \alpha \in ns(A^{[t',t]'})\}$ be the subspace of $\mathbb{R}^{|\mathcal{G}|}$ corresponding to these $\alpha$. $\mathcal{C}^{[t,t']}$ is a k-dimensional vector space with a basis represented by the $k \times |\mathcal{G}|$ matrix $B^{[t',t]} := A^{[t']'}N^{[t',t]}[I_{|\mathcal{Z}|}, \mathbf{0}_{|\mathcal{Z}|\times|\mathcal{Z}|}]$.
3. If $k \geq 1$, we now determine the intersection of $\mathcal{C}^{[t,t']}$ with the unit cube. This

is done by looping over the $2^{|\mathcal{G}|}-1$ non-zero vectors $c$ in $\{0,1\}^{|\mathcal{G}|}$, and checking whether $c \in \mathcal{C}^{[t,t']}$ (when $B$ has full row rank, this can be done e.g. by checking that $B^{[t',t]^+}B^{[t',t]}c = c$, where $B^+$ is the Moore-Penrose pseudo-inverse of $B$).

Note that since the computational problem as a whole is symmetric with respect to permutations of the (arbitrary) treatment labels, we can focus on binary collections containing the two treatment values $t' = 1$ and $t = 0$, and then generate new binary collections by then applying all re-labelings to the treatment values.

| $\lvert\mathcal{T}\rvert$ | $\lvert\mathcal{Z}\rvert$ | $\lvert\mathcal{C}_{\lvert\mathcal{Z}\rvert}\rvert$ | # $\alpha$'s (i.e. $(\lvert\mathcal{C}_{\lvert\mathcal{Z}\rvert}\rvert)^{2\lvert\mathcal{Z}\rvert}$) | $\lvert\mathcal{G}^m\rvert = \lvert\mathcal{T}\rvert^{\lvert\mathcal{Z}\rvert}$ | # selection models (i.e. $2^{\lvert\mathcal{G}^m\rvert}$) |
|---|---|---|---|---|---|
| 2 | 2 | 3 | 81 | 4 | 16 |
| 3 | 2 | 3 | 81 | 8 | 256 |
| 2 | 3 | 7 | 117,649 | 9 | 512 |
| 3 | 3 | 7 | 117,649 | 27 | $1.34{\cdot}10^8$ |
| 4 | 3 | 7 | 117,649 | 81 | $2.42{\cdot}10^{24}$ |
| 4 | 4 | 16 | $4.29 \cdot 10^9$ | 256 | $1.16{\cdot}10^{77}$ |

**Table E.1:** Comparison of the computational complexity of Algorithms 1 and 2

Table E.1 compares the complexity of Algorithms 1 and 2. It does not account for the full computational cost of running each algorithm (e.g. computations within each choice of $\alpha$ in the case of Algorithm 2, or within a selection model in the case of Algorithm 1), but it is nevertheless clear that Algorithm 1 quickly becomes infeasible, while there remains hope for Algorithm 2 with $|\mathcal{Z}| = |\mathcal{T}| = 4$.

Algorithm 2:

Begin with a given instrument support $\mathcal{Z}$ and set of treatments $\mathcal{T}$.

**Part One: generate binary collections by $\alpha$**

1. Loop over all vectors $2 \cdot |\mathcal{Z}|$-component vectors $\alpha$ having components in the set $\mathcal{C}_{|\mathcal{Z}|}$ (there are $(|\mathcal{C}_{|\mathcal{Z}|}|)^{2|\mathcal{Z}|}$ of these)
2. With $t' = 1$ and $t = 0$ fixed (as with Algorithm 1), construct the matrix $A^{[t',t]} := \begin{bmatrix} A^{[t']} \\ A^{[t]} \end{bmatrix}$ where now each of $A^{[t']}$ and $A^{[t]}$ representing the full set of conceivable response types $\mathcal{G}^m$ (having $|\mathcal{G}^m| = |\mathcal{T}|^{|\mathcal{Z}|}$ columns). Compute for each $\alpha$ the row vector $\alpha' A^{[t',t]}$.
3. Consider the columns $g$ of $\alpha' A^{[t',t]}$ that take the value of 0, and call this set $\mathcal{G}^0(\alpha)$. Note that $\mathcal{G}^0(\alpha)$ is the set of $g$ for which $[\alpha_1(\alpha)' A^{[t']}]_g = [\alpha_0(\alpha)' A^{[t]}]_g$ (using the notation introduced in Algorithm 1).
4. Now find the set $\mathcal{G}(\alpha) \subseteq \mathcal{G}^0(\alpha)$ such that $[\alpha_1(\alpha)' A^{[t']}]_g \in \{0,1\}$ for all $g \in \mathcal{G}(\alpha)$.

Only response types $g$ in the set $\mathcal{G}^0(\alpha)$ can exist in a binary collection having $\alpha^{[t]} = \alpha_0(\alpha)$ and $\alpha^{[t']} = \alpha_1(\alpha)$. Further, the set $\mathcal{G}(\alpha)$ is *maximal* (given $\alpha$) in the sense that we get a binary collection from $\alpha$ for $t, t'$ for any selection model $\mathcal{G} \subseteq \mathcal{G}(\alpha)$.

5. Some of the binary collections (indexed by $\alpha$) constructed in this way will be redundant in the following sense. Define $c(\alpha) = \alpha_1(\alpha)' A^{[t']}$, and let vectors $\alpha$ and $\beta$ be two $2|\mathcal{Z}|$-component vectors such that $c(\alpha) = c(\beta)$ but $\mathcal{G}(\alpha) \subset \mathcal{G}(\beta)$. Then $\beta$ delivers the same largest complier group as $\alpha$ but while allowing for a strictly larger selection model. In this case remove $\alpha$, since the identification result for $\beta$ nests that of $\alpha$. If $c(\alpha) = c(\beta)$ as above and $\alpha$ and $\beta$ deliver *the same* maximal selection model, i.e. $\mathcal{G}(\alpha) = \mathcal{G}(\beta)$, then drop whichever vector has more non-zero elements than the other, i.e. drop $\alpha$ if $||\alpha||_0 > ||\beta||_0$ where $||\cdot||_0$ indicates the $\ell_0$ norm. If $||\alpha||_0 = ||\beta||_0$, then keep whichever vector has a smaller $\ell_2$ norm is kept (this choice is arbitrary).

**Part Two: organize by selection model and pare redundancies**

1. Extend the binary collections obtained in Part One of the algorithm for $(t', t) = (1, 0)$ to all other choices of $t' > t$. Binary collections can now be indexed by the tuple $(t', t, \alpha)$. Any $(1, 0, \alpha)$ obtained in Part One above yields a binary collection for $(t', t, \alpha)$ with the same vector $\alpha$, with the response types suitably re-defined based on relabeling the treatment values.

2. Now collect all binary combinations that share a maximal selection model $\mathcal{G}$, which based on the last step may allow treatment effects that contemplate differing treatment contrasts (e.g. treatment value 2 vs. 0 or treatment 1 vs. 0) to be associated with the same selection model.

3. We now have a list of selection models $\mathcal{G}$ that admit of at least one binary collection, and for each such $\mathcal{G}$ a list of these binary collections. Recall that each selection model can be expressed by the matrix $A$. To distill out selection models with a unique structure, eliminate any redundancies where one selection model can be transformed into another by re-labeling treatment values, or by permuting the labels of the instrument values and re-ordering the columns of $A$.

# F Illustrative examples from the brute force search

### F.0.1 Binary treatment and binary instrument

With a binary treatment and binary instrument, the brute-force search reveals that there are exactly two choice models that admit of outcome-nonrestrictive identification of treat-

ment effects. The first is the classic LATE model of Imbens and Angrist (1994) (Example 1), and the second is Example 2 from the main text.

Consider first Example 1. Row reduction of the matrices $A^{[1]}$ and $A^{[0]}$ given in Example 1 in Section 3 yields:

$$rs(A^{[1]}) = span\left\{\begin{pmatrix}0\\1\\0\end{pmatrix}, \begin{pmatrix}0\\0\\1\end{pmatrix}\right\} \quad \text{and} \quad rs(A^{[0]}) = span\left\{\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\0\\1\end{pmatrix}\right\}$$

This leads to the two planes depicted in Figure 1.

*Remark:* In the binary-binary LATE model, the rank $k$ of $A^{[t]}$ is $k = |\mathcal{Z}| = 2$ for either $t = 0$ or $t = 1$, and in either case $|rs(A^{[t]}) \cap \{0,1\}^n| = 2$. This does not meet the upper bound of $2^k = 4$ from Melo and Winter (2019) (see Footnote 5). However the result does imply that there can be no more than $2^{|\mathcal{Z}|}$ binary combinations, even though typically $2^{|\mathcal{Z}|} < 2^{|\mathcal{G}|}$ and there are $2^{|\mathcal{G}|}$ potential values of $c$ to consider ex-ante.

In the case of example 2,[12] the matrix $A$ becomes:

| | compliers | defiers |
|---|---|---|
| $\mathbf{z} = \mathbf{0}$ | 0 | 1 |
| $\mathbf{z} = \mathbf{1}$ | 1 | 0 |

The rowspaces of $A^{[1]}$ and $A^{[0]}$ are the same and both span $\mathbb{R}^2$: $rs(A^{[1]}) = rs(A^{[0]}) = span\left\{\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}\right\}$. Thus, given the results of Section 3, we know that treatment effect parameters that are outcome-nonrestrictive identified correspond to any non-zero vertex of the unit cube in $\mathbb{R}^2$, as depicted in Figure 4 below. Note that by Theorem 1, $\mathbb{E}[D_i|Z_i = 1]$ and $1 - \mathbb{E}[D_i|Z_i = 0]$ are both measures for the same population parameter $P(c(G_i) = 1) = P(i \text{ is complier})$. Thus $\mathbb{E}[D_i|Z_i = 1] = 1 - \mathbb{E}[D_i|Z_i = 0]$ can be used as an overidentification restriction for this choice model (there is no such restriction for the LATE model that simply rules out defiers).

### F.0.2 3 treatments, binary instrument

Now suppose that the instrument is binary and $\mathcal{T} = \{0,1,2\}$. With no restrictions on selection behavior, there are $3^{|\mathcal{Z}|} = 9$ conceivable response types. Table 1 reports that in this case there are five selection models that afford a total of five distinct outcome-nonrestrictive identification results. These results are all listed in Appendix K.

As an example that may be empirically relevant, consider a selection model that has $T(z)$ increasing in $z$, but rules out always-1 takers and individuals that skip from $t = 0$

[12] Goff and Lee (2024) apply this choice model to study the effect of an NFL team deferring the kickoff on the game outcome, with the kickoff coin flip that decides which team is given the option to defer as the instrument. If it is common knowledge between the teams whether receiving the kickoff is beneficial in that particular game, then a simple model of optimizing play would predict that each game will either be a "complier" or a "defier".

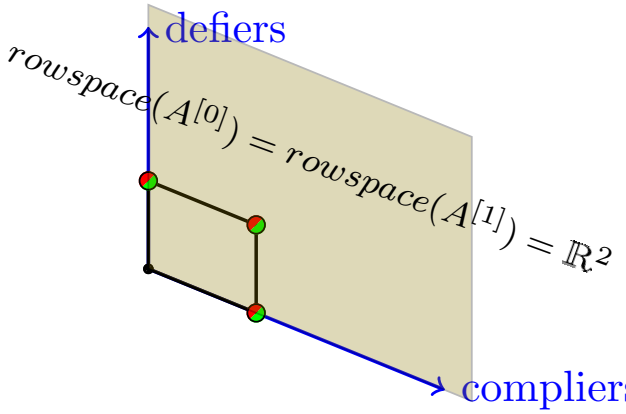


**Figure 4:** A geometric depiction of the model with compliers and defiers only. The vectors $c = (0,1)'$, $c = (1,0)'$ and $c = (1,1)'$ all belong to both $rs(A^{[1]})$ and $rs(A^{[0]})$ and hence the LATE for either response type or the ATE are identified. As in Figure 4, the split-shading of a given vertex (red/green in color) of the unit square indicates that it lies in $rs(A^{[0]}) \cap rs(A^{[1]})$ and is not equal to the zero vector.

to $t = 2$ when $z$ is increased from 0 to 1. This restriction leaves four response types: always-0 takers, always-2 takers, individuals who move from treatment 0 to treatment 1, and individuals who move from treatment 1 to treatment 2. SM.3.2.1 reported in Appendix K reveals that two treatment effects $\Delta_c^{t,t'}$ are identified in this choice model. For example, the quantity

$$\frac{\mathbb{E}[Y_i \cdot D_i^{[1]} | Z_i = 1]}{\mathbb{E}[D_i^{[1]} | Z_i = 1]} - \frac{\mathbb{E}[Y_i \cdot D_i^{[0]} | Z_i = 0] - \mathbb{E}[Y_i \cdot D_i^{[0]} | Z_i = 1]}{\mathbb{E}[D_i^{[0]} | Z_i = 0] - \mathbb{E}[D_i^{[0]} | Z_i = 1]}$$

corresponds to the binary collection with $\alpha_0 = (0,1)'$ and $\alpha_1 = (1,-1)'$, and identifies $\mathbb{E}[Y_i(1) - Y_i(0) | T_i(0) = 0, T_i(1) = 1]$. The treatment effect $\mathbb{E}[Y_i(2) - Y_i(1) | T_i(0) = 1, T_i(1) = 2]$ is identified by a similar estimand.

This selection model could represent a setting in which $t = 2$ represents passing a test outright, $t = 1$ passing the test "provisionally", and $t = 0$ failing. Suppose that a reform implemented for some schools lowers the score threshold $\tau_o$ for an outright pass to the old threshold $\tau_p$ for a provisional pass, while further lowering $\tau_p$, as depicted below:

$$\begin{array}{lccccc} (\mathbf{z = 0}) & \longleftarrow & \tau_p(0) & \longrightarrow & \tau_o(0) & \longrightarrow \\ (\mathbf{z = 1}) & \longleftarrow \ \tau_p(1) & \longrightarrow & \tau_o(1) & \longrightarrow & \end{array}$$

Students will then belong to one of the four types described above, depending on their test score. The quantity $\mathbb{E}[Y_i(1) - Y_i(0) | T_i(z) = z]$ represents the average effect of moving from a fail to a provisional pass among the students who are brought in to a provisional pass by the grading reform.

*The selection model of Kline and Walters (2016):* Another observation from the $|\mathcal{T}| = 3, |\mathcal{Z}| = 2$ setting is that the selection model of Kline and Walters (2016) (KW) does not appear in the catalog of Appendix K. KW study a setting in which the binary instrument is an offer to choose $t = 2$, while treatments $t = 0$ and $t = 1$ are always available even

if $z = 0$. On the grounds of revealed preference, KW impose that $T_i(1) \neq T_i(0) \implies T_i(1) = 2$ which results in a selection model with five response types. KW show that the parameter $\mathbb{E}[Y_i(2) - Y_i(T_i(0))|T_i(1) = 2, T_i(0) \neq 2]$ is then identified. The quantity $T_i(0)$ represents an individual's next preferred alternative to $t = 2$, which may vary across those $i$ for whom $T_i(1) = 2, T_i(0) \neq 2$. As a result, this parameter does not fit the form of the general family of treatment effect parameters $\Delta_c^{t,t'}$ introduced in Section A. Indeed, the brute force search confirms that unfortunately *no* parameters of the form $\Delta_c^{t,t'}$ between two fixed treatmets $t$ and $t'$ are identified in the KW selection model.

### F.0.3 Binary treatment, 3 instrument values

Suppose now that treatment is binary and $\mathcal{Z} = \{0, 1, 2\}$. Table 1 reports that in this case there are 11 selection models that afford a total of 30 distinct outcome-nonrestrictive identification results, listed in Appendix K. One observation that emerges when we extend the analysis to instruments that take more than two values is that, there now exist binary collections that require coefficients $\alpha_z$ that do not belong to the set $\{-1, 0, 1\}$. Although this is entirely consistent with Proposition 2 for $|\mathcal{Z}| \geq 3$, a reasonable conjecture ex-ante might have been that $\alpha_z \in \{-1, 0, 1\}$ always holds, given the preponderance of this pattern in known identification results (see for example all of the results surveyed in Appendix G).

For the sake of exposition, let us consider a "judge-IV" setting in which defendants $i$ receive a bail decision from a randomly-assigned judge $z$, with $t = 1$ indicating that the defendent remains incarcerated and $t = 0$ that they are released on bail. Suppose that the defendants are one of three types $g \in \mathcal{G}$ (typically unobserved to the researcher), comprising the columns of the table below. "Prepared" defendants dress formally and speak politely in their bail hearing, perhaps also presenting evidence that they are not a danger to the community. "Unprepared" defendants do not make such efforts. A third category of "flight-risk" defendants are thought to be particularly capable of and likely to fail to appear for trial if they are granted bail (while this is not true of the first two groups, e.g. due to strong personal ties to the jurisdiction or insufficient financial means to leave town).

| | prepared | unprepared | flight-risk |
|---|---|---|---|
| $\mathbf{z = 0}$ (standard) | 0 | 0 | 1 |
| $\mathbf{z = 1}$ (character) | 0 | 1 | 0 |
| $\mathbf{z = 2}$ (skeptics) | 1 | 0 | 1 |

The above table summarizes selection behavior when the judges also belong to one of three types, represented across rows. "Standard" judges are only concerned with failure to appear, and keep only the flight-risk defendants incarcerated. "Character" judges instead attempt to infer the risk of a defendant to public safety on the basis of the defendant's presentation and arguments to their character made in the bail hearing, but

do not attempt to assess whether the defendant is likely to skip town. "Skeptic" judges are also sensitive to judgments about presentation, but in the opposite direction: they are suspicious of defendants precisely when they seem to be making a case that they are not dangerous. They deny bail for the prepared defendants, and also deny bail for flight-risk defendants.[13]

Note that this model does not satisfy the strong LATE monotonicity assumption typically invoked in judge-IV settings, which has been challenged on empirical grounds (Frandsen et al., 2023; Sigstad, 2023). If there were no skeptic ($z = 2$) judges, then this model would instead consist of compliers, defiers, and never-takers, which we have already seen permits no outcome-nonrestrictive identification results for treatment effects. However, the presence of the skeptics aids here in identification, as we can then identify the average effect of incarceration among two groups $\mathbb{E}[Y_i(1)-Y_i(0)|G_i \in \{unprepared, flight\text{-}risk\}]$ by $\frac{\mathbb{E}[Y_iD_i|Z_i=0]+\mathbb{E}[Y_iD_i|Z_i=1]}{\mathbb{E}[D_i|Z_i=0]+\mathbb{E}[D_i|Z_i=1]} - \frac{-\mathbb{E}[Y_iD_i|Z_i=0]+\mathbb{E}[Y_iD_i|Z_i=1]+2\cdot\mathbb{E}[Y_iD_i|Z_i=2]}{-\mathbb{E}[D_i|Z_i=0]+\mathbb{E}[D_i|Z_i=1]+2\cdot\mathbb{E}[D_i|Z_i=2]}$, corresponding to the binary collection with $\alpha_1 = (1, 1, 0)$ and $\alpha_0 = (-1, 1, 2)'$.[14] Note that using this result requires judge types to be observable, or estimable from judges each seeing many cases.

### F.0.4 3 treatment values, 3 instrument values

In the case of $\mathcal{Z} = \mathcal{T} = \{0, 1, 2\}$, the brute force approach returns 251 distinct binary collections spread across 251 unique selection models. These results nest for example two identification results presented in Kirkeboen, Leuven and Mogstad (2016) (KLM). KLM consider an unordered treatment which represents a student's field of study, where students are "assigned" to a given field, i.e $Z_i = j$ represents an incentive to choose field $j$. Proposition 2 of KLM presents three special cases in which a two stage least squares estimand with indicators for treatments 1 and 2 instrumented by indicators for $Z_i = 1$ and $Z_i = 2$ recovers causally interpretable coefficients. While their first result (restricting treatment effects to be homogeneous) is not outcome-nonrestrictive, the other two results are.

For example, the second result in KLM Proposition 2 shows that if preferences are further restricted so that $D_i^{[2]}(1) = D_i^{[2]}(0)$ and $D_i^{[1]}(2) = D_i^{[1]}(0)$ for all $i$ (an offer to one program does not affect whether or not the student chooses the other program), then $\mathbb{E}[Y_i(1) - Y_i(0)|D_i^{[1]}(1) > D_i^{[1]}(0)]$ and $\mathbb{E}[Y_i(2) - Y_i(0)|D_i^{[2]}(2) > D_i^{[2]}(0)]$ are each identified.[15] Let us consider how the first of these results appears in the comprehensive search (the second result proceeds similarly). Upon a relabeling of treatment/instrument values and removing one response type,[16] selection model SM.3.3.63 in the cat-

[13] This selection model is equivalent to SM.2.3.4 in Appendix K, after permuting instrument/treatment labels. Note that this model is merely illustrative: more types and nuance in their definitions could add some realism.

[14] A coefficient of two is inevitable for $t = 0$, since we need $\alpha_1 = 1$ (using the $\alpha_z$ notation) for $c(flightrisk) = 0$, $\alpha_0 = -\alpha_1$ to get $c(prepared) = 0$, but can only achieve $c(unprepared) = 1$ if $\alpha_2 + \alpha_1 = 1$.

[15] Throughout, KLM also maintain a version of unordered monotonicity (cf. Heckman and Pinto 2018) in which $D_i^{[1]}(1) \geq D_i^{[1]}(0)$ and $D_i^{[2]}(2) \geq D_i^{[2]}(0)$: an offer of admission never causes a student to select out of that field.

[16] In particular, label the treatments $(0, 1, 2)$ as $(1, 0, 2)$, swap instrument values 1 and 2, and drop column 6. All results for the $3 \times 3$ case are enumerated in a working paper version of this paper: `https://arxiv.org/abs/2406.02835`.

alog amounts to the following: $A = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 2 & 0 \\ 0 & 1 & 0 & 1 & 1 & 2 & 2 \\ 0 & 0 & 2 & 2 & 1 & 2 & 1 \end{bmatrix}$. This selection model has seven response types, whereas the choice model considered by KLM contains only the first six columns of A. In the larger selection model with all seven groups, the treatment effect $\mathbb{E}[Y_i(1) - Y_i(0)|T_i(1) \neq T_i(0)]$ is identified by the binary collection with $\alpha_0 = (-2, 1, 1)'$ and $\alpha_1 = (1, -1, 0)'$.[17] Thus, we have seen that KLM's choice model can be relaxed to allow an additional response type, with the same estimand that identifies $\mathbb{E}[Y_i(1) - Y_i(0)|D_i^{[1]}(1) > D_i^{[1]}(0)]$ in their more restrictive model identifying $\mathbb{E}[Y_i(1) - Y_i(0)|T_i(1) \neq T_i(0)]$ more generally. Code available from the author allows one to check in general whether a given selection model can be relaxed in this way, using the catalog of identification results (available for $|\mathcal{T}|, |\mathcal{Z}| \leq 3$).

### F.1 4 treatment values, 4 instrument values: spillover effects within pairs

Although the $|\mathcal{T}| = |\mathcal{Z}| = 4$ case is not included in the brute-force search of Table 1 (due to the computational burden), it remains easy to check for outcome-nonrestrictive identification in any given choice model using the results of Section 3. This section presents an alternative application in the $4 \times 4$ case to supplement the study of interaction effects from Section 5.

Consider a setting in which each unit $i$ has one "neighbor" $n(i)$, and we allow for violations of SUTVA within neighbor pairs $(i, n(i))$. This can be accommodated by expanding the set of treatments $\mathcal{T}$ to accommodate values of such pairs, so that $Y_i = Y_i(T_i, T_{n(i)})$ where $T_{n(i)}$ is the treatment of the neighbor of unit $i$, indexed by $n(i)$. I consider the case in which treatment $T$ itself is binary, so that $\tilde{T}_i := (T_i, T_{n(i)})$ may take one of four values $(0,0), (1,0), (0,1), (1,1)$. Following the notation in Section 5, we denote these pair-level "treatments" as $0, A, B, C$, where $\tilde{T}_i = 0$ indicates that neither unit is treated, $\tilde{T}_i = A$ that only unit $i$ is treated, $\tilde{T}_i = B$ that only their neighbor is treated, and $\tilde{T}_i = C$ that both $i$ and their neighbor is treated.

For each $i$, let $Z_i$ be a binary instrument that reflects whether $i$ is "assigned" to receive treatment. See Kang and Imbens (2016) and Vazquez-Bare (2023) for related setups. Let $\tilde{T}_i(z, z')$ reflect the treatments for the pair as a function of treatment assignments $(z, z')$ for the pair. Let $\tilde{Z}_i := (Z_i, Z_{n(i)})$ be the pairs realized treatment assignment, which can take any of four counterfactual values $z \in \tilde{\mathcal{Z}} = \{0, A, B, C\}$. I maintain throughout two assumptions about selection: i) first, that the $T_i$ component of $\tilde{T}_i(z, z')$ only depends on $z$, and that the $T_{n(i)}$ component of $\tilde{T}_i(z, z')$ only depends on $z'$; and ii) secondly, that each selection uptake is monotonic such that $T_i(1) \geq T_i(0)$, where we write $T_i(z, z')$ as $(T_i(z), Y_{n(i)}(z'))$.

[17] Accordingly, $\mathbb{E}[Y_i(1) - Y_i(0)|T_i(1) \neq T_i(0)] = \frac{\mathbb{E}[Y_i \cdot D_i^{[1]}|Z_i=2] + \mathbb{E}[Y_i \cdot D_i^{[1]}|Z_i=1] - 2 \cdot \mathbb{E}[Y_i \cdot D_i^{[1]}|Z_i=0]}{\mathbb{E}[D_i^{[1]}|Z_i=2] + \mathbb{E}[D_i^{[1]}|Z_i=1] - 2 \cdot \mathbb{E}[D_i^{[1]}|Z_i=0]} - \frac{\mathbb{E}[Y_i \cdot D_i^{[0]}|Z_i=1] - \mathbb{E}[Y_i \cdot D_i^{[0]}|Z_i=0]}{\mathbb{E}[D_i^{[0]}|Z_i=1] - \mathbb{E}[D_i^{[0]}|Z_i=0]}$. Identification of $\mathbb{E}[Y_i(2) - Y_i(0)|T_i(2) \neq T_i(0)]$ is analogous.

These restrictions leave nine response types, enumerated in the table below:

| **assigned** ↓ | (nt,nt) | (cm,cm) | (cm,nt) | (nt,cm) | (at,at) | (at,cm) | (cm,at) | (nt,at) | (at,nt) |
|---|---|---|---|---|---|---|---|---|---|
| 0=(0,0) | 0 | 0 | 0 | 0 | C | A | B | B | A |
| A=(1,0) | 0 | A | A | 0 | C | A | C | B | A |
| B=(0,1) | 0 | B | 0 | B | C | C | B | B | A |
| C=(1,1) | 0 | C | A | B | C | C | C | B | A |

Given a function $c(\cdot)$, the local average direct effect of one's own treatment on their outcome is:

$$\begin{aligned} LADTE &:= \mathbb{E}[Y_i(1, Z_{n(i)}) - Y_i(0, Z_{n(i)})|c(G_i) = 1] \\ &= P(Z_{n(i)} = 1) \cdot \mathbb{E}[Y_i(C) - Y_i(B)|c(G_i) = 1] + P(Z_{n(i)} = 0) \cdot \mathbb{E}[Y_i(A) - Y_i(0)|c(G_i) = 1] \end{aligned}$$

using that $P(Z_{n(i)} = 1|c(G_i) = 1) = P(Z_{n(i)} = 1)$ by independence.

Similarly, the local average spillover (indirect) effect is

$$\begin{aligned} LASTE &:= \mathbb{E}[Y_i(Z_i, 1)) - Y_i(Z_i, 0)|c(G_i) = 1] \\ &= P(Z_i = 1) \cdot \mathbb{E}[Y_i(C) - Y_i(A)|c(G_i) = 1] + P(Z_i = 0) \cdot \mathbb{E}[Y_i(B) - Y_i(0)|c(G_i) = 1] \end{aligned}$$

Since the distributions of $Z_i$ and $Z_{n(i)}$ are identified, we can then point identify the LADTE and the LASTE provided that we can identify $\mu_t^c$ for all of $t \in \{0, A, B, C\}$. An application of Theorems 1 and 2 shows that this is possible without restricting outcomes in the above selection model if and only if $c(g) = \mathbb{1}(g = cm, cm)$.

This can be shown by direct enumeration of the $2^9$ possible functions $c : \mathcal{G} \to \{0, 1\}$. The vector forms $\alpha_t$ of the resulting coefficient functions $\alpha^{[t]}(z)$ are:

$$\begin{aligned} \alpha_0 = (+1, -1, -1, +1)', \quad \alpha_A = (-1, +1, +1, -1)' \\ \alpha_B = (-1, +1, +1, -1)', \quad \alpha_C = (+1, -1, -1, +1)' \end{aligned}$$

Similar to the case of complementarities, the selection model therefore implies the overidentification restriction that

$$\begin{aligned} p &:= P(T_i = 0|Z_i = (1,1)) - P(T_i = 0|Z_i = (0,1)) - P(T_i = 0|Z_i = (1,0)) + P(T_i = 0|Z_i = (0,0)) \\ &= -P(T_i = A|Z_i = (1,1)) + P(T_i = A|Z_i = (0,1)) + P(T_i = A|Z_i = (1,0)) - P(T_i = A|Z_i = (0,0)) \\ &= -P(T_i = B|Z_i = (1,1)) + P(T_i = B|Z_i = (0,1)) + P(T_i = B|Z_i = (1,0)) - P(T_i = B|Z_i = (0,0)) \\ &= P(T_i = C|Z_i = (1,1)) - P(T_i = C|Z_i = (0,1)) - P(T_i = C|Z_i = (1,0)) + P(T_i = C|Z_i = (0,0)) \end{aligned}$$

for some value $p \in [0, 1]$, which identifies $P(g = \text{cm,cm})$.

# G Recovering existing identification results as binary combinations and collections

The notions of binary combinations and binary collections reveals a common structure among several existing IV point identification results.

## G.1 Example 1: LATE monotonicity and marginal treatment effects

Here I extend my analysis of the monotonicity assumption of Imbens and Angrist (1994) to cases with more than two instrument values. Treatment remains binary $\mathcal{T} = \{0, 1\}$. Since $D_i^{[0]}(z) = 1 - D_i^{[1]}(z)$, we can focus on the single treatment indicator $D_i(z) := D_i^{[1]}(z)$. Imbens and Angrist (1994) assume that:

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

Suppose $z, z'$ are a pair such that the former case of assumption IAM obtains, and define a binary combination with $K = 2$, $z_1 = z'$, $\alpha_1 = 1$, $z_2 = z$, $\alpha_2 = -1$. Then Eq. (4) from the main paper yileds

$$\mathbb{E}[Y_i(1)|D_i(z') > D_i(z)] = \frac{\mathbb{E}\left[Y_i \cdot D_i | Z_i = z'\right] - \mathbb{E}\left[Y_i \cdot D_i | Z_i = z\right]}{\mathbb{E}\left[D_i | Z_i = z'\right] - \mathbb{E}\left[D_i | Z_i = z\right]} \tag{19}$$

and similarly

$$\mathbb{E}[Y_i(0)|D_i(z') > D_i(z)] = \frac{\mathbb{E}\left[Y_i \cdot (1 - D_i) | Z_i = z'\right] - \mathbb{E}\left[Y_i \cdot (1 - D_i) | Z_i = z\right]}{\mathbb{E}\left[(1 - D_i) | Z_i = z'\right] - \mathbb{E}\left[(1 - D_i) | Z_i = z\right]}$$

Combining, we have that $\mathbb{E}[Y_i(1) - Y_i(0)|D_i(z') > D_i(z)] = \frac{\mathbb{E}[Y_i|Z_i=z'] - \mathbb{E}[Y_i|Z_i=z]}{\mathbb{E}[D_i|Z_i=z'] - \mathbb{E}[D_i|Z_i=z]}$, which is Theorem 1 of Imbens and Angrist (1994).

Suppose that $\mathcal{Z}$ is continuous and for all $u \in [0, 1]$ there exists a $z \in \mathcal{Z}$ such that $P(z) := P(D_i = 1|Z_i = z) = u$. Let $U_i = \inf_{z \in \mathcal{Z}}\{P(z) : D_i(z) = 1\}$. Given IAM, $U_i$ plays the role of $G_i$, indicating the "first" instrument value (as ordered by the propensity score function $P(z)$) at which $i$ would take treatment. For any given $u$, let $z$ be a point in $\mathcal{Z}$ such that $P(z) = u$ and take a sequence $z_j$ in $\mathcal{Z}$ such that $z_j \to z$ as $j \to \infty$. Taking the limit of Eq. (19) we have that:

$$\mathbb{E}[Y_i(1)|U_i = u] = \lim_{j \to \infty} \frac{\mathbb{E}\left[Y_i \cdot D_i | Z_i = z_j\right] - \mathbb{E}\left[Y_i \cdot D_i | Z_i = z\right]}{\mathbb{E}\left[D_i | Z_i = z_j\right] - \mathbb{E}\left[D_i | Z_i = z\right]} = \frac{d}{du}\mathbb{E}\left[Y_i \cdot D_i | P(Z_i) = u\right]$$

and similarly for $Y_i(0)$, allowing us to identify marginal treatment effects (cf. Heckman et al. (2006)).

## G.2 Example 2: Vector monotonicity (Goff 2024)

Goff (2024) considers a binary treatment and finite $\mathcal{Z} \subseteq \mathcal{Z}_1 \times \mathcal{Z}_2 \times \ldots \mathcal{Z}_J$, and the following monotonicity assumption.

**Assumption 2 (vector monotonicity).** *There exists an ordering $\geq_j$ on $\mathcal{Z}_j$ for each $j \in \{1 \dots J\}$ such that for all $z, z' \in \mathcal{Z}$, if $z \geq z'$ component-wise according to the $\{\geq_j\}$, then $D_i(z) \geq D_i(z')$ for all $i$.*

Theorem 1 of Goff (2024) shows that average counterfactual means are identified under vector monotonicity for groups defined by the condition $c(G_i, Z_i) = 1$, where $c$ satisfies a condition called "Property M" and $Z_i$ has full rectangular support. His Proposition 6 shows that Property M is equivalent to $c(G_i, Z_i) = \sum_{k=1}^{K} \alpha_k \cdot D_i^{[t]}(z_k(Z_i))$ where $K$ is an even number no greater than $J/2$ and $\alpha_k = (-1)^k$ with $z_{k+1}(z) \geq z_k(z)$ component-wise according to the orders $\geq_j$, for all $k$. In what follows I for simplicity focus on the special case of target parameters in which $c$ depends on $G_i$ only, and not additionally on $Z_i$. See Appendix H for a discussion of other parameters.

In the case of two binary instruments $J = 2$, there are six selection types compatible with vector monotonicity, with names introduced by Mogstad et al. 2021:

| | $Z_1$ comp. | $Z_2$ comp. | eager-comp. | reluctant-comp. | n.t. | a.t. |
|---|---|---|---|---|---|---|
| $\mathbf{z} = (\mathbf{0}, \mathbf{0})'$ | 0 | 0 | 0 | 0 | 0 | 1 |
| $\mathbf{z} = (\mathbf{1}, \mathbf{0})'$ | 1 | 0 | 1 | 0 | 0 | 1 |
| $\mathbf{z} = (\mathbf{0}, \mathbf{1})'$ | 0 | 1 | 1 | 0 | 0 | 1 |
| $\mathbf{z} = (\mathbf{1}, \mathbf{1})'$ | 1 | 1 | 1 | 1 | 0 | 1 |

With this table defining matrix $A^{[1]}$, some algebra shows that for example $c = (1, 0, 0, 1, 0, 0)'$ occurs in the rowspace of both $A^{[1]}$ and $A^{[0]}$. One way to see this is to work out the row reduced echelon forms of $A^{[1]}$ and $A^{[0]}$, which preserve their row-spaces and are:

$$rref(A^{[1]}) = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \qquad rref(A^{[0]}) = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

From the first row of each of the reduced echelon forms, we can see immediately that e.g. the average treatment effect among $Z_1$ and reluctant compliers is outcome-nonrestrictive identified. Adding all four rows we see that the average treatment effect among all of the four compliers types is outcome-nonrestrictive identified, what Goff (2024) calls the "all-compliers LATE". Goff (2024) shows how these and similar point identification results generalize to any number of instruments under vector monotonicity.

### G.3 Example 3: Unordered Monotonicity, Heckman and Pinto (2018)

Heckman and Pinto (2018) (HP) consider a finite $\mathcal{Z}$ and assume what they call *unordered monotonicity* (UM) for a multi-valued treatment:

**Assumption UM.** *For any $t \in \mathcal{T}$ and $z, z' \in \mathcal{Z}$, either $D_i^{[t]}(z) \geq D_i^{[t]}(z')$ for all $i$ or $D_i^{[t]}(z') \geq D_i^{[t]}(z)$ for all $i$.*

Given UM, let us fix a $t \in \mathcal{T}$ and label the points in $\mathcal{Z}$ as $z_m$ for $m = 1, 2 \ldots |\mathcal{Z}|$, where the points are labeled in increasing order of the propensity score for treatment $t$: $P_t(z_{m+1}) \geq P_t(Z_m)$, where we define $P_t(z) = \mathbb{E}[D_i^{[t]}|Z_i = z]$. The order is not important in the case of ties. Let $\Sigma_{ti} = |\{z \in \mathcal{Z} : T_i(z) = t\}|$ be the number of $z \in \mathcal{Z}$ for which $i$ takes treatment $t$. Note that $\Sigma_{ti}$ is exactly equal to $|\mathcal{Z}| - m + 1$ for the smallest $m$ such that $D_i^{[t]}(z_m) = 1$. Thus we have a binary combination for any treatment $t$ and value $s \in \{0, 1, \ldots |\mathcal{Z}|\}$: in particular $D_i^{[t]}(z_m) - D_i^{[t]}(z_{m-1}) \in \{0, 1\}$ for all $i$, and is equal to $i$ for those units having $\Sigma_{ti} = s$.

It then follows immediately from Eq. (4), as in the IAM case with a binary treatment (cf. Eq. 19), that $E[Y_i(t)|\Sigma_{ti} = s]$ is identified for any $s = 1 \ldots |\mathcal{Z}|$ as:

$$\mathbb{E}[Y_i(t)|\Sigma_{ti} = s] = \begin{cases} \frac{\mathbb{E}\left[Y_i \cdot D_i^{[t]}|Z_i = z_m\right] - \mathbb{E}\left[Y_i \cdot D_i^{[t]}|Z_i = z_{m-1}\right]}{\mathbb{E}\left[D_i^{[t]}|Z_i = z_m\right] - \mathbb{E}\left[D_i^{[t]}|Z_i = z_{m-1}\right]} & \text{if } s < |\mathcal{Z}| \\ \frac{\mathbb{E}\left[Y_i \cdot D_i^{[t]}|Z_i = z_1\right]}{\mathbb{E}\left[D_i^{[t]}|Z_i = z_1\right]} & \text{if } s = |\mathcal{Z}| \end{cases} \tag{20}$$

where $m = |\mathcal{Z}| - s + 1$.

This provides a simple proof of HP's Theorem T-6, which shows that $E[Y_i(t)|\Sigma_{ti} = s]$ is identified. HP show that

$$E[Y_i(t)|\Sigma_{ti} = s] = \frac{c' {A^{[t]}}^+ Q_Z}{c' {A^{[t]}}^+ P_Z} \tag{21}$$

where $B^+$ is the Moore-Penrose pseudo-inverse of a matrix $B$, $Q_Z$ is a vector of $\mathbb{E}[Y_i D_i^{[t]}|Z_i = z]$ across $z$ and $P_Z$ is a vector of $\mathbb{E}[D_i^{[t]}|Z_i = z]$ across $z$. Here $c$ corresponds to our parameter of interest (indexed by the pair $(t, s)$), with an entry of one if $\Sigma_{tg} = s$ for that selection type and zero otherwise.

To see the equivalence between this result and (20), we can take advantage of the structure of $A^{[t]}$ under UM to replace it with a smaller matrix whose inverse is very simple. Note that any two selection types $g$ sharing a value of $\Sigma_{ti}$ will have identical entries in $c$, and will have identical corresponding columns in the matrix $A^{[t]}$. This implies that they will have identical rows in ${A^{[t]}}^+$. We can remove the redundant columns of $A^{[t]}$ by indexing columns by values of $\Sigma_{ti}$ rather than by full response vectors $g$, and similarly indexing elements of $c$ by values of $\Sigma_{ti}$. This yields the same vector $c' {A^{[t]}}^+$ as before, up to a scalar factor that counts the number of values of $g$ such that $\Sigma_{ti} = s$. However, this factor cancels out in the numerator and denominator of (21). With this modification, $A^{[t]}$ is now a $|\mathcal{Z}|$ by $|\mathcal{Z}| + 1$ matrix and $c$ is now a standard basis vector equal to one in its $s^{th}$ element (and zero elsewhere).

Let us now order the rows of this modified $A^{[t]}$ according to $z_1$, $z_2$, etc, and it's columns in decreasing order of $\Sigma_{ti}$. With this ordering, $A^{[t]}$ is simply a lower triangular matrix of ones, appended to the right by a single column of zeros. It can then be verified that rows $s = 2 \ldots (|\mathcal{Z}| - 1)$ of ${A^{[t]}}^+$ are of the form $(0, \ldots, -1, 1, \ldots 0)'$ with $s - 2$ zeroes on the

left (while the first row is composed of a single 1 in the first column, and the last row is all zeros).[18] Note that given the definition of $c$, $c'A^{[t]+}$picks out the $s^{th}$ row of $A^{[t]+}$ in (21), and we have that (20) and (21) are equivalent.

Remark 7.1 of HP observes that given the above, treatment effects can be identified if (in my notation) for some $s, s'$ and $t, t' \in \mathcal{T}$, $\mathbb{1}(\Sigma_{ti} = s) = \mathbb{1}(\Sigma_{t'i} = s')$ almost surely, since then we can identifiy $\mathbb{E}[Y_i(t') - Y_i(t)|\Sigma_{t'i} = s'] = \mathbb{E}[Y_i(t')|\Sigma_{t'i} = s'] - \mathbb{E}[Y_i(t)|\Sigma_{ti} = s]$. The idea of binary collections can be though of as a generalization of this type of result beyond the case of unordered monotonicity.

### G.4 Example 4: Lee and Salanié (2018)

Lee and Salanié (2018) (LS) consider a class of models in which unit $i$'s selection type depends upon a $J$-dimensional vector $V_i \in [0,1]^J$ and a vector valued function $Q : \mathcal{Z} \to \mathcal{Q}$ where $\mathcal{Q} \subseteq \mathbb{R}^J$. Selection is assumed to follow:

$$D_i^{[t]}(z) = \sum_{l \subseteq \{1 \dots J\}} c_l^t \cdot \prod_{j \in l} \mathbb{1}(V_{ji} \leq Q_j(z)) \tag{22}$$

for some set of coefficients $c_l^t$ defined over the subsets of $\{1 \dots J\}$, for each $t \in \mathcal{T}$, and where $V_{ji}$ is the $j^{th}$ component of $V_i$, and $Q_j$ the $j^{th}$ component of $Q$. This model nests the marginal treatment effects (MTE) framework when we have a binary treatment and $J = 1$, in which case we may let $Q(z) = \mathbb{E}[D_i|Z_i = z]$ be the propensity score function.

The second part of LS's Theorem 3.1 shows that under support/regularity conditions:

$$E[Y_i(t)|V_i = q] = \frac{\frac{\partial^J}{\partial_{q_1} \dots \partial_{q_J}} \mathbb{E}[Y_i D_i^{[t]}|Q(Z_i) = q]}{\frac{\partial^J}{\partial_{q_1} \dots \partial_{q_J}} \mathbb{E}[D_i^{[t]}|Q(Z_i) = q]} \tag{23}$$

Now let's see how this result can also be obtained through Theorem 1. For any vector $q \in \mathbb{R}^J$, let $S_i(q) := \{j \in \{1 \dots J\} : V_{ji} \leq q_j\}$ be the set of indices for which $V_{ji} \leq q_j$. Then $D_i^{[t]}(z) = \sum_{l \subseteq S_i(Q(z))} c_l^t$. Note that $D_i^{[t]}(z)$ only depends on $z$ through $Q(z)$. Thus, we could for each $q$ consider an arbitrary value $z \in \mathcal{Z}$ such that $Q(z) = q$, call it $Q^{-1}(q)$, and think of $D_i^{[t]}$ as a function $D_i^{[t]}(Q^{-1}(q))$ of $q$.

Let us consider a binary combination constructed to capture all units such that $V_i$ belongs to a rectangle $(q, q+h_1] \times (q, q+h_2] \cdots \times (q, q+h_J]$ in $\mathbb{R}^J$ for some "corner" location $q \in \mathbb{R}^J$ and widths $h_1 \dots h_J$. For any $s \subseteq \{1 \dots J\}$, let $h_s := \sum_{j \in s} h_j \mathbf{e}_j$, where $\mathbf{e}_j$ is the $j^{th}$ standard basis vector. This takes the form of a binary combination $(\alpha, t)$ having $K = 2^J$ and coefficients $\alpha_k = (-1)^{|s_k|}/\lambda$ for a certain scalar $\lambda$. The corresponding instrument values are $z_k = Q^{-1}(q + h_s)$ given an arbitrary ordering $s_1 \dots s_K$ on the $K$ distinct subsets of $\{1 \dots J\}$. Below we will verify that the corresponding linear combination of

[18] E.g. the modified forms with $|\mathcal{Z}| = 4$ are $A^{[t]} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}$, $A^{[t]+} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix}$.

the $D_i^{[t]}(Q^{-1}(q))$ is equal to $c(G_i)$ with probability one, where we let $c(G_i)$ be an indicator for $V_i \in (q, q+h_1] \cdots \times (q, q+h_J]$. Via Eq. (4) in the main text, we can thus identify $\mathbb{E}[Y_i(t)|V_i \in (q, q+h_1] \cdots \times (q, q+h_J]]$ as:

$$\mathbb{E}[Y_i(t)|c(G_i)=1] = \frac{\sum_{s\subseteq\{1\dots J\}}(-1)^{|s|}\cdot\mathbb{E}\left[Y_i\cdot D_i^{[t]}|Z_i = Q^{-1}(q+h_s)\right]}{\sum_{s\subseteq\{1\dots J\}}(-1)^{|s|}\cdot\mathbb{E}\left[D_i^{[t]}|Z_i = Q^{-1}(q+h_s)\right]} \tag{24}$$

The scalar $\lambda$ depends on the selection mechanism (22) and is $\lambda := \sum_{s\subseteq\{1\dots J\}}(-1)^{|s|}\sum_{l\subseteq s} c_l^t$.[19]

We now verify that with this notation $c(G_i) = \sum_{k=1}^{2^J}\alpha_k \cdot D_i^{[t]}(z_k)$. That is:

$$c(G_i) = \frac{1}{\lambda}\sum_{s\subseteq\{1\dots J\}}(-1)^{|s|}\cdot D_i^{[t]}\left(Q^{-1}(q+\sum_{j\in s}h_j\mathbf{e}_j)\right) = \frac{1}{\lambda}\sum_{s\subseteq\{1\dots J\}}(-1)^{|s|}\sum_{l\subseteq S_i(q+h_S)}c_l^t \tag{25}$$

Note that for any $S'\subseteq S$, $S_i(q+h'_S)\subseteq S_i(q+h_S)$. Thus $S_i(q+h_{\{1\dots J\}})$ is the "largest" $S_i(q+h_s)$ and $S_i(q)$ is the smallest. Define:

$$A_i = S_i(q+h_{\{1\dots J\}}) - S_i(q) = \{j\in\{1\dots J\} : q < V_{ji} \leq q+h_{\{1\dots J\}}\}$$

$A_i$ is simply the set of dimensions in which $V_{ji}$ falls within the rectangle starting at $q$ with widths $h_{\{1\dots J\}}$. Now comes the crucial step: we'll now show that (25) is zero for any individual $i$ for which $A_i$ does not contain all of $\{1\dots J\}$. Indeed, if there were any $j\notin A_i$, each set $S$ in the first summation of (25) that did not contain $j$ would be canceled out by the set $S\cup j$, because $(-1)^{|S\cup j|} = -(-1)^{|S|}$, while $S_i(q+h_S) = S_i(q+h_{S\cup j})$. Pairing all sets in this way, we see that evaluates to zero unless $A_i = \{1\dots J\}$. Now, $A_i = \{1\dots J\}$ implies that $S_i(q) = \emptyset$, and we can now write $c(G_i)$ as:

$$c(G_i) = \mathbb{1}(A_i = \{1\dots J\})\cdot\frac{1}{\lambda}\cdot\left(\sum_{s\subseteq\{1\dots J\}}(-1)^{|s|}\sum_{l\subseteq s}c_l^t\right) = \mathbb{1}(A_i=\{1\dots J\})$$

observing that we've defined $c$ to be equal to the quantity in parentheses, which depends on the selection model but not on $i$.

[19] We can simplify this expression of $\lambda$ as follows. Note that given 22) the coefficients $c_l^t$ must be such that $\sum_{l\subseteq s}c_l^t\in\{0,1\}$ for any $s\subseteq\{1\dots J\}$. Let $S_t$ be the collection of $s$ such that it is equal to one. This is the collection of subsets of the thresholds that when crossed correspond to taking treatment $t$. Then $\lambda = \sum_{s\in S_t}(-1)^{|s|}$. We can derive an alternative expression for $\lambda$ by making use of the identity that for any $\sum_{f\subseteq S}(-1)^{|f|}=0$ for any $S\neq\emptyset$. Then:

$$\lambda = \sum_{l\subseteq\{1\dots J\}}c_l^t\sum_{s\supseteq l}(-1)^{|s|} = \sum_{l\subseteq\{1\dots J\}}c_l^t\left(\cancel{\sum_{s\subseteq\{1\dots J\}}(-1)^{|s|}} - \cancel{\sum_{s\subseteq l}(-1)^{|s|}} + (-1)^{|l|}\right) = \sum_{l\subseteq\{1\dots J\}}(-1)^{|l|}\cdot c_l^t$$

LS's Theorem 3.1 considers the $J^{th}$ order derivative

$$\begin{aligned}
&\frac{\partial^J}{\partial_{q_1}\dots\partial_{q_J}}\mathbb{E}[Y_i D_i^{[t]}|Q(Z_i)=q] \\
&\qquad = \frac{\partial^J}{\partial_{q_2}\dots\partial_{q_J}}\lim_{h_1\downarrow 0}\frac{1}{h_1}\Big(\mathbb{E}[Y_i D_i^{[t]}|Q(Z_i)=q+h_1\mathbf{e}_1]-\mathbb{E}[Y_i D_i^{[t]}|Q(Z_i)=q]\Big) \\
&\qquad = \lim_{h_1\dots h_J\downarrow 0}\frac{1}{\prod_{j=1}^J h_j}\cdot\sum_{s\subseteq\{1\dots J\}}(-1)^{|s|}\cdot\mathbb{E}\left[Y_i D_i^{[t]}\middle| Q(Z_i)=q+\sum_{j\in s}h_j\mathbf{e}_j\right]
\end{aligned}$$

and takes the ratio:

$$\frac{\frac{\partial^J}{\partial_{q_1}\dots\partial_{q_J}}\mathbb{E}[Y_i D_i^{[t]}|Q(Z_i)=q]}{\frac{\partial^J}{\partial_{q_1}\dots\partial_{q_J}}\mathbb{E}[D_i^{[t]}|Q(Z_i)=q]}=\lim_{h_1\dots h_J\downarrow 0}\frac{\sum_{s\subseteq\{1\dots J\}}(-1)^{|s|}\cdot\mathbb{E}\left[Y_i D_i^{[t]}\middle| Q(Z_i)=q+\sum_{j\in s}h_j\mathbf{e}_j\right]}{\sum_{s\subseteq\{1\dots J\}}(-1)^{|s|}\cdot\mathbb{E}\left[D_i^{[t]}\middle| Q(Z_i)=q+\sum_{j\in s}h_j\mathbf{e}_j\right]}$$

LS's result (23) thus considers the limit of Eq. (24) as the width of the rectangle goes to zero in all dimensions.

### G.5 Example 5: Unordered (generalized) partial monotonicity

We can define a generalization of vector and partial monotonicity to settings with multi-valued treatments, also nesting unordered monotonicity:

**Assumption UPM.** *For any $t\in\mathcal{T}$, there exists a partial order $\succeq_t$ on $\mathcal{Z}$, such that if $z'\succeq_t z$, $D_i^{[t]}(z')\geq D_i^{[t]}(z)$ for all $i$.*

Note that even in the case of a binary treatment, UPM represents a generalization of partial monotonicity (PM), defined by Mogstad et al. (2019) for settings with multiple instruments. UPM allows for an arbitrary partial order on $\mathcal{Z}$, while PM considers a partial order that is based on holding all instruments but one at fixed values.

Assumption UPM implies that for any such $z,z'$: $D_i^{[t]}(z')-D_i^{[t]}(z)\in\{0,1\}$ and thus

$$\mathbb{E}[Y_i(t)|D_i^{[t]}(z')>D_i^{[t]}(z)]=\frac{\mathbb{E}[Y_i D_i^{[t]}|Z_i=z']-\mathbb{E}[Y_i D_i^{[t]}|Z_i=z]}{\mathbb{E}[D_i^{[t]}|Z_i=z']-\mathbb{E}[D_i^{[t]}|Z_i=z]}$$

UPM holds, for example, when instruments correspond to choice sets and agents choose rationally from them, as in Arora et al. (2021). In such a setting instrument values $z$ are subsets of the treatments $\mathcal{T}$ that are available to the agent, and $D_i^{[t]}(z)\geq D_i^{[t]}(z')$ whenever ($z/t\subseteq z'/t$ and $t\in z$ if $t\in z'$). In words, $D_i^{[t]}(z)$ is weakly increasing with respect to the inclusion of $t$ in $z$ (since $i$ can only choose $t$ if it is available), and weakly decreasing with respect to the inclusion of any $t'\neq t$ in $z$ (since $i$ may prefer $t'$ to $t$).

### G.6 Example 6: Pairwise notions of monotonicity

Sun and Wüthrich (2024) proposes a notion of IV-validity that is specific to two values $z,z'$ of the instrument (which may be a vector). This includes the standard LATE model

assumptions (independence, exclusion, and monotonicity). However, if independence and exclusion are maintained, the notion of pairwise valid instruments reduces to what we might call *pairwise-monotonicity*, i.e. that $D_i^{[t]}(z') \geq D_i^{[t]}(z')$ almost surely, or vice versa.

van't Hoff, Lewbel and Mellace (2023) consider a notion of "limited monotonicity" for settings with multiple binary instruments and a binary treatment, which in the notation above corresponds to a setting in which $z' = (1, \dots 1)$ and $z = (0, \dots 0)$. van't Hoff (2023) extends this notion to ordered treatments that need not be binary.

In the context of "judge designs" where the instrument is a scalar continuous measure of "leniency" with respect to a binary treatment, Sigstad (2023) and Sigstad (2024) introduce a notion of "extreme-pair" monotonicity $D_i(\bar{j}) \geq D_i(\underline{j})$ almost surely, where $\bar{j}$ is the strictest judge, and $\underline{j}$ the most lenient.

In the case of a binary treatment $D_i$, the above papers point out that under a limited version of "monotonicity" between a pair of values $z, z'$, a particular local average treatment effect can be identified from a simple Wald estimand:

$$\mathbb{E}[Y_i(1) - Y_i(0) | D_i(z') > D_i(z)] = \frac{\mathbb{E}[Y_i | Z_i = z'] - \mathbb{E}[Y_i | Z_i = z]}{\mathbb{E}[D_i | Z_i = z'] - \mathbb{E}[D_i | Z_i = z]}$$

This corresponds to a binary collection in which $\alpha_{z'} = 1$ and $\alpha_z = -1$ for $t = 1$ , while $\alpha_{z'} = -1$ and $\alpha_z = 1$ for $t = 0$.

## H Letting local causal parameters depend on $Z_i$

Let $z_k : \mathcal{Z} \to \mathcal{Z}$ be a function that maps an instrument value $Z_i$ to some possibly different value in $\mathcal{Z}$. Non-constant functions $z_k(\cdot)$ will allow us to nest parameters such as the average treatment effect on the treated, as well as some parameters from Goff (2024). In that paper $z_k(z)$ could for instance change one component of $z$, and the $\alpha_k$ and $z_k(\cdot)$ can be chosen so that $c(G_i, Z_i) := \sum_k \alpha_k \cdot D_i^{[t]}(z_k(Z_i))$ only takes values of zero or one, i.e. $\alpha_z = \{\alpha_k, z_k(z)\}_{k=1}^K$ yields a binary combination for any $z \in \mathcal{Z}$. Then, by the law of iterated expectations: $\mathbb{E}\left[Y_i(t) \,|c(G_i, Z_i) = 1\right] = \sum_{z \in \mathcal{Z}} P(Z_i = z) \cdot \mathbb{E}[Y_i(t) | c(G_i, Z_i) = 1]$ where each term in the summand is identified by (4) and the distribution of $Z_i$.

Let us maintain the assumption that the support of the instruments $\mathcal{Z}$ is discrete and finite. Consider any counterfactual mean of the form $\theta = \mathbb{E}[Y_i(t) | c(G_i, Z_i) = 1]$ where now $c : \mathcal{G} \times \mathcal{Z} \to \{0, 1\}$. By the law of iterated expectations over $Z_i$ and independence Eq. (2), we can write $\theta$ as:

$$\begin{aligned} \theta &= \sum_{z \in \mathcal{Z}} P(Z_i = z | c(G_i, z) = 1) \cdot \mathbb{E}[Y_i(t) | c(G_i, z) = 1, Z_i = z] \\ &= \sum_{z \in \mathcal{Z}} P(Z_i = z) \cdot \mathbb{E}[Y_i(t) | c_z(G_i) = 1] \end{aligned} \tag{26}$$

where we let $c_z(g)$ denote $c(g, z)$. Eq (26) shows that $\theta$ can be written as a convex combination of $|\mathcal{Z}|$ counterfactual means of the form $\mu_c^t$ considered by Theorems 1 and 2,

with complier groups $c_z(\cdot)$ that depend on $z$. It is clear then by Theorem 1, a sufficient condition for $\theta$ to be outcome-nonrestrictive identified is that $c_z$ lies in the rowspace of matrix $A^{[t]}$ for each $z \in \mathcal{Z}$.

Theorem 2 similarly extends to the more general class of functions $c(G_i, Z_i)$, provided that the family $\mathscr{P}_Z$ of distributions over the instruments allows for degenerate distributions at each value of $Z_i$. Then $c_z$ must lie in the rowspace of $A^{[t]}$ for all $z \in \mathcal{Z}$. If it were not, then for some $z \in \mathcal{Z}$, $c_z \notin rs(A^{[t]})$ and hence $\mu^t_{c_z}$ is not outcome-nonrestrictive identified, by Theorem 2. For a degenerate distribution $P_{\mathcal{Z}}$ that sets $P(Z_i = z) = 1$, $\theta = \mu^t_{c_z}$ and hence $\theta$ is not outcome-nonrestrictive identified if $c_z \notin rs(A^{[t]})$. With this extension Theorem 2 of this paper nests Theorem 2 of Goff (2024) as a special case, and expands its reach even in the case that vector monotonicity is maintained, if the outcome variable is continuous.

# I Partial identification when $c \notin rowspace(A^{[t]})$

## I.1 Relationship to Bai, Huang, Moon, Shaikh and Vytlacil (2024)

Bai et al. (2024) (BHMSV) study the identifying power for ATEs and unconditional counterfactual means of a restriction on selection that they call *generalized monotonicity* (GM). In my notation, GM says that for a given $\mathcal{P}_{latent}$ and each $t \in \mathcal{T}$, there exists an instrument value $z^* = z^*(t, \mathcal{P}_{latent})$ such that

$$P(D_i(z^*) \neq t \text{ and } D_i(z) = t \text{ for some } z \in \mathcal{Z}) = 0 \tag{27}$$

according to $\mathcal{P}_{latent}$. That is, no individual takes treatment $t$ when $z \neq z^*$ unless they also do when $z = z^*$. BHMSV show that GM or any strengthening of it (that does not restrict outcomes) does not reduce the size of identified sets for unconditional parameters of the form $\mathbb{E}[Y_i(t)]$, when the outcome variable has finite support $\mathcal{Y}$ and the instruments are also finite.

While GM nests many notions of monotonicity from the literature that have been used for positive point identification results, it generalizes them in a different way than the criterion $c \in rs(A^{[t]})$ of the present paper does. While $c \in rs(A^{[t]})$ ensures point identification of $\mathbb{E}[Y_i(t)|c_{G_i} = 1]$, GM represents a double-edged sword when the parameter of interest is an unconditional mean or ATE with $c = (1, 1, \dots 1)'$. Using Theorem 2 of this paper, we can see that GM is in fact sufficient to establish either that i) $\mathbb{E}[Y_i(t)]$ is point identified in an outcome-nonrestrictive way; or ii) that it is *not* point identified in an outcome-nonrestrictive way. Which of these cases i) or ii) holds can be determined by the observable distribution $\mathcal{P}_{obs}$, and does not depend on $\mathcal{G}$ beyond it satisfying GM.

Let $\tilde{\mathcal{G}}(\mathcal{P}_{latent})$ be the support of $G_i$ under $\mathcal{P}_{latent}$, and note that (27) is equivalent to:

$$\text{For all } g \in \tilde{\mathcal{G}}(\mathcal{P}_{latent}): \quad A^{[t]}_{z^*,g} = 0 \implies A^{[t]}_{z,g} = 0 \text{ for all } z \in \mathcal{Z} \tag{28}$$

Consider a given distribution of observables $\mathcal{P}_{obs}$. Either $P(T_i = t|Z_i = z^*) = 1$ or $P(T_i = t|Z_i = z^*) < 1$ according to $\mathcal{P}_{obs}$. If the first case holds, then $\mathbb{E}[Y_i(t)] = \mathbb{E}[Y_i|Z_i = z^*]$ and $\mathbb{E}[Y_i(t)]$ is thus point-identified without requiring any restrictions on selection. Thus, assuming GM or any strengthening of it cannot reduce the identified set for $\mathbb{E}[Y_i(t)]$ further, unless it results in model rejection.

If on the other hand $P(T_i = t|Z_i = z^*) < 1$ and GM holds, then there must exist a $g \in \tilde{\mathcal{G}}(\mathcal{P}_{latent})$ such that $A^{[t]}_{z^*,g} = 0$. Therefore by (28), for this $g$ it must be that $A^{[t]}_{z,g} = 0$ for all $z \in \mathcal{Z}$, i.e. there are never-takers with respect to treatment $t$. This in turn implies that $(1, 1, \ldots 1)' \notin rs(A^{[t]})$, precisely the case in which we know that $\mathbb{E}[Y_i(t)]$ is *not* outcome-nonrestrictive point identified, by Theorem 2.

By showing that the bounds on $\mathbb{E}[Y_i(t)]$ in partially identified settings are not improved by imposing restrictions stronger than GM, BHMSV underscore the importance of: i) focusing on other parameters of interest beyond the ATE (i.e. $c \neq (1, 1, \ldots 1)'$) when one is willing to impose restrictions on selection; and ii) finding restrictions on selection that are outside of the scope of GM. Indeed, many of the selection models reported in Appendix K below do not satisfy GM, yet are sufficient for point identification of more localized treatment effect parameters than the ATE (and in some cases the ATE as well).

The remainder of this section provides more detail to build intuition about the connection between BHMSV's result and the proof of Theorem 2 in this paper. For a given $\mathcal{P}_{obs}$, let us write the identified set for $\mathbb{E}[Y_i(t)]$ under model $M$ as

$$\Theta(\mathcal{P}_{obs}, M) = \{\theta(\mathcal{P}) : \phi(\mathcal{P}) = \mathcal{P}_{obs} \text{ and } \mathcal{P} \in M\}$$

Given BHMSV's assumption that $\mathcal{Y}$ is finite, let us for each $y \in \mathcal{Y}$ define $x^y$ to be a $|\mathcal{G}|$-component vector with components $x^y_g = P(Y_i(t) = y|G_i = g)$ and $\beta$ to be a $|\mathcal{Z}|$-component vector with components $\beta^y_z = P(Y_i = y, T_i = t|Z_i = z)$. The restriction $\phi(\mathcal{P}) = \mathcal{P}_{obs}$ corresponds to the set of solutions to finite system of linear equations $A^{[t]}x^y = \beta^y$, for each $y \in \mathcal{Y}$. Given $|\mathcal{Y}| < \infty$, we can collect these into a single finite linear system $\mathcal{A}^{[t]}\tilde{x} = \tilde{\beta}$, where $\mathcal{A}^{[t]}$ is a block diagonal matrix of $A^{[t]}$ copied $|\mathcal{Y}|$ times, $\tilde{x}$ is a $|\mathcal{Y}| \times |\mathcal{G}|$ component vector, and $\tilde{\beta}$ is a $|\mathcal{Y}| \times |\mathcal{Z}|$ component vector. The set $\mathcal{X} := \{\tilde{x}(\mathcal{P}) : \phi(\mathcal{P}) = \mathcal{P}_{obs}\}$ is thus a vector space, where we let $\tilde{x}(\mathcal{P})$ represent $\tilde{x}$ as a function of the distribution of model fundamentals $\mathcal{P}$.

Whether GM or any strengthening of it reduces the identified set for $\mathbb{E}[Y_i(t)]$ thus depends upon whether the action of $\theta(\cdot)$ on the $\mathcal{P} \in M$ such that $\tilde{x}(\mathcal{P}) \in \mathcal{X}$ reduces $\Theta(\mathcal{P}_{obs}, M)$ relative to a case with no restrictions on selection. Eq. (12) from the proof of Theorem 2 suggests that $\Theta(\mathcal{P}_{obs}, M)$ satisfies

$$\Theta(\mathcal{P}_{obs}, M) \subseteq \left\{ \mathbb{1}'(A^{[t]})^+\beta + \sum_{g'} [\mathbb{1}'(I - (A^{[t]})^+A^{[t]})]_{g'} \cdot w_{g'} : w \in \mathbb{R}^{|\mathcal{G}|} \right\} \tag{29}$$

where $\mathbb{1} := (1, 1, \ldots 1)'$ and $\beta$ a $|\mathcal{Z}|$-component vector with components $\beta_z = \mathbb{E}[Y_i \cdot \mathbb{1}(T_i =$

$t)|Z_i = z]$. GM implies that the set of the RGS is not a singleton if $P(T_i = t|Z_i = z^*) < 1$. The subset relation appearing in (29) reflects that, as in Theorem 2, some $\tilde{x}$ for which $\mathcal{A}^{[t]}\tilde{x} = \tilde{\beta}$ may not be attainable from $\mathcal{P}$ that are valid distributions and reflect any further assumptions of the model $M$, for example that $Y_i$ has bounded support.

BHMSV show that if $M$ does not restrict outcomes, $\Theta(\mathcal{P}_{obs}, M)$ is in fact equal to the identified set under no selection restrictions, which is (given the finite support $\mathcal{Y}$):

$$\{\beta_{z^*} - \min\{\mathcal{Y}\} \cdot P(T_i \neq t|Z_i = z^*), \beta_{z^*} + \max\{\mathcal{Y}\} \cdot P(T_i \neq t|Z_i = z^*)\}$$

An interesting question for further study is in what manner the result of BHMSV extends to the more general class target parameters indexed by vectors $c$ that may differ from $(1, 1, \ldots, 1)'$. A reasonable conjecture would be that if, given $\mathcal{P}_{obs}$, a class of restrictions on selection cannot change the fact that $c \notin rs(A^{[t]})$ , there is limited scope for such restrictions to reduce the size of the identified set for $\mu_c^t$.

### I.2 Partial identification in general

Accordingly, consider an arbitrary $c \in \{0, 1\}^{|\mathcal{G}}$ where we may have that $c \notin rs(A^{[t]})$. By similar logic as above, we can deduce that the identified set $\Theta(\mathcal{P}_{obs}, M)$ for $\mu_c^t$ satisfies:

$$\Theta(\mathcal{P}_{obs}, M) \subseteq \frac{1}{P(c(G_i) = 1)} \cdot \left\{ c'(A^{[t]})^+\beta + \sum_{g'} [c'(I - (A^{[t]})^+A^{[t]})]_{g'} \cdot w_{g'} : w \in \mathbb{R}^{|\mathcal{G}|} \right\}$$

The RHS may again be an outer set for $\Theta(\mathcal{P}_{obs}, M)$, for example when $Y_i$ has bounded support. An added complication now, as compared to unconditional means, is that the probability $P(c(G_i) = 1)$ is no longer known to be equal to one, and our only identifying information for it is that $\sum_{g \in \mathcal{G}} A_{gz}^{[t]} = d_z$ for all $z \in \mathcal{Z}$, where $d_g := P(T_i = t|Z_i = z)$.

## J Supplemental material for the application to interaction effects

### J.1 Motivating the restriction imposed by Proposition 4

We can rationalize the restriction $\mathcal{G} \subseteq \mathcal{G}^{sep}$ made in Proposition 4 by supposing that individuals choose *separately* whether to receive treatment $A$ or $B$, rather than as a single joint decision. Let $S(z)$ denote the set of treatments among $\{A, B\}$ offered to an individual when their instrument realization is $z \in \{\text{neither}, \text{A}, \text{B both}\}$. That is, $S(\text{neither}) = \emptyset$, $S(\text{A}) = \{A\}$, $S(\text{B}) = \{B\}$, $S(\text{both}) = \{A, B\}$.

**Definition.** *We say that the population exhibits* ***separable choices*** *if their counterfactual selection satisfies for each $z \in \{$neither, A, B both$\}$:*

$$T_i(z) = \{t \in S(z) : U_i(t) \geq 0\}$$

*where treatment $C$ is here understood as the set of treatments $\{A, B\}$, and treatment $0$ is understood as the null set $\emptyset$.*

Separable (counterfactual) choices says that individuals choose treatment $A$ if and only if $U_i(A) \geq 0$ and $B$ if and only if $U_i(B) \geq 0$, subject to the options offered to them. This implies that $T_i(\text{both}) = C \implies T_i(A) = A$ and $T_i(B) = B$, and similarly that $T_i(A) = A$ and $T_i(B) = B \implies T_i(\text{both}) = C$. This eliminates exactly the remaining five groups displayed in gray in Table 2.

### J.2 Identification with covariates

Suppose that instead of (2) we have

$$\{Z_i \perp\!\!\!\perp (\tilde{Y}_i, G_i)\}|X_i \tag{30}$$

where $X_i$ are observed covariates that are unaffected by treatment. This holds, for example, if the instruments are independent of these covariates jointly with the latent heterogeneity $(\tilde{Y}_i, G_i)$ across individual: $Z_i \perp\!\!\!\perp (\tilde{Y}_i, G_i, X_i)$.

Consider a binary combination $(t, \alpha)$ such that $\sum_k \alpha_k D_i^{[t]}(z_k) = c^{[t,\alpha]}(G_i)$ for all $i$ where $c^{[t,\alpha]}(G_i) \in \{0, 1\}$ for all $g \in \mathcal{G}$. I do not consider the case in which $\sum_k \alpha_k D_i^{[t]}(z_k) = c^{[t,\alpha]}(G_i, X_i)$ for some function $c^{[t,\alpha]}$ that depends both on $G_i$ and $X_i$, though such an extension would be possible. By the steps that establish Eq. (4) in the uncondional case, (4) generalizes to

$$\mathbb{E}\left[Y_i(t) \left| c^{[t,\alpha]}(G_i) = 1, X_i = x\right.\right] = \frac{\sum_{k=1}^K \alpha_k \cdot \mathbb{E}\left[Y_i \cdot D_i^{[t]} | Z_i = z_k, X_i = x\right]}{\sum_{k=1}^K \alpha_k \cdot \mathbb{E}\left[D_i^{[t]} | Z_i = z_k, X_i = x\right]} \tag{31}$$

for any value $x$. Notice that although $P(c^{[t,\alpha]}(G_i) = 1|X_i = x) = \mathbb{E}[c^{[t,\alpha]}(G_i)|X_i = x]$ might vary with $x$, it is identified by the denominator of the above for each: $P(c^{[t,\alpha]}(G_i) = 1|X_i = x) = \sum_{k=1}^K \alpha_k \cdot \mathbb{E}\left[D_i^{[t]} | Z_i = z_k, X_i = x\right]$. Consequently, the overall counterfactual

mean that does not condition on $x$ is identified as

$$\begin{aligned}
\mathbb{E}[Y_i(t)|c^{[t,\alpha]}(G_i)=1] &= \int dF_{X|c^{[t,\alpha]}(G)=1}(x)\cdot \mathbb{E}[Y_i(t)|c^{[t,\alpha]}(G_i)=1, X_i=x] \\
&= \int dF_{X|c^{[t,\alpha]}(G)=1}(x)\cdot \frac{\sum_{k=1}^K \alpha_k \cdot \mathbb{E}\left[Y_i\cdot D_i^{[t]}|Z_i=z_k, X_i=x\right]}{P(c^{[t,\alpha]}(G_i)=1|X_i=x)} \\
&= \int dF_X(x)\cdot \frac{\sum_{k=1}^K \alpha_k \cdot \mathbb{E}\left[Y_i\cdot D_i^{[t]}|Z_i=z_k, X_i=x\right]}{P(c^{[t,\alpha]}(G_i)=1)} \\
&= \frac{\sum_{k=1}^K \alpha_k \cdot \mathbb{E}\left[\mathbb{E}\left[Y_i\cdot D_i^{[t]}|Z_i=z_k, X_i\right]\right]}{P(c^{[t,\alpha]}(G_i)=1)} \\
&= \frac{\sum_{k=1}^K \alpha_k \cdot \mathbb{E}\left[\mathbb{E}\left[Y_i\cdot D_i^{[t]}|Z_i=z_k, X_i\right]\right]}{\sum_{k=1}^K \alpha_k \cdot \mathbb{E}\left[\mathbb{E}[D_i^{[t]}|Z_i=z_k, X_i]\right]}
\end{aligned}$$

applying Bayes' rule, echoing an argument for the LATE model by Frölich (2007). See also Appendix A of Goff (2024). Given a binary collection, we can use these results to identify treatment effects that either do or do not condition on $X_i$.

Note that the conditional independence assumption 30 further allows us to identify the distribution of covariates $X_i$ among "compliers" for whom $c^{[t,\alpha]}(G_i)=1$ given a binary combination $(t,\alpha)$. Suppose that $X_i$ has $M$ components so that $X_i\in\mathbb{R}^M$. Then for any Borel set $\mathcal{B}$ of $\mathbb{R}^M$ we have that, by (30):

$$\begin{aligned}
\sum_k \alpha_k \cdot \mathbb{E}[\mathbb{1}(X_i\in\mathcal{B})\cdot P(T_i=t|Z_i=z_k, X_i)] &= \sum_k \alpha_k \cdot \mathbb{E}[\mathbb{1}(X_i\in\mathcal{B})\cdot \mathbb{E}[D_i^{[t]}(z_k)|X_i, Z_i=z_k]] \\
&= \sum_k \alpha_k \cdot \mathbb{E}[\mathbb{1}(X_i\in\mathcal{B})\cdot \mathbb{E}[D_i^{[t]}(z_k)|X_i]] \\
&= \mathbb{E}\left[\mathbb{1}(X_i\in\mathcal{B})\cdot \mathbb{E}\left[\sum_k \alpha_k\cdot D_i^{[t]}(z_k)\middle| X_i\right]\right] \\
&= \mathbb{E}[\mathbb{1}(X_i\in\mathcal{B})\cdot \mathbb{E}[c^{[t,\alpha]}(G_i)|X_i]] \\
&= \mathbb{E}[\mathbb{E}[\mathbb{1}(X_i\in\mathcal{B})\cdot c^{[t,\alpha]}(G_i)|X_i]] \\
&= \mathbb{E}[\mathbb{1}(X_i\in\mathcal{B})\cdot c^{[t,\alpha]}(G_i)] \\
&= P(c^{[t,\alpha]}(G_i)=1)\cdot P(X_i\in\mathcal{B}|c^{[t,\alpha]}(G_i)=1)
\end{aligned}$$

Meanwhile

$$\begin{aligned}
\sum_k \alpha_k \cdot \mathbb{E}[P(T_i=t|Z_i=z_k, X_i)] &= \sum_k \alpha_k \cdot \mathbb{E}[\mathbb{1}(X_i\in\mathcal{B})\cdot \mathbb{E}[D_i^{[t]}(z_k)|X_i, Z_i=z_k]] \\
&= \sum_k \alpha_k \cdot \mathbb{E}[\mathbb{E}[D_i^{[t]}(z_k)|X_i]] = \mathbb{E}\left[\mathbb{E}\left[\sum_k \alpha_k\cdot D_i^{[t]}(z_k)\middle| X_i\right]\right] \\
&= \mathbb{E}[\mathbb{E}[c^{[t,\alpha]}(G_i)|X_i]] = \mathbb{E}[c^{[t,\alpha]}(G_i)] = P(c^{[t,\alpha]}(G_i)=1)
\end{aligned}$$

And thus

$$P(X_i \in \mathcal{B}|c^{[t,\alpha]}(G_i) = 1) = \frac{\sum_k \alpha_k \cdot \mathbb{E}[\mathbb{1}(X_i \in \mathcal{B}) \cdot P(T_i = t|Z_i = z_k, X_i)]}{\sum_k \alpha_k \cdot \mathbb{E}[P(T_i = t|Z_i = z_k, X_i)]} \tag{32}$$

This implies, for example, that we can identify the mean of $X_i$ among the $c^{[t,\alpha]}(G_i) = 1$ sub-population as

$$\mathbb{E}[X_i \in \mathcal{B}|c^{[t,\alpha]}(G_i) = 1] = \frac{\sum_k \alpha_k \cdot \mathbb{E}[X_i \cdot P(T_i = t|Z_i = z_k, X_i)]}{\sum_k \alpha_k \cdot \mathbb{E}[P(T_i = t|Z_i = z_k, X_i)]}$$

which generalizes the seminal result of Abadie (2003) for the case of the binary treatment, binary instrument LATE model.

If we have a binary collection $\{(t, \alpha^{[t]})\}_{t\in\psi}$, then Eq. (32) yields overidentification restrictions since it implies that

$$\frac{\sum_k \alpha_z^{[t]} \cdot \mathbb{E}[\mathbb{1}(X_i \in \mathcal{B}) \cdot P(T_i = t|Z_i = z_k, X_i)]}{\sum_k \alpha_k^{[t]} \cdot \mathbb{E}[P(T_i = t|Z_i = z_k, X_i)]} = \frac{\sum_k \alpha_z^{[t']} \cdot \mathbb{E}[\mathbb{1}(X_i \in \mathcal{B}) \cdot P(T_i = t'|Z_i = z_k, X_i)]}{\sum_k \alpha_k^{[t']} \cdot \mathbb{E}[P(T_i = t'|Z_i = z_k, X_i)]}$$

for any $t, t' \in \psi$. Note that this restriction is trivially satisfied for the binary collection that isolates compliers in the binary instrument, binary treatment LATE model.

### J.3 Details on empirical estimates including strata covariates

Consider a binary combination $(t, \alpha)$ for a given treatment $t$, with associated function $c$. As shown in Appendix J.2, when Equation 2 holds conditional on covariates $X_i$ we have:

$$\begin{aligned}\mathbb{E}[Y_i(t)|c(G_i) = 1] &= \frac{\sum_{k=1}^K \alpha_k \cdot \mathbb{E}\left[\mathbb{E}\left[Y_i \cdot D_i^{[t]}|Z_i = z_k, X_i\right]\right]}{P(c(G_i) = 1)} \\ &= \frac{\mathbb{E}\left[\sum_{k=1}^K \alpha_k \cdot \mathbb{E}\left[Y_i \cdot D_i^{[t]}|Z_i = z_k, X_i\right]\right]}{P(c(G_i) = 1)}\end{aligned} \tag{33}$$

where

$$P(c(G_i) = 1) = \sum_{k=1}^K \alpha_k \cdot \mathbb{E}\left[\mathbb{E}[D_i^{[t]}|Z_i = z_k, X_i]\right] = \mathbb{E}\left[\sum_{k=1}^K \alpha_k \cdot \mathbb{E}[D_i^{[t]}|Z_i = z_k, X_i]\right] \tag{34}$$

In the empirical application of Angelucci and Bennett (2024), randomization is performed within nine strata, which represents a discrete $X_i$ taking on nine values. To simplify estimation, I assume that the expectations $\mathbb{E}[Y_i \cdot D_i^{[t]}|Z_i, X_i]$ and $\mathbb{E}[D_i^{[t]}|Z_i, X_i]$ additively separable in $Z_i$ and $X_i$:

$$\mathbb{E}[Y_i \cdot D_i^{[t]}|Z_i, X_i] = \beta_{\text{both}}^{[t]} \cdot \mathbb{1}(Z_i = \text{both}) + \beta_A^{[t]} \cdot \mathbb{1}(Z_i = \text{just A}) + \beta_B^{[t]} \cdot \mathbb{1}(Z_i = \text{just B}) + \sum_{s=1}^9 \lambda_s^{[t]} \cdot \mathbb{1}(X_i = s) \tag{35}$$

and

$$\mathbb{E}[D_i^{[t]}|Z_i, X_i] = \gamma_{\text{both}}^{[t]} \cdot \mathbb{1}(Z_i = \text{both}) + \gamma_A^{[t]} \cdot \mathbb{1}(Z_i = \text{just A}) + \gamma_B^{[t]} \cdot \mathbb{1}(Z_i = \text{just B}) + \sum_{s=1}^{9} \rho_s^{[t]} \cdot \mathbb{1}(X_i = s) \tag{36}$$

i.e. linear regression equations with a full set of strata fixed effects (with none omitted) and instead omitting a dummy variable for $\mathbb{1}(Z_i = \text{neither})$.

The four estimates of $p := P(G_i = \text{complier})$ based on the choice model $\mathcal{G}^{sep}$ given in (10) then become, using Eqs (34) and (36):

$$p = \left\{ \gamma_{\text{both}}^{[C]} + \sum_{s=1}^{9} \rho_s^{[C]} \cdot P(X_i = s) \right\} = \left\{ \gamma_A^{[A]} - \gamma_{\text{both}}^{[A]} \right\} = \left\{ \gamma_A^{[A]} - \gamma_{\text{both}}^{[A]} \right\} = \left\{ \gamma_{\text{both}}^{[0]} - \gamma_A^{[0]} - \gamma_B^{[0]} \right\} \tag{37}$$

Treatment effect estimates are then based on the following expressions using (35)-(36):

$$\begin{aligned}
&\mathbb{E}[Y_i(C)|i \text{ is complier}] = \frac{\beta_{\text{both}}^{[C]} + \sum_{s=1}^{9} \lambda_s^{[C]} \cdot P(X_i = s)}{\gamma_{\text{both}}^{[C]} + \sum_{s=1}^{9} \rho_s^{[C]} \cdot P(X_i = s)}, \quad \mathbb{E}[Y_i(A)|i \text{ is complier}] = \frac{\beta_A^{[A]} - \beta_{\text{both}}^{[A]}}{\gamma_A^{[A]} - \gamma_{\text{both}}^{[A]}} \\
&\mathbb{E}[Y_i(B)|i \text{ is complier}] = \frac{\beta_B^{[B]} - \beta_{\text{both}}^{[B]}}{\gamma_B^{[B]} - \gamma_{\text{both}}^{[B]}}, \qquad \mathbb{E}[Y_i(0)|i \text{ is complier}] = \frac{\beta_{\text{both}}^{[0]} - \beta_A^{[0]} - \beta_B^{[0]}}{\gamma_{\text{both}}^{[0]} - \gamma_A^{[0]} - \gamma_B^{[0]}}
\end{aligned} \tag{38}$$

and the local average interaction effect among compliers $LAIE$ is estimated accordingly. Some involved algebra shows that the expressions in (38) recover the results for complier average treatment effects in Theorem 1 of Blackwell (2017), given one-sided noncompliance.

### J.4 GMM estimation

Note that given the overidentification of $p := P(G_i = \text{complier})$, any of the local counterfactual means (38) could be estimated by swapping out an alternative estimate of $p$ in the denominator. In principle, we can increase efficiency by estimating treatment effects as well as $LAIE$ while imposing Eq. (37), in a generalized method of moments (GMM) estimation approach. Column (4) of Table 3 implements this. Given the logic of Corollary 2, GMM estimation of $LAIE$ combines the ITT regression (39) with the first-stage regressions (36), and imposing (37) as additional moments. For the treatment effect estimates $\mathbb{E}[Y_i(t) - Y_i(0)|i \text{ is complier}]$ for $t \in \{0, A, B\}$, GMM estimation combines regressions (35) for treatments $t$ and 0 with the first-stage regressions and (37). All GMM estimates use the two-step GMM estimator, starting from an initial identity weight-matrix, and requesting a cluster robust final weight-matrix and standard errors.

### J.5 Deriving the expression $\theta^{ITT}/p$ for local average interaction effect

Consider first the case with no covariates. We have using Eqs. (4) and (9):

$$
\begin{aligned}
LAIE &= \mathbb{E}[Y_i(C)|c(G_i)=1] - \mathbb{E}[Y_i(A)|c(G_i)=1] - \mathbb{E}[Y_i(B)|c(G_i)=1] + \mathbb{E}[Y_i(0)|c(G_i)=1] \\
&= \frac{\mathbb{E}[Y_i \cdot D_i^{[C]}|Z_i=\text{both}]}{\mathbb{E}[D_i^{[C]}|Z_i=\text{both}]} - \frac{\mathbb{E}[Y_i \cdot D_i^{[A]}|Z_i=\text{just A}] - \mathbb{E}[Y_i \cdot D_i^{[A]}|Z_i=\text{both}]}{\mathbb{E}[D_i^{[A]}|Z_i=\text{just A}] - \mathbb{E}[D_i^{[A]}|Z_i=\text{both}]} \\
&\quad - \frac{\mathbb{E}[Y_i \cdot D_i^{[B]}|Z_i=\text{just A}] - \mathbb{E}[Y_i \cdot D_i^{[B]}|Z_i=\text{both}]}{\mathbb{E}[D_i^{[B]}|Z_i=\text{just B}] - \mathbb{E}[D_i^{[B]}|Z_i=\text{both}]} \\
&+ \frac{\mathbb{E}[Y_i \cdot D_i^{[0]}|Z_i=\text{both}] - \mathbb{E}[Y_i \cdot D_i^{[0]}|Z_i=\text{just A}] - \mathbb{E}[Y_i \cdot D_i^{[0]}|Z_i=\text{just B}] + \mathbb{E}[Y_i \cdot D_i^{[0]}|Z_i=\text{neither}]}{\mathbb{E}[D_i^{[0]}|Z_i=\text{both}] - \mathbb{E}[D_i^{[0]}|Z_i=\text{just A}] - \mathbb{E}[D_i^{[0]}|Z_i=\text{just B}] - \mathbb{E}[D_i^{[0]}|Z_i=\text{neither}]} \\
&= \frac{1}{p} \cdot \Big\{ \mathbb{E}[Y_i \cdot D_i^{[C]}|Z_i=\text{both}] - \mathbb{E}[Y_i \cdot D_i^{[A]}|Z_i=\text{just A}] + \mathbb{E}[Y_i \cdot D_i^{[A]}|Z_i=\text{both}] \\
&\qquad - \mathbb{E}[Y_i \cdot D_i^{[B]}|Z_i=\text{just B}] + \mathbb{E}[Y_i \cdot D_i^{[B]}|Z_i=\text{both}] + \mathbb{E}[Y_i \cdot D_i^{[0]}|Z_i=\text{both}] \\
&\qquad - \mathbb{E}[Y_i \cdot D_i^{[0]}|Z_i=\text{just A}] - \mathbb{E}[Y_i \cdot D_i^{[0]}|Z_i=\text{just B}] + \mathbb{E}[Y_i \cdot D_i^{[0]}|Z_i=\text{neither}] \Big\} \\
&= \frac{1}{p} \cdot \Big\{ \mathbb{E}[Y_i \cdot (D_i^{[0]} + D_i^{[A]} + D_i^{[B]} + D_i^{[C]})|Z_i=\text{both}] - \mathbb{E}[Y_i \cdot (D_i^{[0]} + D_i^{[A]})|Z_i=\text{just A}] \\
&\qquad - \mathbb{E}[Y_i \cdot (D_i^{[0]} + D_i^{[B]})|Z_i=\text{just B}] + \mathbb{E}[Y_i \cdot D_i^{[0]}|Z_i=\text{neither}] \Big\} \\
&= \frac{1}{p} \cdot \{ \mathbb{E}[Y_i|Z_i=\text{both}] - \mathbb{E}[Y_i|Z_i=\text{just A}] - \mathbb{E}[Y_i|Z_i=\text{just B}] + \mathbb{E}[Y_i|Z_i=\text{neither}] \} = \frac{\theta^{ITT}}{p}
\end{aligned}
$$

where $\theta^{ITT} := \gamma_3 - \gamma_1 - \gamma_2$ from the ITT regression Eq. (7). In the above I have used Eq. (10) in the second step, then combined terms, and finally using that $(D_i^{[0]} + D_i^{[A]} + D_i^{[B]} + D_i^{[C]}) = 1$, that $(D_i^{[0]} + D_i^{[A]})$ conditional on $Z_i = \text{just A}$ (given one-sided noncompliance), that $(D_i^{[0]} + D_i^{[B]})$ conditional on $Z_i = \text{just B}$, and that $D_i^{[0]}$ conditional on $Z_i = \text{neither}$.

With covariates $X_i$, the standard intent-to-treat regression generalizes (7) by adding a linear function in the covariates that includes a constant:

$$Y_i = \gamma_1 \cdot \mathbb{1}(Z_i = A) + \gamma_2 \cdot \mathbb{1}(Z_i = B) + \gamma_3 \cdot \mathbb{1}(Z_i = C) + \pi' X_i + \nu_i \tag{39}$$

In this case, $\theta^{ITT} := \gamma_3 - \gamma_1 - \gamma_2$ is equal to

$$\mathbb{E}[Y_i|Z_i=\text{both}, X_i] - \mathbb{E}[Y_i|Z_i=\text{just A}, X_i] - \mathbb{E}[Y_i|Z_i=\text{just B}, X_i] + \mathbb{E}[Y_i|Z_i=\text{neither}, X_i]$$

with probability one (i.e. for all $X_i$). The same steps as above show that, using Eqs. (33)

and (34):

$$
\begin{aligned}
LAIE = \frac{1}{p} \cdot \mathbb{E}\Big\{ & \mathbb{E}[Y_i \cdot D_i^{[C]} | Z_i = \text{both}, X_i] - \mathbb{E}[Y_i \cdot D_i^{[A]} | Z_i = \text{just A}, X_i] \\
& + \mathbb{E}[Y_i \cdot D_i^{[A]} | Z_i = \text{both}, X_i] - \mathbb{E}[Y_i \cdot D_i^{[B]} | Z_i = \text{just A}, X_i] - \mathbb{E}[Y_i \cdot D_i^{[B]} | Z_i = \text{both}, X_i] \\
& - \mathbb{E}[Y_i \cdot D_i^{[0]} | Z_i = \text{both}, X_i] + \mathbb{E}[Y_i \cdot D_i^{[0]} | Z_i = \text{just A}, X_i] + \mathbb{E}[Y_i \cdot D_i^{[0]} | Z_i = \text{just B}, X_i] \\
& - \mathbb{E}[Y_i \cdot D_i^{[0]} | Z_i = \text{neither}, X_i] \Big\} = \frac{1}{p} \cdot \mathbb{E}[\gamma_3 - \gamma_1 - \gamma_2] = \frac{\theta^{ITT}}{p}
\end{aligned}
$$

provided that Eq. (39) is correctly specified for the conditional mean $\mathbb{E}[Y_i | Z_i, X_i]$.

### J.6 Setting up the linear program to test for offending types

This section considers the identification of bounds on the proportion of the population that belongs to a certain set of response types $\mathcal{G}^*$, within a larger selection model $\mathcal{G}$. This method is implemented in Section 5 to discuss whether first stage selection information is consistent with the choice model $\mathcal{G} \subseteq \mathcal{G}^{sep}$, under the maintained assumption that $\mathcal{G} \subseteq \mathcal{G}^{WARP}$. Thus for the remainder of this section we assume that $\mathcal{G} = \mathcal{G}^{WARP}$ defined in Section 5. This section also ignores the randomization strata $X_i$, which is valid for testing "first-stage" restrictions if the response-type distribution is common across strata.

For any set of response types $\mathcal{G}^* \subseteq \mathcal{G}^{WARP}$, we can partially identify $P(G_i \in \mathcal{G}^*)$ as $P(G_i \in \mathcal{G}^*) \in [LB^*, UB^*]$ where

$$LB^* = \min_{x \in \mathbb{R}^9} w'x \quad \text{subject to } \mathcal{A}x = \beta \text{ and } x \geq 0 \tag{40}$$

$$UB^* = \max_{x \in \mathbb{R}^9} w'x \quad \text{subject to } \mathcal{A}x = \beta \text{ and } x \geq 0 \tag{41}$$

with $w$ a $9 \times 1$ vector (where $|\mathcal{G}^{WARP}| = 9$) with components $w_g = \mathbb{1}(g \in \mathcal{G}^*)$, and the constraint $x \geq 0$ is read as all components of the vector $x$ must be weakly positive. If $LB^*$ were found to be strictly positive with $\mathcal{G}^*$ chosen to be $\mathcal{G}^{WARP} - \mathcal{G}^{sep}$, this would constitute evidence that the restriction $\mathcal{G} \subseteq \mathcal{G}^{sep}$ is not satisfied, assuming that $\mathcal{G} \subseteq \mathcal{G}^{WARP}$.

The $16 \times 9$ matrix $\mathcal{A}$ can be obtained from the matrices $A^{[t]}$, and the 16-component

vector $\beta$ estimated from the data:

$$\mathcal{A} = \begin{bmatrix} A^{[0]} \\ A^{[A]} \\ A^{[B]} \\ A^{[both]} \end{bmatrix} = \left[\begin{array}{ccccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \end{array}\right] \qquad \beta = \left(\begin{array}{c} P(T_i = 0 | Z_i = \text{neither}) \\ P(T_i = 0 | Z_i = \text{just A}) \\ P(T_i = 0 | Z_i = \text{just B}) \\ P(T_i = 0 | Z_i = \text{both}) \\ \hline P(T_i = A | Z_i = \text{neither}) \\ P(T_i = A | Z_i = \text{just A}) \\ P(T_i = A | Z_i = \text{just B}) \\ P(T_i = A | Z_i = \text{both}) \\ \hline P(T_i = B | Z_i = \text{neither}) \\ P(T_i = B | Z_i = \text{just A}) \\ P(T_i = B | Z_i = \text{just B}) \\ P(T_i = B | Z_i = \text{both}) \\ \hline P(T_i = C | Z_i = \text{neither}) \\ P(T_i = C | Z_i = \text{just A}) \\ P(T_i = C | Z_i = \text{just B}) \\ P(T_i = C | Z_i = \text{both}) \end{array}\right)$$

Point estimates of the bounds $LB^*$ and $UB^*$ are readily obtained by solving the linear programs (40) and (41) with the sample estimator $\hat{\beta}$.

Given sampling error in $\hat{\beta}$ however, we would like to construct a valid confidence interval for the partially identified parameter $P(G_i \in \mathcal{G}^*) = w'x$ given its representation as a solution to the linear program $Ax = \beta, x \geq 0$. This problem is considered by Fang et al. (2023), and I use the `fsst` command in the `lpinfer` package in `R` to generate a confidence interval for the parameters $P(G_i \in \mathcal{G}^*)$ considered in the main text. The required inputs for `fsst` are the matrix $\mathcal{A}$, the vector $\beta$ (specified as a function of the data, as the FSST procedure makes use of estimates of $\beta$ in bootstrap samples).

### J.6.1 Results for Angelucci and Bennett (2024)

In addition to the results for $\mathcal{G}^* = \mathcal{G}^{WARP} - \mathcal{G}^{sep}$ reported in the main text, I here present some further estimates. A 90% confidence interval using the method of Fang and Santos (2018) (FSST) does rule out zero but is otherwise similar at $[0, 0.8281]$ (as opposed to $[0, 0.8297]$ for the 95% interval). However, the p-value for the null hypothesis that $P(G_i \in \mathcal{G}^{WARP} - \mathcal{G}^{sep}) = 0$ puts it just on the margin of being included in the 90% confidence interval. The `lpinfer` package in `R` also allows for statistical inference on solutions to problems (40) and (41) using methods introduced by Romano and Shaikh (2008) and Cho and Russell (2024). The method of Cho and Russell (2024) yields $[0.02, 0.86]$ as a 95% confidence interval. The method of Romano and Shaikh (2008) yields $[0, 0.84]$ as a 95% confidence interval. Confidence intervals for the share of the favor-B type (which the point estimates suggest may be the largest offending type) only are similar to

the confidence intervals for all offending types in $\mathcal{G}^{WARP} - \mathcal{G}^{sep}$. Overall, the results offer little evidence against the assumption that $\mathcal{G}^{sep}$ represents the true choice model, and any violations that cannot be ruled out appear to be minor. This supports the conclusion that complementarities between the two treatments are identified among compliers without restricting outcomes, in line with Proposition 4.

In the above calculations, I do not condition on the nine strata used by Angelucci and Bennett (2024) for randomization. This could be implemented by expanding $\mathcal{A}$ and $\beta$ to have $16 \times 9$ rows each, rather than 16. However the above results are valid if the response-type distribution is common across strata, and under this assumption allow for a much more efficient use of the available sample.

### J.7 Financial incentives and support for academic achievement

Angrist, Lang and Oreopoulos (2009) (ALO) report results from an the Project STAR intervention that cross-randomized academic support and financial incentives on academic achievement among first-year students at a large Canadian university. In this setting, I let treatment A represent the Student Support Program (SSP): a program which gave students access to peer advisers and supplemental instruction. I let treatment B represent the Student Fellowship Program (SFP), which made students eligible for merit scholarships based on good performance during the first year courses.

The STAR intervention randomized 250 students into an arm that was offered access to the SSP only ($Z_i =$ just A), another 250 students to be offered access to the SFP only ($Z_i =$ just B), and a third group of 150 students that was offered access to both programs ($Z_i =$ both). A control group of 1,006 students were offered neither ($Z_i =$ neither).

I use the replication data from ALO, which tracks program takeup as well as student performance among those students included in Project STAR. Treatment uptake for treatment A (SSP) is observable, and I define it as having attended a facilitated study groups or having met with an advisor. For treatment B, I follow ALO in defining treatment takeup as having responded to their invitation to sign up for the assigned treatment. ALO define compliance with respect to SSP (treatment A) similarly as having given their consent by simply signing up for their assigned treatment. With this definition however, no individual offered both treatments could opt for one treatment alone.[20] Since further information is available on whether individuals actually take part in SSP activities, I make use of this additional information.

I test the overidentification restriction of $\mathcal{G} \subseteq \mathcal{G}^{sep}$ in this setting as described in Eq. (38), however note that in the present setting there are no randomization strata that need to be controlled for. The four point estimates for $p = P(i \text{ is complier})$ are 41%, 21%, 51%, and 34%, respectively. A test for equality of the four estimates returns a chi-squared

[20] Given WARP, this would then limit the choice model to the groups n.t., complier, only both, A+, and B+ from Table 2. This group yields a rather uninteresting selection model when intersected with $\mathcal{G}^{sep}$ (leaving just never takers and compliers) in order to afford outcome non-restrictive identification of complementarity between the treatments. However, defining compliance as ALO do also rejects the overidentification restriction (37), with a chi-squared statistic (with 2 degrees of freedom) of 28.66.

statistic (with 2 degrees of freedom) of 26.16, a p-value of 0.0000.[21] Thus in contrast to the application of Angelucci and Bennett (2024), we find in the ALO context that we can clearly reject the choice model $\mathcal{G} \subseteq \mathcal{G}^{sep}$ that is required to identify complementarity between the two treatment effects in an outcome-agnostic manner.

# K Catalog of outcome-nonrestrictive identification results

The following results are for various small values of $|\mathcal{Z}|$ $|\mathcal{T}|$. Results for $|\mathcal{Z}| = |\mathcal{T}| = 3$ are available upon request from the author (these add roughly 40 pages of output).

For a given $\mathcal{T}$ and $\mathcal{Z}$, binary collections are organized by selection models, given unique identifiers of the format `SM.`$|\mathcal{T}|$`.`$|\mathcal{Z}|$`.s`, where `s` is an index of the various selection models in that setting. Within each selection model, binary collections are enumerated by ascending numbers $i), ii)$ etc. Each binary collection is presented via the coefficient vectors $\alpha_{t'}$ and $\alpha_t$ (following the notation of Sec. 4.2 but keeping $t$ and $t'$ explicit).

Binary collections that share a common maximal selection model $\mathcal{G}(\alpha)$ organized under that selection model, and are not re-listed for $\mathcal{G} \subseteq \mathcal{G}(\alpha)$. Further, some binary collections for $\mathcal{G}$ might be listed under a $\mathcal{G}(\alpha)$ that nests $\mathcal{G}$ only after suitable re-labeling of the treatments and instruments. It is for this reason that the set of binary collections listed under a given selection model may not be closed under addition, even when adding the $c$ for two such collections results in another vector composed of all zeroes and ones.

For example, consider the $A$ matrices for SM.2.3.8 and SM2.3.1 below:

$$\texttt{SM.2.3.8}: \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \texttt{SM.2.3.1.swapped}: \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

where by `SM.2.3.1.swapped` I indicate that I have swapped the first and third rows of the A matrix listed in the catalog that follows for SM.2.3.1. This swapping corresponds to a re-labelling of the instrument values.

SM.2.3.8 consists of two selection types, and the catalog shows that it admits of binary collection i) with $\alpha_1 = (0.5, 0.5, 0)'$ and $\alpha_0 = (0, 0, 1)'$ yielding $c = (1, 0)'$ as well as binary collection ii) with $\alpha_1 = (0, 0, 1)'$ and $\alpha_0 = (0.5, 0.5, 0)'$ yielding $c = (1, 0)'$. This implies that SM.2.3.8 also admits of a binary collection yielding $c = (1, 1)'$, with $\alpha_0 = \alpha_1 = (0.5, 0.5, 1)'$.

The reason that this third binary collection is not listed under SM.2.3.8 is that SM.2.3.8 is not maximal for it: unlike collections i) and ii) which just include one of the two types in SM.2.3.8, identification of the average treatment effect for both of the types in SM.2.3.8 holds in the less restrictive selection model SM.2.3.1.swapped, which contains the selection types of SM.2.3.8 in its first and third columns. The sole binary collection listed under SM.2.3.1 corresponds to $c = (1, 1)'$ in SM.2.3.8.

[21] Inferential methods for the linear program described in J.6 at the 95% level suggest that at least about 15% of the population belongs to groups in $\mathcal{G}^{WARP} - \mathcal{G}^{sel}$, provided that $WARP$ holds.

## K.1 2 treatments, 2 instrument values

**SM.2.2.1**

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

i) $(t', t) = (1, 0)$; $\alpha_{t'} = (-1, 1)'$; $\alpha_t = (1, -1)'$; $c = (0, 1, 0)'$

**SM.2.2.2**

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

i) $(t', t) = (1, 0)$; $\alpha_{t'} = (0, 1)'$; $\alpha_t = (1, 0)'$; $c = (0, 1)'$

ii) $(t', t) = (1, 0)$; $\alpha_{t'} = (1, 0)'$; $\alpha_t = (0, 1)'$; $c = (1, 0)'$

iii) $(t', t) = (1, 0)$; $\alpha_{t'} = (1, 1)'$; $\alpha_t = (1, 1)'$; $c = (1, 1)'$

## K.2 3 treatments, 2 instrument values

**SM.3.2.1**

$$A = \begin{bmatrix} 0 & 0 & 1 & 2 \\ 0 & 1 & 2 & 2 \end{bmatrix}$$

i) $(t', t) = (1, 0)$; $\alpha_{t'} = (0, 1)'$; $\alpha_t = (1, -1)'$; $c = (0, 1, 0, 0)'$

**SM.3.2.2**

$$A = \begin{bmatrix} 0 & 0 & 1 & 2 \\ 0 & 1 & 1 & 2 \end{bmatrix}$$

i) $(t', t) = (1, 0)$; $\alpha_{t'} = (-1, 1)'$; $\alpha_t = (1, -1)'$; $c = (0, 1, 0, 0)'$

**SM.3.2.3**

$$A = \begin{bmatrix} 1 & 2 & 0 & 1 & 2 \\ 0 & 0 & 1 & 2 & 2 \end{bmatrix}$$

i) $(t', t) = (1, 0)$; $\alpha_{t'} = (0, 1)'$; $\alpha_t = (1, 0)'$; $c = (0, 0, 1, 0, 0)'$

**SM.3.2.4**

$$A = \begin{bmatrix} 2 & 0 & 1 & 2 \\ 0 & 1 & 1 & 2 \end{bmatrix}$$

i) $(t', t) = (1, 0)$; $\alpha_{t'} = (-1, 1)'$; $\alpha_t = (1, 0)'$; $c = (0, 1, 0, 0)'$

**SM.3.2.5**

$$A = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 2 \end{bmatrix}$$

i) $(t', t) = (1, 0)$; $\alpha_{t'} = (1, 1)'$; $\alpha_t = (1, 1)'$; $c = (1, 1, 0)'$

## K.3 2 treatments, 3 instrument values

**SM.2.3.1**

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

i) $(t', t) = (1, 0)$; $\alpha_{t'} = (1, 1, 0)'$; $\alpha_t = (-1, 1, 2)'$; $c = (1, 0, 1)'$

**SM.2.3.2**

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

i) $(t', t) = (1, 0)$; $\alpha_{t'} = (1, 1, -2)'$; $\alpha_t = (-1, -1, 2)'$; $c = (0, 1, 1, 0)'$

**SM.2.3.3**

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

i) $(t', t) = (1, 0)$; $\alpha_{t'} = (0.5, 0.5, 0.5)'$; $\alpha_t = (1, 1, 1)'$; $c = (1, 1, 1)'$

ii) $(t', t) = (1, 0)$; $\alpha_{t'} = (1, 0, 0)'$; $\alpha_t = (0, 1, 1)'$; $c = (1, 1, 0)'$

iii) $(t', t) = (1, 0)$; $\alpha_{t'} = (0, 1, 0)'$; $\alpha_t = (1, 0, 1)'$; $c = (1, 0, 1)'$

iv) $(t', t) = (1, 0)$; $\alpha_{t'} = (0.5, 0.5, -0.5)'$; $\alpha_t = (0, 0, 1)'$; $c = (1, 0, 0)'$

v) $(t', t) = (1, 0)$; $\alpha_{t'} = (0, 0, 1)'$; $\alpha_t = (1, 1, 0)'$; $c = (0, 1, 1)'$

vi) $(t', t) = (1, 0)$; $\alpha_{t'} = (0.5, -0.5, 0.5)'$; $\alpha_t = (0, 1, 0)'$; $c = (0, 1, 0)'$

vii) $(t', t) = (1, 0)$; $\alpha_{t'} = (-0.5, 0.5, 0.5)'$; $\alpha_t = (1, 0, 0)'$; $c = (0, 0, 1)'$

**SM.2.3.4**

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

i) $(t', t) = (1, 0)$; $\alpha_{t'} = (2, 1, -1)'$; $\alpha_t = (0, 1, 1)'$; $c = (1, 1, 0)'$

**SM.2.3.5**

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

i) $(t', t) = (1, 0)$; $\alpha_{t'} = (0, 1, 1)'$; $\alpha_t = (0, 1, 1)'$; $c = (1, 1, 1, 1)'$

ii) $(t', t) = (1, 0)$; $\alpha_{t'} = (1, 0, 0)'$; $\alpha_t = (-1, 1, 1)'$; $c = (0, 1, 0, 1)'$

iii) $(t', t) = (1, 0)$; $\alpha_{t'} = (0, 1, 0)'$; $\alpha_t = (0, 0, 1)'$; $c = (1, 1, 0, 0)'$

iv) $(t', t) = (1, 0)$; $\alpha_{t'} = (0, 0, 1)'$; $\alpha_t = (0, 1, 0)'$; $c = (0, 0, 1, 1)'$

v) $(t', t) = (1, 0)$; $\alpha_{t'} = (-1, 1, 1)'$; $\alpha_t = (1, 0, 0)'$; $c = (1, 0, 1, 0)'$

**SM.2.3.6**

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

i) $(t', t) = (1, 0)$; $\alpha_{t'} = (2, -1, -1)'$; $\alpha_t = (-2, 1, 1)'$; $c = (0, 1, 1, 0)'$

**SM.2.3.7**

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

i) $(t', t) = (1, 0)$; $\alpha_{t'} = (1, -1, 0)'$; $\alpha_t = (0, 0, 1)'$; $c = (1, 0, 0)'$

ii) $(t', t) = (1, 0)$; $\alpha_{t'} = (0, -1, 1)'$; $\alpha_t = (1, 0, 0)'$; $c = (0, 1, 0)'$

**SM.2.3.8**

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

i) $(t', t) = (1, 0)$; $\alpha_{t'} = (0.5, 0.5, 0)'$; $\alpha_t = (0, 0, 1)'$; $c = (1, 0)'$

ii) $(t', t) = (1, 0)$; $\alpha_{t'} = (0, 0, 1)'$; $\alpha_t = (0.5, 0.5, 0)'$; $c = (0, 1)'$

**SM.2.3.9**

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

i) $(t', t) = (1, 0)$; $\alpha_{t'} = (1, 1, 0)'$; $\alpha_t = (0, 0, 1)'$; $c = (1, 1, 0)'$

ii) $(t', t) = (1, 0)$; $\alpha_{t'} = (1, 1, 1)'$; $\alpha_t = (0.5, 0.5, 0.5)'$; $c = (1, 1, 1)'$

iii) $(t', t) = (1, 0)$; $\alpha_{t'} = (1, 0, 0)'$; $\alpha_t = (-0.5, 0.5, 0.5)'$; $c = (1, 0, 0)'$

iv) $(t', t) = (1, 0)$; $\alpha_{t'} = (0, 1, 0)'$; $\alpha_t = (0.5, -0.5, 0.5)'$; $c = (0, 1, 0)'$

v) $(t', t) = (1, 0)$; $\alpha_{t'} = (1, 0, 1)'$; $\alpha_t = (0, 1, 0)'$; $c = (1, 0, 1)'$

vi) $(t', t) = (1, 0)$; $\alpha_{t'} = (0, 1, 1)'$; $\alpha_t = (1, 0, 0)'$; $c = (0, 1, 1)'$

vii) $(t', t) = (1, 0)$; $\alpha_{t'} = (0, 0, 1)'$; $\alpha_t = (0.5, 0.5, -0.5)'$; $c = (0, 0, 1)'$

**SM.2.3.10**

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

i) $(t', t) = (1, 0)$; $\alpha_{t'} = (1, 0, -1)'$; $\alpha_t = (-1, 0, 1)'$; $c = (0, 1, 0, 1, 0, 0)'$

**SM.2.3.11**

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

i) $(t', t) = (1, 0)$; $\alpha_{t'} = (0, 1, 0)'$; $\alpha_t = (-1, 0, 1)'$; $c = (0, 1, 0)'$

ii) $(t', t) = (1, 0)$; $\alpha_{t'} = (0, 0, 1)'$; $\alpha_t = (-1, 1, 0)'$; $c = (0, 0, 1)'$

## K.4 3 treatments, 3 instrument values

Omitted for brevity (251 selection models). See `https://arxiv.org/abs/2406.02835`.