# Online appendix to *Sharp tests of instrument validity for general outcome distributions*: Sharpness in the three instruments case, proofs for section 4 and examples of shape restrictions.

## A.1 Proof for subsection 3.3

***Proof of theorem 3.3.***

**1⟹2:**

The proof follows immediately from theorem 3.2.

**2⟹1:**

Now I show that to falsify the model, it is not necessary to consider distributions over response types. It suffices to construct distributions over instrument-response types that fit the observed conditional treatment probabilities, and then, ensure that these are consistent with the marginals at different combinations of $Z$ and $X$. In other words, there is no need to enumerate response types. I state this result for the general, $K$ instrument values case. This is stated and proven in appendix A.1.1 proposition A.1.1.

Now, I show that condition 2 from theorem 3.3 implies condition 2 from proposition A.1.1.

For any $S \subset \mathcal{Z}$, let $\mathcal{I}_{S,x_l}$ denote the set of instrument-response types that realize $x_l$ if and only if $z \in S$ (see the formal definition in the proof of proposition A.1.1). Let $\boldsymbol{p}$ be a vector of instrument-response type probabilities that satisfies condition 2 from theorem 3.3. Define $p(\mathcal{I}_{S,x_l})$ as:

$$p(\mathcal{I}_{S,x_l}) = \sum_{t^Z : t^Z(z)=x_l \ \ \forall z \in S} p(t^z)$$

By construction, the terms $p(\mathcal{I}_{S,x_l})$ satisfy condition 73.

It remains to construct functions that are consistent with condition 72. To this end, I first state the following lemma:

First I state the following lemma.

**Lemma A.1.1.** *Given nonnegative functions $f_0$, $f_1$ and $f_2$, define the functions:*

$$
\begin{aligned}
f_a &= f_0 - \min\{f_0, f_1 + f_2\} \\
f_b &= f_1 - \min\{f_1, f_0 + f_2\} \\
f_c &= f_2 - \min\{f_2, f_0 + f_1\} \\
f_g &= \big(\min\{f_0, f_1 + f_2\} - \max\{f_1, f_2\}\big)\mathbf{1}\big\{\min\{f_0, f_1 + f_2\} - \max\{f_1, f_2\}\big\} \\
f_d &= \big(\min\{f_0, f_1\} - f_2\big)\mathbf{1}\big\{\min\{f_0, f_1\} - f_2\big\} \\
f_e &= \big(\min\{f_0, f_2\} - f_1\big)\mathbf{1}\big\{\min\{f_0, f_2\} - f_1\big\} \\
f_i &= \big(\min\{f_1, f_2\} - f_0\big)\mathbf{1}\big\{\min\{f_1, f_2\} - f_0\big\} \\
f_j &= \min\{f_0, f_1, f_2\} - f_g - f_h - f_i
\end{aligned}
$$

*Then:*

1. *All functions are positive.*
2. $f_j + f_i + f_g + f_h + f_d = \min\{f_0, f_1\}$.
3. $\min\{f_0, f_1 + f_2\} = f_g + f_e + \min\{f_0, f_1, f_2\}$.

The proof of the lemma is given at the end of this appendix.

Lemma A.1.1 decomposes the nonnegative functions $f_0$, $f_1$ and $f_2$ into different nonnegative functions. Consequently, it decomposes its integral into the integral of different nonnegative functions. Each of these functions is associated with a unique region in the area below $f_0$, $f_1$ and $f_2$. The goal of the proof is to distribute the mass assigned to different groups $I_{S,x_l}$ across these areas. Then, potential outcomes for these types can be defined as their share of the total area in their corresponding region.

In what follows, let $f_{Y,X|Z=z_k}$ represent function $f_k$ in lemma A.1.1. And for $letter \in \{a, b, c, d, e, f, g, h, i, j\}$ let $A_{letter,x_l}$ represent $\int_{\mathcal{Y}} f_{letter}(y)dy$.

Abusing notation, for any $z_k, z_{k'}, z_{k''} \in \mathcal{Z}$ I will denote the terms $p(\mathcal{I}_{\{z_k\},x_l})$, $p(\mathcal{I}_{\{z_k,z_k,z_{k'}\},x_l})$ and $p(\mathcal{I}_{\{z_k,z_{k'},z_{k''}\},x_l})$ as $p(\mathcal{I}_{k,x_l})$, $p(\mathcal{I}_{kk',x_l})$ and $p(\mathcal{I}_{kk'k'',x_l})$ respectively.

The inequalities in 8, 9, 10 and 11 can then be recast as:

$$p(I_{0,x_l}) + p(I_{01,x_l}) + p(I_{02,x_l}) + p(I_{012,x_l}) = \Psi(C_0) \tag{45}$$
$$p(I_{1,x_l}) + p(I_{01,x_l}) + p(I_{12,x_l}) + p(I_{012,x_l}) = \Psi(C_1) \tag{46}$$
$$p(I_{2,x_l}) + p(I_{02,x_l}) + p(I_{12,x_l}) + p(I_{012,x_l}) = \Psi(C_2) \tag{47}$$
$$p(I_{01,x_l}) + p(I_{012,x_l}) \leq \Psi(C_{01}) \tag{48}$$
$$p(I_{02,x_l}) + p(I_{012,x_l}) \leq \Psi(C_{02}) \tag{49}$$
$$p(I_{12,x_l}) + p(I_{012,x_l}) \leq \Psi(C_{12}) \tag{50}$$
$$p(I_{012,x_l}) \leq \Psi(C_{012}) \tag{51}$$
$$p(I_{01,x_l}) + p(I_{02,x_l}) + p(I_{012,x_l}) \leq \Psi(C_{01,02}) \tag{52}$$
$$p(I_{01,x_l}) + p(I_{12,x_l}) + p(I_{012,x_l}) \leq \Psi(C_{01,12}) \tag{53}$$
$$p(I_{02,x_l}) + p(I_{12,x_l}) + p(I_{012,x_l}) \leq \Psi(C_{02,12}) \tag{54}$$

Note that there is a set of such equations for every $x_l \in \mathcal{X}$.

In what follows, for fixed $x_l$ and any nonempty $S \subset \mathcal{Z}$, I show how to construct quantities $\breve{q}(\mathcal{I}_{S,x_l})$ from the quantities $p(\mathcal{S}_{S,x_l})$. Then, I will define a distribution of the outcome for every group $\breve{q}(\mathcal{I}_{S,x_l})$ that coincides with the observed distributions $f_{Y,X|Z=z_k}(y, x_l)$ for every $z_k \in \mathcal{Z}$.

Each quantity $\breve{q}(\mathcal{I}_{S,x_l})$ will be associated with one unique region in the following figure. The heigh of these regions at any given $y \in \mathcal{Y}$ will be the contribution of $\breve{q}(\mathcal{I}_{S,x_l})$ to the subdensities $f_{Y,X|Z=z_k}(y, x_l)$.

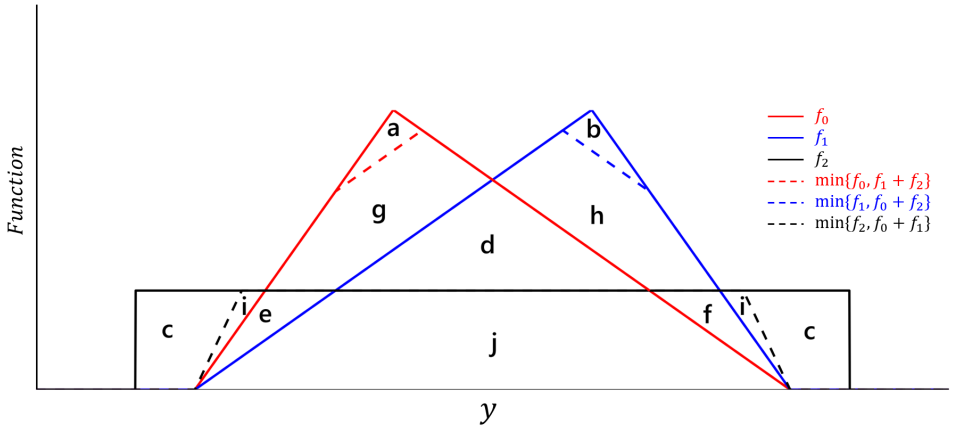
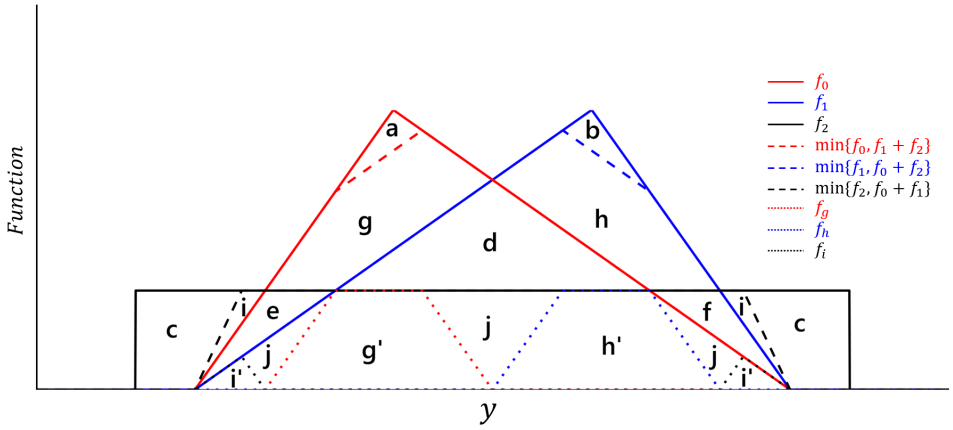



Figure 8




***Notes:***

*The solid lines in both figures represent the subdensities $f_{Y,X|Z=z_k}(y, x_l)$ for $k = 0, 1, 2$ and the partitions of the area below them that allows to construct a distribution of potential outcomes that is consistent with the observed joint distribution of $Y, X$ and $Z$, the instrument validity assumptions, and shape restrictions. The label $f_k$ denotes the subdensity $f_{Y,X|Z=z_k}(y, x_l)$. The dashed lines represent functions that are obtained from transforming the primitive functions $f_k$. The long dashes represent functions that take the form $\min\{f_k, \sum_{k' \neq k} f_{k'}\}$ . In the second figure, the functions $f_g$, $f_h$ and $f_i$ are represented by the lines with short dashes and are defined as $f_g = \min\{f_0, f_1 + f_2\} - \min\{f_0, f_1, f_2\}$, $f_h = \min\{f_1, f_0 + f_2\} - \min\{f_0, f_1, f_2\}$ and $f_i = \min\{f_2, f_0 + f_1\} - \min\{f_0, f_1, f_2\}$. The areas labeled from a to i correspond to the partition of the total area below at least one of the functions $f_k$. To construct a valid distribution of potential outcomes, I assign sufficient-taker types to each of the areas depicted. Then, I assign outcomes to such groups that are proportional to the heights of the area to which they where assigned. All areas are delimited by one or more of the depicted functions. The areas g and $g'$ have exact same area and height at any value of y. Equivalently, the areas h and $h'$; and the areas i and $i'$ have the same area and height. This allows assigning the same area twice, when necessary.*

**Lemma A.1.2.** *a Properties: Definitions Additions.*

**Step 1: (assign mass from $I_{0,x_l}$ to areas $a$, $b$ and $c$)**

Focus on the instrument value $z_0$. Note that by definition $\Psi(C_0) \geq \Psi(C_{01,02})$ Subtracting equation 45 and inequality 52:

$$p(I_{0,x_l}) \geq \Psi(C_0) - \Psi(C_{01,02}) \tag{55}$$

Set:

$$\breve{a}_{0,x_l} := \Psi(C_0) - \Psi(C_{01,02}), \tag{56}$$

In this step, I assigned a share of the units in $p(I_{0,x_l})$ to area $a$. $\breve{a}_{0,x_l}$ denotes that units of type $I_{0,x_l}$ are being assigned to area $a$.

Note that $p(I_{0,x_l}) - \breve{a}_{0,x_l} \geq 0$ and that subtracting equations 56 and 45 yields:

$$\left[p(I_{0,x_l}) - \breve{a}_{0,x_l}\right] + p(I_{01,x_l}) + p(I_{02,x_l}) + p(I_{012,x_l}) = \Psi(C_{01,02}). \tag{57}$$

Analogous constructions hold for $z_1$ and $z_2$ yielding definitions:

$$\breve{b}_{1,x_l} := \Psi(C_1) - \Psi(C_{01,12}), \tag{58}$$

$$\breve{c}_{2,x_l} := \Psi(C_2) - \Psi(C_{02,12}). \tag{59}$$

And:

$$\left[p(I_{1,x_l}) - \breve{a}_{1,x_l}\right) + p(I_{01,x_l}) + p(I_{12,x_l}) + p(I_{012,x_l}) = \Psi(C_{01,12}), \tag{60}$$

$$\left[p(I_{2,x_l}) - \breve{a}_{2,x_l}\right) + p(I_{02,x_l}) + p(I_{12,x_l}) + p(I_{012,x_l}) = \Psi(C_{02,12}). \tag{61}$$

**Step 2: (assign mass from $I_{02,x_l}$ and $I_{0,x_l}$ to areas $g$ and $e$)**
Focusing on $z_0$, note that subtracting equation 57 and inequality 48 yields:

$$\left[p(I_{0,x_l}) - \breve{a}_{0,x_l}\right] + p(I_{02,x_l}) \geq \Psi(C_{01,02}) - \Psi(C_{01}).$$

Moreover, lemma A.1.2 implies that $\Psi(C_{01,02}) - \Psi(C_{01}) = A_{g,x_l} + A_{e,x_l}$.

Set:

$$\breve{g}_{02,x_l} := \begin{cases} A_{g,x_l} & if \quad p(I_{02,x_l}) \geq A_{g,x_l} \\ p(I_{02,x_l}) & if \quad p(I_{02,x_l}) < A_{g,x_l} \end{cases}$$

$$\breve{e}_{02,x_l} := \begin{cases} A_{e,x_l} & if \quad p(I_{02,x_l}) - A_{g,x_l} \geq A_{e,x_l} \\ p(I_{02,x_l}) - A_{g,x_l} & if \quad p(I_{02,x_l}) - A_{g,x_l} < A_{e,x_l} \end{cases}$$

$$\breve{g}_{0,x_l} := A_{g,x_l} - \breve{g}_{02,x_l}$$

$$\breve{e}_{0,x_l} := A_{e,x_l} - \breve{e}_{02,x_l}$$

The total mass assigned in this step amounts to the nonnegative quantity $\Psi(C_{01,02}) - \Psi(C_{01})$. The quantity $\breve{g}_{0,x_l}$ represents mass from $p(I_{0,x_l})$ that was assigned to area $g$. Equivalently, $\breve{g}_{02,x_l}$ represents mass from $p(I_{02,x_l})$ that was assigned to area $g$. Analogous interpretations hold for $\breve{e}_{0,x_l}$ and $\breve{e}_{02,x_l}$.

Subtracting $\breve{g}_{0,x_l} + \breve{g}_{02,x_l} + \breve{e}_{0,x_l} + \breve{e}_{02,x_l}$ from equation 57 yields:

$$\begin{aligned} &\left[p(I_{0,x_l}) - \breve{a}_{0,x_l} - \breve{g}_{0,x_l} - \breve{e}_{0,x_l}\right] + p(I_{01,x_l}) \\ &+ \left[p(I_{02,x_l}) - \breve{g}_{02,x_l} - \breve{e}_{02,x_l}\right] + p(I_{012,x_l}) = \Psi(C_{01}) \end{aligned} \tag{62}$$

**Step 3: (assign mass from $I_{12,x_l}$ and $I_{1,x_l}$ to areas $h$ and $f$)**

Step 3 mimics step 2, but focuses on areas $h$ and $d$ and types $I_{12,x_l}$ and $I_{1,x_l}$.

Note that subtracting equation 60 and inequality 48 yields:

$$\left[p(I_{1,x_l}) - \breve{a}_{1,x_l}\right] + p(I_{12,x_l}) \geq \Psi(C_{01,12}) - \Psi(C_{01}). \tag{63}$$

Moreover, lemma A.1.2 implies that $\Psi(C_{01,12}) - \Psi(C_{01}) = A_{h,x_l} + A_{f,x_l}$.

Set:

$$\breve{h}_{12,x_l} := \begin{cases} A_{h,x_l} & if \quad p(I_{12,x_l}) \geq A_{h,x_l} \\ p(I_{12,x_l}) & if \quad p(I_{12,x_l}) < A_{h,x_l} \end{cases} \tag{64}$$

$$\breve{f}_{12,x_l} := \begin{cases} A_{f,x_l} & if \quad p(I_{12,x_l}) - A_{h,x_l} \geq A_{f,x_l} \\ p(I_{12,x_l}) - A_{h,x_l} & if \quad p(I_{12,x_l}) - A_{h,x_l} < A_{f,x_l} \end{cases} \tag{65}$$

$$\breve{h}_{1,x_l} := A_{h,x_l} - \breve{h}_{12,x_l} \tag{66}$$

$$\breve{f}_{0,x_l} := A_{f,x_l} - \breve{f}_{12,x_l} \tag{67}$$

The total mass assigned in this step amounts to the nonnegative quantity $\Psi(C_{01,12}) - \Psi(C_{01})$. The quantity $\breve{h}_{1,x_l}$ represents mass from $p(I_{1,x_l})$ that was assigned to area $h$. Equivalently, $\breve{h}_{12,x_l}$ represents mass from $p(I_{12,x_l})$ that was assigned to area $h$. Analogous interpretations hold for $\breve{f}_{1,x_l}$ and $\breve{f}_{12,x_l}$.

Subtracting $\breve{h}_{1,x_l} + \breve{h}_{12,x_l} + \breve{f}_{1,x_l} + \breve{f}_{12,x_l}$ from equation 57 yields:

$$\begin{aligned} &\left[p(I_{1,x_l}) - \breve{c}_{1,x_l} - \breve{h}_{1,x_l} - \breve{f}_{1,x_l}\right] + p(I_{01,x_l}) \\ &+ \left[p(I_{12,x_l}) - \breve{h}_{12,x_l} - \breve{f}_{12,x_l}\right] + p(I_{012,x_l}) = \Psi(C_{01}) \end{aligned} \tag{68}$$

**Step 4: (assign mass from $I_{01,x_l}$ to area $d$)**
Define:

$$\breve{d}_{01,x_l} = \begin{cases} A_{d,x_l} & if \quad p(I_{01,x_l}) \geq A_{d,x_l} \\ p(I_{01,x_l}) & if \quad p(I_{01,x_l}) < A_{d,x_l} \end{cases}$$

**Step 5: (assign mass from $I_{02,x_l}$ and $I_{12,x_l}$) to area $i$**

So far, mass from $p(I_{02,x_l})$ has only be assigned to $\breve{e}_{02,x_l}$ and $\breve{g}_{02,x_l}$. Note that only if $\breve{e}_{0,x_l}$ and $\breve{g}_{0,x_l}$ are both equal to zero there can still be mass from $p(I_{02,x_l})$ to assign. Thus, set:

$$\breve{i}_{02,x_l} = \begin{cases} A_{i,x_l} & if \quad p(I_{02,x_l}) - A_{e,x_l} + A_{g,x_l} \geq A_{i,x_l} \\ p(I_{02,x_l}) - A_{e,x_l} + A_{g,x_l} & if \quad p(I_{02,x_l}) - A_{e,x_l} + A_{g,x_l} < A_{i,x_l} \end{cases}$$

Analogously, mass from $p(I_{12,x_l})$ has only be assigned to $\breve{f}_{12,x_l}$ and $\breve{h}_{12,x_l}$. Note that

only if $\breve{f}_{1,x_l}$ and $\breve{h}_{1,x_l}$ are both equal to zero there can still be mass from $p(I_{12,x_l})$ to assign. Thus, set:

$$\breve{i}_{12,x_l} = \begin{cases} A_{i,x_l} & if \quad p(I_{12,x_l}) - A_{f,x_l} + A_{h,x_l} \geq A_{i,x_l} \\ p(I_{12,x_l}) - A_{f,x_l} + A_{h,x_l} & if \quad p(I_{12,x_l}) - A_{f,x_l} + A_{h,x_l} < A_{i,x_l} \end{cases}$$

Note that, perhaps un-intuitively, this assignment allows for the possibility that the area $A_{i,x_l}$ is assigned twice.

Four different cases span from this step:

- Case 1: $\breve{i}_{02,x_l} = A_{i,x_l}$ and $\breve{i}_{12,x_l} = A_{i,x_l}$.
- Case 2: $\breve{i}_{02,x_l} = A_{i,x_l}$ and $\breve{i}_{12,x_l} < A_{i,x_l}$.
- Case 3: $\breve{i}_{02,x_l} < A_{i,x_l}$ and $\breve{i}_{12,x_l} = A_{i,x_l}$.
- Case 4: $\breve{i}_{02,x_l} < A_{i,x_l}$ and $\breve{i}_{12,x_l} < A_{i,x_l}$.

I consider each case separately.

**Case 1-Step 6 (exhaust mass from $I_{12,,x_l}$, $I_{012,,x_l}$ and $I_{2,x_l}$ )** In this case, $A_{g,x_l}$, $A_{e,x_l}$ and $A_{i,x_l}$ have been assigned mass from $p(I_{02,x_l})$. Equivalently, $A_{h,x_l}$, $A_{f,x_l}$ and $A_{i,x_l}$ have been assigned mass from $p(I_{12,x_l})$.

Then, equality 47 implies that:

$$\begin{aligned} A_{j,x_l} =& p(I_{012,x_l}) + \left[p(I_{02,x_l}) - \breve{g}_{02,x_l} - \breve{e}_{02,x_l} - \breve{i}_{02,x_l}\right] \\ &+ \left[p(I_{12,x_l}) - \breve{h}_{12,x_l} - \breve{f}_{12,x_l} - \breve{i}_{12,x_l}\right] + \left[p(I_{2,x_l} - \breve{c}_{2,x_l}\right] \end{aligned}$$

Thus, assign:

$$\begin{aligned} \breve{j}_{012,x_l} &= \frac{p(I_{012,x_l})}{A_{j,x_l}} \\ \breve{j}_{02,x_l} &= \frac{p(I_{012,x_l})}{A_{j,x_l}} = \frac{p(I_{02,x_l}) - \breve{g}_{02,x_l} - \breve{e}_{02,x_l} - \breve{i}_{02,x_l}}{A_{j,x_l}} \\ \breve{j}_{12,x_l} &= \frac{p(I_{12,x_l}) - \breve{h}_{12,x_l} - \breve{f}_{12,x_l} - \breve{i}_{12,x_l}}{A_{j,x_l}} \\ \breve{j}_{2,x_l} &= \frac{p(I_{2,x_l}) - \breve{c}_{2,x_l}}{A_{j,x_l}} \end{aligned}$$

Note that after this step:

$$\begin{aligned} p(I_{012,x_l}) &= \breve{j}_{012,x_l} \\ p(I_{02,x_l}) &= \breve{g}_{02,x_l} + \breve{e}_{02,x_l} + \breve{i}_{02,x_l} + \breve{j}_{02,x_l} \\ p(I_{12,x_l}) &= \breve{h}_{12,x_l} + \breve{f}_{12,x_l} + \breve{i}_{12,x_l} + \breve{j}_{12,x_l} \\ p(I_{2,x_l}) &= \breve{c}_{2,x_l} + \breve{j}_{2,x_l} \end{aligned} \tag{69}$$

That is, the mass of the density $f_{Y,X|Z=z_2}$, and of all units that realize $x_l$ when $Z = z_2$ are exhausted.

**Case 1-Step 7 (assign all remaining mass)** Only mass from $p(I_{0,x_l})$, $p(I_{1,x_l})$ and $p(I_{01,x_l})$ remains to be assigned.

Note that:

$$\begin{aligned} \Psi(C_0) - &\breve{g}_{02,x_l} - \breve{e}_{02,x_l} - \breve{d}_{01,x_l} - \breve{i}_{02,x_l} - \breve{j}_{02,x_l} - \breve{j}_{012,x_l} \\ &= A_{g,x_l} + A_{h,x_l} + \left(A_{d,x_l} - \breve{d}_{01,x_l}\right) + \left(A_{j,x_l} - \breve{j}_{02,x_l} - \breve{j}_{012,x_l}\right) \\ \Psi(C_1) - &\breve{h}_{12,x_l} - \breve{f}_{12,x_l} - \breve{d}_{01,x_l} - \breve{i}_{12,x_l} - \breve{j}_{12,x_l} - \breve{j}_{012,x_l} \\ &= A_{g,x_l} + A_{h,x_l} + \left(A_{d,x_l} - \breve{d}_{01,x_l}\right) + \left(A_{j,x_l} - \breve{j}_{01,x_l} - \breve{j}_{012,x_l}\right) \end{aligned}$$

That is, the areas $\Psi(C_0)$ and $\Psi(C_1)$ (that correspond to $\mathbb{P}[X = x_l|Z = z_0]$ and $\mathbb{P}[X = x_l|Z = z_1]$) have a common unassigned area:

$$U_{x_l} = A_{g,x_l} + A_{h,x_l} + \left(A_{d,x_l} - \breve{d}_{01,x_l}\right) + \min\{A_{j,x_l} - \breve{j}_{02,x_l} - \breve{j}_{012,x_l}, A_{j,x_l} - \breve{j}_{01,x_l} - \breve{j}_{012,x_l}\}$$

Furthermore, by hypothesis, $p(I_{01,x_l}) - \breve{d}_{01,x_l}$ (that is the yet unassigned mass associated with types $I_{01,x_l}$) must be less or equal than $U$.

Then set:

$$\breve{g}_{x_l,01} = \begin{cases} A_{g,x_l} & if \quad p(I_{01,x_l}) - \breve{d}_{01,x_l} \geq A_{g,x_l}, \\ p(I_{01,x_l}) - \breve{d}_{01,x_l} & if \quad p(I_{01,x_l}) - \breve{d}_{01,x_l} < A_{g,x_l}, \end{cases}$$

$$\breve{h}_{x_l,01} = \begin{cases} A_{h,x_l} & if \quad p(I_{01,x_l}) - \breve{d}_{01,x_l} - \breve{g}_{x_l,01} \geq A_{h,x_l}, \\ p(I_{01,x_l}) - \breve{d}_{01,x_l} - \breve{g}_{01,x_l} & if \quad p(I_{01,x_l}) - \breve{d}_{01,x_l} - \breve{g}_{01,x_l} < A_{h,x_l} \end{cases}$$

$$\breve{j}_{x_l,01} = \begin{cases} U_{j,x_l} & if \quad r(I_{01,x_l}) = U_{j,x_l} \\ r(I_{01,x_l}) \quad if \quad r(I_{01,x_l}) \leq U_{j,x_l} \end{cases}$$

Where $U_{j,x_l} = \min\{A_{j,x_l} - \breve{j}_{02,x_l} - \breve{j}_{012,x_l}, A_{j,x_l} - \breve{j}_{01,x_l} - \breve{j}_{012,x_l}\}$ and $r(I_{01,x_l}) = p(I_{01,x_l}) - \breve{d}_{01,x_l} - \breve{g}_{01,x_l}, - \breve{h}_{01,x_l}$.

Any remaining unassigned mass necessarily comes from $p(I_{0,x_l})$ and $p(I_{1,x_l})$. It must, and can be assigned directly to $\breve{d}_{0,x_l}$, $\breve{g}_{0,x_l}$, $\breve{h}_{0,x_l}$, $\breve{j}_{0,x_l}$, $\breve{d}_{1,x_l}$, $\breve{g}_{1,x_l}$, $\breve{h}_{1,x_l}$ and $\breve{j}_{1,x_l}$.

This exhaust all the mass in the solution to the linear system, and all the regions in figure 8.

**Case 2-Step 6 (exhaust mass from $I_{12,,x_l}$, $I_{012,,x_l}$ and $I_{2,x_l}$ )**

Note that case 3 is symmetric to this case. So, without loss of generality, the arguments from Case 2 apply to case 3 as well.

In this case, there is still mass from $p(I_{2,x_l})$, $p(I_{12,x_l}), p(I_{012,x_l})$ that remains to be assigned.

Inequality 50 implies that that:

$$\begin{aligned} A_{g,x_l} + A_{j,x_l} &= \Psi(C_{12}) - \breve{f}_{12,x_l} - \breve{h}_{12,x_l} - \breve{i}_{12,x_l} \\ &\geq p(I_{12,x_l}) + p(I_{012,x_l}) \end{aligned}$$

That is, all remaining mass associated with $I_{12,x_l}$ $I_{012,x_l}$ and can be assigned to areas $g$ and $j$.

Set:

$$\breve{g}_{12,x_l} = \begin{cases} A_{g,x_l} & if r(I_{12,x_l}) \geq A_{g,x_l} \\ r(I_{12,x_l}) & if \quad r(I_{12,x_l}) < A_{g,x_l} \end{cases}$$

Where $r(I_{12,x_l}) = p(I_{12,x_l}) - \breve{f}_{12,x_l} - \breve{h}_{12,x_l} - \breve{i}_{12,x_l}$.

Then set:

$$\breve{g}_{12,x_l} = \begin{cases} A_{g,x_l} & if r'(I_{12,x_l}) \geq A_{j,x_l} \\ r(I_{12,x_l}) & if \quad r'(I_{12,x_l}) < A_{j,x_l} \end{cases}$$

Where $r'(I_{12,x_l}) = p(I_{12,x_l}) - \breve{f}_{12,x_l} - \breve{h}_{12,x_l} - \breve{i}_{12,x_l} - \breve{g}_{12,x_l}$. Then, set:

$$\breve{g}_{123,x_l} = \begin{cases} \tilde{r}(A_{g,x_l}) & if \quad p(I_{012,x_l}) \geq \tilde{r}(A_{g,x_l}) \\ p(I_{012,x_l}) & if \quad p(I_{012,x_l}) < \tilde{r}(A_{g,x_l}) \end{cases},$$

where $\tilde{r}(A_{g,x_l}) = A_{g,x_l} - \breve{g}_{12,x_l}$ (note that letters $r$ denote unassigned types mass while

letters $\tilde{r}$ denote areas to which no mass has been assigned).

It only remains to assign the mass from $p(I_{2,x_l})$. There are no restrictions on how to assign this mass other than it being assigned to areas below $f_{Y,X|Z=z_2}(y, x_l)$. Thus, it can an must be assigned to all such empty reminding areas.

**Case 2-Step 7 (assign all remaining mass)**

Note that an area:

$$A_{g,x_l} + \left[A_{i,x_l} - \breve{i}_{02,x_l}\right] + \left[A_{j,x_l} - \breve{j}_{012,x_l}\right] + \left[A_{d,x_l} - \breve{j}_{01,x_l}\right]$$

below $f_{Y,X|Z=z_0}(y, x_l)$ remains to be assigned among types $I_{01}$ and $I_0$.

Equivalently, the area:

$$\left[A_{g,x_l} - \breve{i}_{12,x_l}\right] + \left[A_{j,x_l} - \breve{j}_{12,x_l} - \breve{j}_{012,x_l}\right] + \left[A_{d,x_l} - \breve{j}_{01,x_l}\right],$$

below $f_{Y,X|Z=z_0}(y, x_l)$ remains to be assigned among types $I_{01}$ and $I_1$.

Note that the terms $\left[A_{j,x_l} - \breve{j}_{012,x_l}\right]$ and $\left[A_{d,x_l} - \breve{j}_{01,x_l}\right]$ are identical in both lines. On the other hand:

$$\begin{aligned} A_{g,x_l} &\geq \left[A_{g,x_l} - \breve{i}_{12,x_l}\right] \\ \left[A_{j,x_l} - \breve{j}_{012,x_l}\right] &\geq \left[A_{j,x_l} - \breve{j}_{12,x_l} - \breve{j}_{012,x_l}\right]. \end{aligned}$$

That is, the unassigned area below $f_{Y,X|Z=z_0}(y, x_l)$ is weakly larger and contains the unassigned area below $f_{Y,X|Z=z_1}(y, x_l)$. Thus, this second, smaller area can be freely assigned to types $I_{01,x_l}$.

All remaining mass can and must be assigned to types $I_{0,x_l}$ and $I_{0,x_l}$.

This completes case 2 and, by symmetry, case 3.

**Case 4-Step 6 (exhaust mass from $I_{012,,x_l}$ and $I_{2,x_l}$ )**

All remaining, unassigned mass below $f_{Y,X|Z=z_2}$ must come from $p(I_{012,x_l})$ and $p(I_{2,x_l})$.

Thus, set:

$$\breve{j}_{012,x_l} = \begin{cases} A_{j,x_l} & if \quad p(I_{012,x_l}) \geq A_{j,x_l} \\ p(I_{012,x_l}) & if \quad p(I_{012,x_l}) < A_{j,x_l} \end{cases}$$

$$\breve{g}_{012,x_l} = \begin{cases} A_{g,x_l} & if \quad p(I_{012,x_l}) - \breve{j}_{012,x_l} \geq A_{g,x_l} \\ p(I_{012,x_l}) - \breve{j}_{012,x_l} & if \quad p(I_{012,x_l}) - \breve{j}_{012,x_l} < A_{g,x_l} \end{cases}$$

$$\breve{h}_{012,x_l} = \begin{cases} A_{h,x_l} & if \quad r(I_{012,x_l}) \geq A_{h,x_l} \\ r(I_{012,x_l}) & if \quad r(I_{012,x_l}) < A_{h,x_l} \end{cases}$$

Where $r(I_{012,x_l}) = p(I_{012,x_l}) - A_{j,x_l} - A_{g,x_l}$.

Now define $r'(I_{012,x_l}) = p(I_{012,x_l}) - A_{j,x_l} - A_{g,x_l} - \breve{h}_{012,x_l}$.

If $r'(I_{012,x_l}) > 0$, the only mass left below $f_{Y,X|Z=z_2}$ comes from $p(I_{012,x_l})$ and $p(I_{2,x_l})$, the only areas left are $A_{i,x_l} - \breve{i}_{02,x_l}$ and $A_{i,x_l} - \breve{i}_{12,x_l}$.

If both $\breve{i}_{02,x_l}$ and $\breve{i}_{12,x_l}$ are zero, then, a mass of at most $A_{i,x_l}$ remains unassigned from $p(I_{012,x_l})$. This mass can be assigned to $A_{i,x_l}$ and then $p(I_{2,x_l})$ is freely assigned to the remainder.

If $\breve{i}_{02,x_l} > 0$ then $\breve{e}_{02,x_l} = A_{e,x_l}$. This implies that the unassigned mass from $p(I_{012,x_l})$ cannot exceed $A_{i,x_l} - \breve{i}_{02,x_l}$. All remaining mass must and can be assigned to types in $I_{2,x_l}$.

If both $\breve{i}_{02,x_l}$ and $\breve{i}_{12,x_l}$ are positive, the unassigned mass from $p(I_{012,x_l})$ cannot exceed $A_{i,x_l} - \max\{\breve{i}_{02,x_l}, \breve{i}_{12,x_l}\}$. All remaining mass must and can be assigned to types in $I_{2,x_l}$.

At this point, the unassigned mass below $f_{Y,X|Z=z_0}$ is equal to:

$$U_{z_0,x_l} = A_{j,x_l} + A_{g,x_l} + A_{h,x_l} + A_{i,x_l} - \breve{j}_{012,x_l} - \breve{g}_{012,x_l} - \breve{h}_{012,x_l} - \breve{i}_{012,x_l} - \breve{j}_{02,x_l}.$$

And the unassigned mass below $f_{Y,X|Z=z_0}$ is equal to:

$$U_{z_1,x_l} A_{j,x_l} + A_{g,x_l} + A_{h,x_l} + A_{i,x_l} - \breve{j}_{012,x_l} - \breve{g}_{012,x_l} - \breve{h}_{012,x_l} - \breve{i}_{012,x_l} - \breve{j}_{12,x_l}.$$

The unassigned mass from $p(I_{01,x_l})$ cannot exceed $\min U_{z_0,x_l}, U_{z_1,x_l}$. It can be freely assigned across al remaining areas.

All unassigned mass now corresponds to types $I_{0,x_l}$ and $I_{1,x_l}$. This mass can be freely assigned.

Note that in this last step, if $\min U_{z_0,x_l}, U_{z_1,x_l}$ is minimized at 0, then an area $A_{i,x_l} - \breve{i}_{12,x_l}$ is assigned to types $I_{2,x_l}$. If instead the minimizer is 1, an area $A_{i,x_l} - \breve{i}_{02,x_l}$ is assigned to types $I_{2,x_l}$.

**Final step:** The previous sequential assignment yields equations:

$$
\begin{aligned}
A_{a,x_l} &= \breve{a}_{0,x_l} \\
A_{b,x_l} &= \breve{b}_{1,x_l} \\
A_{c,x_l} &= \breve{c}_{2,x_l} \\
2A_{g,x_l} &= \breve{g}_{0,x_l} + \breve{g}_{02,x_l} + \breve{g}_{01,x_l} + \breve{g}_{012,x_l} \\
A_{g,x_l} &= \breve{g}_{1,x_l} + \breve{g}_{01,x_l} + \breve{g}_{12,x_l} + \breve{g}_{012,x_l} \\
A_{g,x_l} &= \breve{g}_{2,x_l} + \breve{g}_{02,x_l} + \breve{g}_{12,x_l} + \breve{g}_{012,x_l} \\
A_{h,x_l} &= \breve{h}_{0,x_l} + \breve{h}_{02,x_l} + \breve{h}_{01,x_l} + \breve{h}_{012,x_l} \\
2A_{h,x_l} &= \breve{h}_{1,x_l} + \breve{h}_{01,x_l} + \breve{h}_{12,x_l} + \breve{h}_{012,x_l} \\
A_{h,x_l} &= \breve{h}_{2,x_l} + \breve{h}_{02,x_l} + \breve{h}_{12,x_l} + \breve{h}_{012,x_l} \\
A_{i,x_l} &= \breve{i}_{0,x_l} + \breve{i}_{02,x_l} + \breve{i}_{01,x_l} + \breve{i}_{012,x_l} \\
A_{i,x_l} &= \breve{i}_{1,x_l} + \breve{i}_{01,x_l} + \breve{i}_{12,x_l} + \breve{i}_{012,x_l} \\
2A_{i,x_l} &= \breve{i}_{2,x_l} + \breve{i}_{02,x_l} + \breve{i}_{12,x_l} + \breve{i}_{012,x_l} \\
A_{j,x_l} &= \breve{j}_{0,x_l} + \breve{j}_{02,x_l} + \breve{j}_{01,x_l} + \breve{j}_{012,x_l} \\
A_{j,x_l} &= \breve{j}_{1,x_l} + \breve{j}_{01,x_l} + \breve{j}_{12,x_l} + \breve{j}_{012,x_l} \\
A_{j,x_l} &= \breve{j}_{2,x_l} + \breve{j}_{02,x_l} + \breve{j}_{12,x_l} + \breve{j}_{012,x_l}
\end{aligned}
$$

Then, simply assigning a function $\frac{lt,\breve{x}_l}{A_{lt,x_l}} f_{lt,x_l}(y)$ for $lt \in \{a,b,c,d,e,f,g,h,i,j\}$ to type $I_{lt,x_l}$ ensures that condition two of lemma A.1.1 is satisfied. This completes the proof.

□

***Proof of lemma A.1.1.***

**Part 1:** The proof of part 1 is trivial, except for $f_j$.

Note that if $f_0 \leq f_1 + f_2$ then $f_g \max\{0, f_0 - \min\{f_1, f_2\}\} = 0 \leq \min\{f_0, f_1, f_2\}$

If instead $f_0 \geq f_1 + f_2$ then $f_g = \min\{f_1, f_2\} = \min\{f_1, f_2, f_3\}$. Thus $\min\{f_1, f_2, f_3\} \geq f_g$. Analogous arguments show that $f_h \leq min\{f_1, f_2, f_3\}$ and $f_i \leq min\{f_1, f_2, f_3\}$

Moreover, if $f_g > 0$ then $f_0 \geq f_1$ which implies $\max f_0, f_1 > f_1$ and thus $f_i = 0$.

Hence, $f_g > 0$ implies $f_i =$. It also implies $f_h = 0$.

Thus, $f_g + f_i + f_h \leq \min\{f_0, f_1, f_2\}$.

**Part 2:**

First note that $f_j + f_g + f_h + f_i = \min\{f_0, f_1, f_2\}$. So, $f_j + f_i + f_g + f_h + f_d = \min\{f_0, f_1, f_2\} + f_d$

If $f_d = 0$ then $\min\{f_0, f_1, f_2\} = \min\{f_1, f_2\}$ and therefore $f_d + \min\{f_0, f_1, f_2\} = \min\{f_0, f_1\}$.

If $f_d > 0$ then $f_2 = \min\{f_0, f_1, f_2\}$ and thus $f_d + \min\{f_0, f_1, f_2\} = \min\{f_0, f_1\}$.

**Part 3:**

**Case 1 ($f_g = 0$ and $f_e = 0$):** In this case, $\max\{f_1, f_2\} \geq \min\{f_0, f_1 + f_2\}$ and $f_1 \geq f_0$, $f_1 \geq f_2$. Thus, $\min\{f_0, f_1\} = f_0 = \min\{f_0, f_1 + f_2\} = f_g + f_e + \min\{f_0, f_1, f_2\}$.

**Case 2 ($f_g = 0$ and $f_e > 0$):** In this case, $\min\{f_0, f_2\} > f_1$. Thus: $\min\{f_0, f_2\} - f_1 + \min\{f_0, f_1\} = \min\{f_0, f_2\}$. If $f_0 > f_2$, then $g$ would be positive. Thus $f_2 \geq f_0$ which implies $f_g + f_e + \min\{f_0, f_1, f_2\} = f_0 = \min\{f_0, f_1 + f_2\}$.

**Case 3 ($f_g > 0$)**: If $f_0 \geq f_1 \geq f_2$ then $f_e = 0$ and $f_g + \min\{f_0, f_1, f_2\} = \min\{f_0, f_1 + f_2\}$.

If $f_0 \geq f_2 \geq f_1$ then the result holds mechanically.

□

### A.1.1 Auxiliary proposition: No need to enumerate response types

For any treatment value $x_l \in \mathcal{X}$ and any subset of instrument values $S \subset \mathcal{Z}$ define the set of response types that realize $x_l$ when $Z$ takes any value in $S$, but not for any other instrument value, as:

$$I_{S,x_l} := \{t^Z : (t^Z)_k = x_l \iff z_k \in \tilde{Z},\}$$

When $|\tilde{Z}| \geq 2$ I refer to the response types in $I_{S,x_l}$ as the group of $\tilde{Z}$-IFF-$x_l$-takers. That is, the set of response types for which the event $Z_i \in S$ is a necessary and sufficient condition to guarantee that $X_i = x_l$.

The total probability assigned to such a group under a distribution $p \in \Delta(\mathcal{T}_Z)$ over instrument-response types is given by:

$$p(\mathcal{I}_{S,x_l}) := \sum_{t^Z \in \widetilde{IFF}(x_l,\tilde{Z})} p(t^Z) \;\; \forall x_k \in \mathcal{X}, S \subset \mathcal{Z} \setminus \emptyset \tag{70}$$

Note that for every nonempty subset $S$ of $\mathcal{Z}$ a group of $S$-IFF-$x_l$-takers exists. Therefore, for every treatment value, there are $R = 2^{K_Z} - 1$ different IFF-taker groups.

Since the previous equation is linear, it can be represented in vector form as:

$$\boldsymbol{p}_{x_l,I} = C\boldsymbol{p}$$

Where $\boldsymbol{p}$ is the vector representation of $p$ and $C$ an $R$ by $|T_Z|$ matrix of zeroes and ones and $\boldsymbol{p}_{x_l,I}$ is a vector that stacks the terms on the left hand side of equation 70.

Suppose that the model is correctly specified. Then, a distribution over response types $g_0$ exists that is consistent with the observed joint distribution of $(Y_i, X_i, Z_i)$ and satisfies the restrictions from 2.2. As a consequence, for any group of IFF-takers $\mathcal{I}_{S,x_l}$, the distribution of the outcomes that units in this group realize when $X_i = x_l$ can be recovered from $g_0$ as follows:

$$q_{x_l,\tilde{Z}}(y) = \sum_{t^Z \in \mathcal{I}_{S,x_l}} g'_{x_l}(y|t^Z)$$

Where $g'_{x_l}(y|t^Z)$ denotes the Radon-Nikodym derivative of the measure over $(\mathcal{Y}, \mathcal{B}_Y)$ defined by $g\big(\{(t'^Z, t^X) \in \mathcal{T} : t'^Z = t^Z, (t^X)_l \in B_Y\}\big)$ (that is, the distribution of

potential outcomes when the treatment equals $x_l$ for units whose instrument-response type is given by $t^Z$).[22]

Hence, for every $(y, x_l, z_k) \in \mathcal{Y} \times \mathcal{X} \times \mathcal{Z}$ the observed distribution must satisfy:

$$f_{Y,X|Z=z_k}(y, x_l) = \sum_{S \ni z_k} q_{x_l,\tilde{Z}}(y) \ \ a.e.$$

That is, the observed distribution of the outcome when $X_i = x_l$ and $Z_i = z_k$ must coincide with the sum of the subdensities that describe the outcomes of the units in IFF-$x_l$-taker groups that contain $z_k$.

The previous equation is linear in all its components. Thus, all such equations associated to the same treatment value $x_l$ but different instrument values can be stacked as follows:

$$\boldsymbol{\phi}_{x_l}(y) = D\boldsymbol{q}_{x_l}(y) \ \ a.e. \tag{71}$$

Where $\boldsymbol{\phi}_{x_l} : \mathcal{Y} \longrightarrow \mathcal{R}^{|K_Z|}$ is a function that maps every outcome value to a vector that stacks the observed subdensities $f_{Y,X|Z=z_k}(y, x_l)$ across instrument values, $D$ is a $K_Z$ by $R$ matrix comprised by zeroes and ones that adds the subdensities associated to IFF-takers groups to obtain densities over outcomes at different instrument values, and $\boldsymbol{q}_{x_l} : \mathcal{Y} \longrightarrow \mathbb{R}^R$ is a function that maps every outcome value $y \in \mathcal{Y}$ to a vector comprised by all the densities associated to the potential outcomes of IFF-$x_l$ taker groups, evaluated at $y$.

Furthermore, for any function $\boldsymbol{q} : \mathcal{Y} \longrightarrow \mathbb{R}^K$ for some integer $K$, let $\mathcal{S}(\boldsymbol{q})$ denote the vector of entry-wise definite integrals of $\boldsymbol{q}$ over $\mathcal{Y}$:

$$\big(\mathcal{S}(\boldsymbol{q}_{x_l})\big)_i = \int_{\mathcal{Y}} \big(\boldsymbol{q}(y)\big)_i d\mu_Y$$

**Proposition A.1.1.** *The following statements are equivalent:*

1. *A distribution over response types $g \in \Delta(\mathcal{T})$ exists that satisfies assumption 2.1 (instrument validity), assumption 2.2 (instrument-response heterogeneity restrictions), and equation 3 (consistency with the observed treatment and outcome*

[22] The Radon-Nikodym derivative exists because the measure $g\big(\{(t'^Z, t^X) \in \mathcal{T} : t'^Z = t^Z, (t^X)_l \in B_Y\}\big)$ is dominated by the observed outcome distribution which is assumed to be absolutely continuous with respect to $\mu_Y$.

*distribution) exists.*

2. *A distribution over instrument response types $p \in \Delta(\mathcal{T}_Z)$ and entry-wise measurable functions $\boldsymbol{q}_{x_l} : \mathcal{Y} \longrightarrow \mathbb{R}^R$ exist such that $\boldsymbol{p}$ jointly satisfies proposition 2.1, and, for every $x_l \in \mathcal{X}$:*

$$\boldsymbol{\Phi}_{x_l} = D\boldsymbol{q}_{x_l}(y) \quad a.e. \qquad (\textit{Marginal consistency of IFF-takers outcomes}) \tag{72}$$

$$\mathcal{S}(\boldsymbol{q}_{x_l}) = \boldsymbol{p}_{x_l,I} \qquad (\textit{Global consistency of IFF takers probabilities}) \tag{73}$$

***Proof of proposition A.1.1.***

**1$\Longrightarrow$ 2:**

The proof follows from the previopus discussion. By hypothesis, a distribution over response types $g$ exists that is consistent with observed probabilities, and satisfies assumption 2.2.

Let $g_{x_l}(B_Y|t^Z)$ denote the measure that describes the probability that units whose instrument-response type is given by $t^Z$ realize outcomes in the measurable set $B_Y \in \mathcal{B}_Y$ when they realize treatment $x_l$:

$$g_{x_l}(B_Y|t^Z) = g\big(\{(t'^Z, t^X) \in \mathcal{T} : t'^Z = t^Z, (t^X)_l \in B_Y\}\big)$$

If $t^Z$ is such that $(t^Z)_k = x_l$ for some $k = 1, 2, ..., K_Z$, then, for such $k$:

$$\mathbb{P}[Y_i \in B_Y, X_i = x_l | Z_i = z_k] \geq g_{x_l}(B_Y|t^Z)$$

Then, since the observed outcome distribution is absolutely continuous with respect to $\mu_Y$, the measure defined by $g_{x_l}(B_Y|t^Z)$ must also be absolutely continuous with respect to $\mu_Y$. Thus, the Radon-Nikodym derivative of $g_{x_l}(B_Y|t^Z)$ with respect to $\mu_Y$ must exist. Let $g'_{x_l}(B_Y|t^Z) =$ denote one such derivative.

Then, for every nonempty $\tilde{Z} \subset \mathcal{Z}$, define:

$$q_{x_l,\tilde{Z}}(y) = \sum_{t^Z \in \mathcal{I}_{S,x_l}} g'_{x_l}(y|t^Z)$$

The function $q_{x_l,\tilde{Z}}(y)$ defines the distribution of outcomes for the $\tilde{Z}$-IFF-$x_l$-takers.

Then, the probability that an outcome in $B_Y$ and treatment $x_l$ are realized con-

ditional on $Z_i = z_k$ is equal to the sum of the probability that an outcome in $B_Y$ is realized over all the $\tilde{Z}$-IFF-$x_l$-takers groups such that $\tilde{Z}$ contains $z_k$.

$$f_{Y,X|Z=z_k} = \sum_{S \ni z_k} q_{x_l,\tilde{Z}}(y) \, a.e.$$

The previous equation is linear in all its components. Thus, all such equations associated to the same treatment value $x_l$ but different instrument values can be stacked as follows:

$$\boldsymbol{\phi}_{x_l}(y) = D\boldsymbol{q}_{x_l}(y) \;\; a.e. \tag{74}$$

Where $\boldsymbol{\phi}_{x_l} : \mathcal{Y} \longrightarrow \mathcal{R}^{|K_Z|}$ is a function that maps every outcome value to a vector that stacks the observed subdensities $\phi_{Y|x_l,z_k}(y)$ across instrument values, $D$ is a $K_Z$ by $R$ matrix comprised by zeroes and ones that adds the subdensities associated to IFF-takers groups to obtain densities over outcomes at different instrument values, and $\boldsymbol{q}_{x_l} : \mathcal{Y} \longrightarrow \mathbb{R}^R$ is a function that maps every outcome $y \in \mathcal{Y}$ to a vector comprised by all the densities associated to the potential outcomes of IFF $x_l$ taker groups, evaluated at $y$.

Now, for every function $g'_{x_l}(y|t^Z)$:

$$\begin{aligned}\int_{\mathcal{Y}} g'_{x_l}(y|t^Z) d\mu_Y &= g\big(\{(t'^Z, t^X) \in \mathcal{T} : t'^Z = t^Z, (t^X)_l \in \mathcal{Y}\}\big) \\ &= p_g\big(t^Z\big)\end{aligned}$$

Therefore:

$$\begin{aligned}\int_{\mathcal{Y}} q_{x_l,\tilde{Z}}(y) &= \int_{\mathcal{Y}} \sum_{\tilde{G}(x_l,\tilde{Z}) \ni t^Z} g'_{x_l}(y|t^Z) \\ &= \sum_{\tilde{G}(x_l,\tilde{Z}) \ni t^Z} p_g\big(t^Z\big) \\ &= p_g\big(\tilde{G}(x_l,\tilde{Z})\big)\end{aligned}$$

Or, equivalently:

$$\mathcal{S}(\boldsymbol{q}_{\boldsymbol{x_l}}) = C\boldsymbol{p} = \boldsymbol{p}_{x_l,RT}$$

**2$\Longrightarrow$ 1:**

Now I show that from valid distributions for the outcomes of the different $\tilde{Z}$-IFF-$x_l$-takers groups that satisfy aggregate instrument-response probabilities, a distribution over response types that is consistent with the model can be constructed.

Suppose that for every $x_{\in}\mathcal{X}$ a function $q_{x_l}$ and a distribution over instrument response types $\boldsymbol{p} \in \Delta(\mathcal{T}_Z)$ exist such that:

$$\begin{aligned} \boldsymbol{\Phi}_{x_l}(y) &= D\boldsymbol{q}_{x_l}(y) \quad a.e. \\ \mathcal{S}(\boldsymbol{q}_{x_l}) &= \boldsymbol{p}_{x_l,RT} \\ A_Z\boldsymbol{p} &\leq \boldsymbol{r}_Z \\ A_X\boldsymbol{p} &= \boldsymbol{r}_X \end{aligned} \tag{75}$$

For every instrument-response type $t^Z \in \mathcal{T}_Z$, define the set of treatments it can potentially realize as $X(t^Z) = \{x_k \in \mathcal{X} : \exists\ z_l \in \mathcal{Z} s.t. x_k = (t^Z)_l\}$ and let $\tilde{\Gamma}(x_l, t^Z)$ be the unique IFF-$x_l$-takers group to which $t^Z$ belongs (i.e. the unique $\mathcal{I}_{S,x_l}$ such that $\mathcal{I}_{S,x_l} \ni t^Z$). When $(t^z)_k \neq x_l$ for all $k = 0, 1, ..., K_Z$ set $\tilde{\Gamma}(x_l, t^Z) = \emptyset$.

Some response types might never realize certain treatments, nonetheless, to fully specify a distribution over response types it is necessary to define the distribution of those potential outcomes that are never observed. To this end, let $\gamma(y)$ be an arbitrary distribution over $\mathcal{Y}$ that is absolutely continuous with respect to $\mu_Y$.

Then, every response type $t^Z$, define the function $g_{x_l}(y|t^Z)$ as follows:

$$g'_{x_l}(y|t^Z) = \begin{cases} \frac{1}{p(\tilde{\Gamma}(x_l,t^Z))} q_{x_l,\tilde{Z}}(y) & \text{if } p(\tilde{\Gamma}(x_l, t^Z)) > 0 \\ 0 & \text{if } p(\tilde{\Gamma}(x_l, t^Z)) = 0 \text{ and } \tilde{\Gamma}(x_l, t^Z)) \neq \emptyset \\ \gamma(y) & \text{if } \tilde{\Gamma}(x_l, t^Z)) = \emptyset \end{cases}$$

Integrating the functions $g'_{x_l}$ allows to construct the desired distribution over response types. To this end, recall the following result from measure theory:

**Lemma A.1.3.** *Any measurable set of response types $B_{\mathcal{T}} \in \mathcal{B}_{\mathcal{T}}$ can be partitioned into slices as follows:*

$$B_{\mathcal{T}} = \bigcup_{t^Z \in \mathcal{T}_Z} \{t^Z\} \times B_{T_X}^{t^Z}$$

*Where* $B_{T_X}^{t^Z} = B_{Y_0}^{t^Z} \times B_{Y_1}^{t^Z} \times ... \times B_{Y_{L_X-1}}^{t^Z} \in \mathcal{Y}^{L_X}$ *is potentially empty and measurable with respect to the product measure of the* $L_X$ *spaces that correspond to the potential outcomes at each different treatment value* $x_l \in \mathcal{X}$.

*Moreover, each individual set* $= B_{Y_l}^{t^Z}$ *for* $l = 0, 1..., L_X - 1$ *is measurable with respect to* $\mu_Y$.

Then, using the lemma, define:

$$g\Big(B_{\mathcal{T}}\Big) = \sum_{t^Z \in \mathcal{T}_Z} \Big[ p(t^Z) \int\limits_{B_{T_X}^{t^Z}} \Big( \prod_{x_l \in X(t^Z)} g'_{x_l}(y|t^Z) \Big) d\mu_Y \Big]$$

Where $\mu_{T^X}$ denotes the product measure $\mu_{T^X} = \mu_Y \otimes \mu_Y \otimes ... \otimes \mu_Y$ over the space of potential treatment-responses. This is:

$$\int\limits_{B_{T_X}^{t^Z}} \Big( \prod_{x_l \in X(t^Z)} g'_{x_l}(y|t^Z) \Big) d\mu_Y = \int\limits_{B_{Y_0}^{t^Z}} \int\limits_{B_{Y_1}^{t^Z}} ... \int\limits_{B_{Y_{K_Z-1}}^{t^Z}} \prod_{x_l \in X(t^Z)} g'_{x_l}(y|t^Z) d\mu_Y \mu_Y ... \mu_Y$$

It remains to prove that $g\Big(B_{\mathcal{T}}\Big)$ is a distribution over reponse types, this is, it is nonnegative, adds up to 1, satisfies countable additivity, and is consistent with the observed joint distribution of outcomes, treatments and instruments..

Seeing that $g\Big(B_{\mathcal{T}}\Big)$ is nonnegative is trivial since each function $g'_{x_l}(y|t^Z)$ is nonnegative.

Moreover, equation 73 implies that if $p(\tilde{\Gamma}(x_l, t^Z)) > 0$ and $t^Z \in \tilde{G}(x_l, \tilde{Z})$, then

$$\int\limits_{\mathcal{Y}} g'_{x_l}(y|t^Z) d\mu_Y = 1 \tag{76}$$

Therefore:

$$\begin{aligned} g(\mathcal{T}) &= \sum_{t^Z \in \mathcal{T}_Z} \Big[ p(t^Z) \int\limits_{\mathcal{T}_X} \Big( \prod_{x_l \in X(t^Z)} g'_{x_l}(y|t^Z) \Big) d\mu_{T_X} \Big] \\ &= \sum_{t^Z \in \mathcal{T}_Z} \Big[ p(t^Z) \Big] \\ &= 1 \end{aligned}$$

Countable additivity follows mechanically. Let $\{B_{\mathcal{T}}^m\}_{m \in M}$ be a countable collection of

disjoint subsets of $\mathcal{T}$ indexed by $m \in M$. Applying lemma A.1.3, each sets $B^m_{\mathcal{T}}$ can be sliced as follows:

$$B^m_{\mathcal{T}} = \bigcup_{t^Z \in \mathcal{T}_Z} \{t^Z\} \times B^{t^Z,m}_{T_X}$$

Since $B^m \cap B^{m'} = \emptyset$ for all $m, m' \in M$ such that $m \neq m'$ then its must be that $B^{t^Z,m}_{T_X} \cap B^{t^Z,m'}_{T_X} = \emptyset$.

Then:

$$\begin{aligned}
g\Big(\bigcup_{m\in M} B^m_{\mathcal{T}}\Big) &= g\Big(\bigcup_{m\in M}\Big[\bigcup_{t^Z\in\mathcal{T}_Z}\Big\{\{t^Z\}\times B^{t^Z,m}_{T_X}\Big\}\Big]\Big) \\
&= g\Big(\bigcup_{t^Z\in\mathcal{T}_Z}\Big[\bigcup_{m\in M}\Big\{\{t^Z\}\times B^{t^Z,m}_{T_X}\Big\}\Big]\Big) \\
&= g\Big(\bigcup_{t^Z\in\mathcal{T}_Z}\Big[\{t^Z\}\times\Big\{\bigcup_{m\in M} B^{t^Z,m}_{T_X}\Big\}\Big]\Big) \\
&= \sum_{t^Z\in\mathcal{T}_Z} p(t^Z) \int_{\bigcup_{m\in M} B^{t^Z,m}_{T_X}} \Big(\prod_{x_l\in X(t^Z)} g'_{x_l}(y|t^Z)\Big) d\mu_{T_X} \\
&= \sum_{t^Z\in\mathcal{T}_Z}\sum_{m\in M} p(t^Z) \int_{B^{t^Z,m}_{T_X}} \Big(\prod_{x_l\in X(t^Z)} g'_{x_l}(y|t^Z)\Big) d\mu_{T_X} \\
&= \sum_{m\in M}\sum_{t^Z\in\mathcal{T}_Z} p(t^Z) \int_{B^{t^Z,m}_{T_X}} \Big(\prod_{x_l\in X(t^Z)} g'_{x_l}(y|t^Z)\Big) d\mu_{T_X} \\
&= \sum_{m\in M} g(B^m_{\mathcal{T}})
\end{aligned}$$

The first line is comes from application of lemma A.1.3 to every set $B^m_{\mathcal{T}}$, the second simply inverts the order of the unions, the third is a property of cartesian products, the fourth simply applies the definition of $g$ and takes advantage of the fact that the set is already sliced, the fifth line uses countable additivity of the integral, the sixth changes the order of the sums, and the last one uses the definition of $g$ but now applied to each set $B^m_Y$ separately. This proves countable additivity.

It only remains to verify consistency with observed probabilities, let $B^{t^Z}_{T_X} = \mathcal{Y} \times .... \times B_Y \times ... \times \mathcal{Y}$ be the obtained by taking the product of the outcome's support $L_X - 1$ times and letting the $l' - th$ element of the product coincide with $B_Y$ instead

of $\mathcal{Y}$. Then:

$$
\begin{aligned}
g\Big(\big\{(t^Z,t^X)\in\mathcal{T}:(t^Z)_k=x_{l'},(t^X)_l\in B_Y\big\}\Big) &= \sum_{t^Z:(t^Z)_k=x_l}\Big[p(t^Z)\int\limits_{B_{T_X}^{t^Z}}\Big(\prod_{x_l\in X(t^Z)}g'_{x_l}(y|t^Z)\Big)d\mu_Y\Big]\\
&= \sum_{t^Z:(t^Z)_k=x_l}\Big[p(t^Z)\Big(\int\limits_{\mathcal{Y}}\dots\int\limits_{B_Y}\dots\int\limits_{\mathcal{Y}}\prod_{x_l\in X(t^Z)}g'_{x_l}(y|t^Z)\Big)d\mu_Y\Big]\\
&= \sum_{t^Z:(t^Z)_k=x_l}\Big[p(t^Z)\int\limits_{B_Y}g'_{x_{l'}}(y|t^Z)d\mu_Y\Big]\\
&= \sum_{t^Z:(t^Z)_k=x_l}\Big[\frac{p(t^Z)}{p(\tilde{\Gamma}(x_l,t^Z))}\int\limits_{B_Y}q_{x_l,\tilde{Z}}(y)d\mu_Y\Big]\\
&= \sum_{\tilde{Z}\ni x_l}\Big[\sum_{t^Z:\tilde{\Gamma}(x_l,t^Z)=\tilde{Z}}\Big(\frac{p(t^Z)}{p(\tilde{\Gamma}(x_l,t^Z))}\int\limits_{B_Y}q_{x_l,\tilde{Z}}(y)d\mu_Y\Big)\Big]\\
&= \sum_{\tilde{Z}\ni x_l}\Big[\int\limits_{B_Y}q_{x_l,\tilde{Z}}(y)d\mu_Y\Big)\Big]\\
&= \int\limits_{B_Y}f_{Y,X|Z=z_k}(y,x_l)d\mu_Y
\end{aligned}
$$

The first equality is just the definition of $g$, the second expands the integral as a sequence of integrals over the different dimensions, the third uses 76, the fourth uses the definition of $g'$, the fifth decomposes all instrument-response types that realize $x_l$ when $Z_i=z_k$ in terms of the IFF-taker groups to which they belong, the sixth uses the definition of IFF-taker groups, and the last follows from equation 75.

This completes the proof. □

## A.2 Proofs for section 4

### A.2.1 Proof for subsection 4.1

Before providing the proofs I formally define the transportation problem as well as the maximum flow and minimum cut problems.

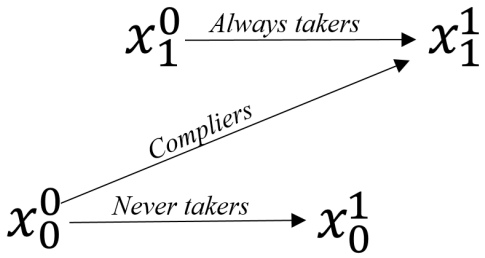
First define the sets $\mathcal{X}_k$ for $k=0,1$ as copies of $\mathcal{X}$ but with elements $x_l^k\in\mathcal{X}_k$ labeled with the superscript $k$ to make explicit that $x_l^0$ refers to treatment $x_l\in\mathcal{X}$

realized when $Z = 0$ and $x_l^1$ to treatment $x_l \in \mathcal{X}$ realized when $Z = 1$.

Next, define the directed graph with vertex set $\mathcal{X}_0 \cup \mathcal{X}_1$ and let the edges be $\mathcal{E}_R \subset \mathcal{X}_1 \times \mathcal{X}_2$, where the ordered pair $(x_l^0, x_{l'}^1) \in \mathcal{X}_0 \times \mathcal{X}_1$ belongs to $\mathcal{E}_R$ if and only if the instrument-response type $t^Z = (x_l, x_{l'})$ does not belong to $R$. In order words, an edge $(x_l^0, x_{l'}^1)$ belongs to the graph if the response type that realizes $x_l$ at $Z = 0$ and $x_{l'}$ at $Z = 1$ is not ruled out by assumption 4.1.

For instance, in the case of the no defiers assumption with a binary treatment, the set $R$ would consist of the singleton $\{(x_1, x_0)\}$. Therefore, the edges $(x_0^0, x_0^1), (x_0^0, x_1^1)$ and $(x_1^0, x_1^1)$ –corresponding to never takers, compliers, and always takers– would belong to $\mathcal{E}_R$, in contrast, the edge $(x_1^0, x_0^1)$ would not. I illustrate this particular case in figure 9.

Figure 9: Graph representation of the no-defiers instrument-response type restriction.




***Notes:*** *This figure depicts the graph representation of the no-defiers/monotonicity assumption from Imbens and Angrist (1994). The vertices with a super-index $k = 0, 1$ represent treatment values realized at instrument values $Z = z_k$. The edges with subindex $l = 0, 1$ represent treatments $x_0$ and $x_1$ respectively. Since there is no edge from $x_1^0$ to $x_0^1$, defiers are ruled out from the model.*

The transportation problem is as follows: every node $x_l^0 \in \mathcal{X}_0$ starts with a mass $\mathbb{P}[X_i = x_l | Z_i = z_0]$ of units. Each unit can travel from its origin node $x_l^0$ through the admissible edges $(x_l^0, x_{l'}^1) \in \mathcal{E}_R$ in order to reach some destination node $x_{l'}^1 \in \mathcal{X}_1$ subject to the constraint that each destination node $x_{l'}$ cannot receive more than $\mathbb{P}[X_i = x_{l'} | Z_i = z_1]$ units. If it is possible to empty all origin nodes $x_l^0$ so that all destination nodes end at full capacity, then, the transportation problem is solved.

In what follows, I prove that if the transportation problem is solved, a distribution over response types exists that is consistent with both the observed joint distribution of $(X, Y)$ and the imposed response type restriction.

Throught the definitions and proofs I alternate between edges of the graphs that define the transportation problem and response types. Alternating is possible because there is a one-to-one correspondence between the response types $t^Z$ that belong to $\mathcal{T}_Z \backslash R$

and edges $e \in \mathcal{E}_R$. This correspondence is given by $t^Z = (x_l, x_{l'}) \longrightarrow e = (x_l^0, x_{l'}^1)$. I denote the unique edge associated to $t^Z \in \mathcal{T}_Z$ as $e_{tz}$ and the unique instrument-response type associated to $e \in \mathcal{E}_R$ as $t_e^Z$. Here I introduce a version of the problem taylored to the instrument validity application.

**Definition A.2.1 (Circulation with demands problem associated to R ).** *For a detailed discussion of discrete circulation and transportation problems see Kleinberg and Tardos (2006) chapter 7.5.*

*Let $R \subset \mathcal{T}_Z$ be a collection of instrument-response types with associated graph $\mathcal{E}_R$. Define demands $d_{(x_l,z)}$ for $x_l^k \in \mathcal{X}_0 \cup \mathcal{X}_1$ as follows:*

$$d_{(x_l,z)} = \begin{cases} -\mathbb{P}[X = x_l | Z = z_k] & \text{if } k = 0 \\ \mathbb{P}[X = x_l | Z = z_k] & \text{if } k = 1 \end{cases}$$

*Furthermore, every edge $e = (x_l^0, x_{l'}^1) \in \mathcal{E}_R$ has a maximum capacity $c_e \in [0, 1]$ defined as $c_e = c_e((x_l^0, x_{l'}^1))$.*

*A candidate solution to the circulation with capacities problem is a function $\tilde{p} : E_R \longrightarrow \mathbb{R}_+$ such that, for each directed edge $e = (x_l^0, x_{l'}^1) \in \mathcal{E}_R$, the number $\tilde{p}(e)$, represents the flow from node $x_l^0$ to node $x_{l'}^1$. The flow $\tilde{p}$ solves the problem if:*

1. *Capacity constraints are satisfied:*

$$0 \leq \tilde{p}(e) \leq c_e \text{ for all } e \in E_R$$

2. *Demand is satisfied at every node (note that supply is represented as negative demand):*

$$d_{x_l^0} = -\sum_{e \in E^{out}(x_l^0)} \tilde{p}(e) \qquad \text{for all } l = 1, 2, ..., L-1$$
$$d_{x_{l'}^1} = \sum_{e \in E^{in}(x_{l'}^1)} \tilde{p}(e) \qquad \text{for all } l = 1, 2, ..., L-1$$

*Where $E^{out}(x_l^0) = \{e \in \mathcal{E}_R : e = (x_l^0, x_{l'}^1) \text{ for some } x_{l'}^1 \in \mathcal{X}_1\}$ and $E^{in}(x_{l'}^1) = \{e \in \mathcal{E}_R : e = (x_l^0, x_{l'}^1) \text{ for some } x_l^1 \in \mathcal{X}_0\}$ represent the sets of edges that depart from $x_l^0$ and end in $x_{l'}^1$ respectively.*

**Definition A.2.2 (Maximum flow problem associated to R).** *Let $R \subset \mathcal{T}_Z$ be a collection of instrument-response types with associated graph $\mathcal{E}_R$. Let $(\mathcal{X}_0 \cup \mathcal{X}_1, \mathcal{E}_R)$ be the graph from definition A.2.1. From theorem 7.50 in Kleinberg and Tardos (2006) it follows that any circulation with demands problem is equivalent to a maximum flow problem. The maximum flow problem associated to R is obtained by applying this principle to the associated circulation with demands problem.*

*The directed graph for the maximum flow problem is defined as follows:*

*Two vertices are added: a source $s^*$ and a sink $t^*$ to obtain the set of vertices $\mathcal{X}_0 \cup \mathcal{X}_1 \cup \{s^*, t^*\}$.*

*A directed edge points from the source to every node $x_l^1$ and from every node $x_{l'}^1$ to the sink:*

$$\mathcal{E}_R^{MaxFlow} = \mathcal{E}_R \bigcup \Big\{(s^*, x_l^0) : x \in \mathcal{X}\Big\} \bigcup \Big\{(x_{l'}^1, t^*\ ) : x \in \mathcal{X}\Big\}$$

*The capacity of all edges of the form $e = (x_l^0, t^*)$ is set to $c_e = \mathbb{P}[X = x_l | Z = z_1]$, the capacity of all edges of the form $e = (s^*, x_{l'}^1\ )$ is set to $c_e = \mathbb{P}[X = x_{l'} | Z = z_0]$. All remaining edges have the same capacities as in definition A.2.1.*

*A flow is a function $\tilde{p} : \mathcal{E}_R^{MaxFlow} \longrightarrow \mathbb{R}_+$ such that every intermediate node (i.e. all nodes except $s^*$ and $t^*$) has exactly the same inflow and outflow. Formally:*

1. *For every $x_l^0 \in \mathcal{X}_0$:*

$$\sum_{e' \in E^{out}(x_l^0)} \tilde{p}(e') = \tilde{p}\big((s^*, x_l^0)\big)$$

2. *For every $x_{l'}^1 \in \mathcal{X}_1$:*

$$\sum_{e' \in E^{in}(x_{l'}^1)} \tilde{p}(e') = \tilde{p}\big((x_{l'}^1, t^*)\big)$$

*Let $\tilde{P}$ denote the set of all flows. The problem is:*

$$\underset{\tilde{p} \in \tilde{P}}{max.} \sum_{x_l^0 \in \mathcal{X}_0} \tilde{p}\big((s^*, x_l^0)\big)$$

$$\text{subject to: } \tilde{p}(e) \leq c_e.$$

**Definition A.2.3** ($s^* - t^*$ **cut of** $\left(\mathcal{X}_0 \cup \mathcal{X}_1 \cup \{s^*, t^*\}, \mathcal{E}_R^{MaxFlow}\right)$). *Let* $\left(\mathcal{X}_0 \cup \mathcal{X}_1 \cup \{s^*, t^*\}, \mathcal{E}_R^{MaxFlow}\right)$ *be the graph from definition A.2.2.*

1. *A cut* $C = (S_C, T_C)$ *is a partition (with elements* $S$ *and* $T$*) of the set* $\mathcal{X}_0 \cup \mathcal{X}_1 \cup \{s^*, t^*\}$ *such that* $s^* \in S_C$ *and* $t^* \in T_C$*. I denote the set of all cuts of* $\left(\mathcal{X}_0 \cup \mathcal{X}_1 \cup \{s^*, t^*\}, \mathcal{E}_R^{MaxFlow}\right)$ *as* $\mathcal{C}(R, \mathcal{X})$*.*

2. *The cut set of* $C = (S_C, T_C)$ *denoted* $\Xi_C$*, is defined as the set of edges that, if removed, would disconect* $S_C$ *from* $T_C$*, formally:*

$$\Xi_C = \{e = (e_1, e_2) \in \mathcal{E}_R^{MaxFlow} | e_1 \in S_C \text{ and } e_2 \in T_C\}$$

3. *The capacity of a cut* $C$*, denoted* $Cap(C)$*, is defined as the sum of the capacities of the edges in* $\Xi_C$*:*

$$Cap(C) = \sum_{e \in \Xi_C} c_e$$

4. *The minimum cut associated to* $R$ *is* $\min_{C \in \mathcal{C}(R,\mathcal{X})} Cap(C)$

**Proposition A.2.1.** *The following statements are equivalent:*

1. *A distribution over instrument response types exists that satisfies instrument independence ( see assumption 2.1), assumption 4.1, and* $p(x_l, x_{l'}) \leq c_e = c_e((x_l^0, x_{l'}^1))$ *exists.*

2. *A solution to the circulation with demands problem from definition A.2.1 exists.*

3. *The maximum flow attainable in the problem from definition A.2.2 is equal to 1.*

4. *The minimum cut attainable in definition A.2.3 is equal to 1.*

***Proof of proposition A.2.1.*** I prove the equivalence between 1 and 2. The equivalence of 2, 3 and 4 follows from well known results in operations research (See Kleinberg and Tardos (2006) theorems 7.9 and 7.50 for reference).

$2 \Longrightarrow 1$ :

Let $\tilde{p}$ be a flow that satisfies the circulation with demands problem associated to $R$.

For any instrument-response type $t^Z = (x_l, x_{l'}) \in \mathcal{T}_Z$ define the distribution over instrument-response types $p$ as follows:

$$p(t^Z) = p\big((x_l, x_{l'})\big) := \begin{cases} \tilde{p}(e_{tz}) & \text{if } e_{tz} \in \mathcal{E}_R \\ 0 & \text{otherwise} \end{cases}$$

Consider any node $x_l^0$:

$$\begin{aligned} -\mathbb{P}[X = x|Z = 0] = d_{x_l^0} & \\ &= -\sum_{e \in E^{out}(x_l^0)} \tilde{p}(e) \\ &= -\sum_{x_{l'} \in \mathcal{X}} p((x_l, x_{l'})) \end{aligned}$$

The first two equalities follow from the definition of the transportation problem, the last from $p(t^Z) = \tilde{p}(e_{tz})$.
An analogous argument shows that for $z = 1$:

$$\mathbb{P}[X = x_l|Z = z_1] = \sum_{x_{l'} \in \mathcal{X}} p((x_{l'}, x_l))$$

Which shows that $p$ is consistent with observed conditional treatment probabilities.

Capacity constraints are trivially satisfied since $\tilde{p}(e_{tz}) \leq c_e = c_e((x_l^0, x_{l'}^1))$.

Finally, from the definition of $\mathcal{E}_R$ it follows that $\tilde{p}$ only assigns nonnegative flows to edges in $e \in \mathcal{E}_R$, therefore, $p$ only assigns nonnegative probabilities to response types $t^Z \in R$. This shows consistency with assumption 4.1.

$\underline{1 \Longrightarrow 2:}$

Let $p$ be a distribution over response types that satisfies assumptions (2.1) and (4.1). For every edge $e = (x_l^0, x_{l'}^1) \in \mathcal{E}_R$ define the flow $\tilde{p}$ as follows:

$$\tilde{p}(e) = \tilde{p}\big((x_l^0, x_{l'}^1)\big) = p(x_l, x_{l'})$$

Note that $e$ is always well defined since every response type is assigned nonnegative probability.

Now, consider an arbitrary node $x_l^0 \in \mathcal{X}_0$:

$$\begin{aligned} d_{x_l^0} &= -\mathbb{P}[X = x_l | Z = z_0] \\ &= -\sum_{x_{l'} \in \mathcal{X}} p((x_l, x_{l'})) \\ &= -\sum_{e \in E^{out}(x_l^0)} \tilde{p}(e) \end{aligned}$$

An analogous argument shows that for $x_{l'}^1$:

$$d_{(x_l,z)} = \sum_{e \in E^{in}(x_{l'}^1)} \tilde{p}(e)$$

Thus, demand is satisfied at every node. Again, capacity constraints are trivially satisfied since $\tilde{p}((x_l^0, x_{l'}^1)) \leq c_e = c_e((x_l^0, x_{l'}^1))$. This completes the proof.

□

***Proof of theorem 4.1.***

**Part 1:** First note that the restriction $c_{tz} = 1$ simply states that instrument-response probabilities are not bounded from above other than by the requirement that they are less than 1. This is, response types in $R$ can occur with arbitrary frequency in the population.

$\Longrightarrow$

$$\begin{aligned} \mathbb{P}[X \in S | Z = z_0] &= p(\{t^Z \in \mathcal{T}^Z : t^Z = (x_l, x_{l'}) \in S \times \mathcal{X}\}) \\ &= \sum_{x_{l'} \in \mathcal{X}} p(\{t^Z \in \mathcal{T}^Z : t^Z = (x_l, x_{l'}) \in S \times \{x_{l'}\}\}) \\ &= \sum_{x_{l'} \in \{x_{l'} \in \mathcal{X} : (s, x_{l'}) \notin R \text{ and } s \in S\}} (\{t^Z \in \mathcal{T}^Z : t^Z = (x_l, x_{l'}) \in S \times \{x_{l'}\}\}) \\ &= \sum_{x_{l'} \in L_{\geq R}(S)^C} p(\{t^Z \in \mathcal{T}^Z : t^Z = (x_l, x_{l'}) \in S \times \{x_{l'}\}\}) \qquad (77) \\ &\leq \sum_{x_{l'} \in L_{\geq R}(S)^C} \mathbb{P}[X = x_{l'} | Z = 1] \\ &= \mathbb{P}[X \in L_{\geq R}(S)^C | Z = z_1] \end{aligned}$$

The first equality comes from the consistency of $p$ with observed treatment probabili-

ties, the second from decomposing the probability in terms of all the different response types that realize $x_l$ when $Z = 0$, the third equality comes from the fact that $p$ assigns zero probability to response types outside of $R$, and the fourth noting that if $(x_l, x_{l'}) \notin R$ then $x_{l'} \in L_{\geq^R}(S)^C$. The inequality comes from noting that $p$ is consistent with observed probabilities and $\{x_{l'}\}$ could also be realized when $Z = 1$ by types whose potential outcome when $Z = 0$ lies outside of $S$. The last equality comes from adding disjoint probabilities across the elements of $L_{\geq^R}(S)^C$.

$\Longleftarrow$

I prove the counterpositive. If no $p$ that satisfies the hypotheses exists, then, a violation of equation 12 must exist. To this end I use the minimum cut characterization.

First, note that by setting $C = \{e \in \mathcal{E}_R : e = (s^*, x_l^0) \textit{ for some } x_l \in \mathcal{X}_0\}$ a cut of capacity one can always be attained. Therefore, edges of the form $(x_l^0, x_{l'}^1)$ will never be part of a minimum cut since their capacity is 1 and removing a single edge will not induce a cut. Thus, minimum cut sets will only contain edges of the form $(s^*, x_l^0)$ or $(x_{l'}^1, t^*)$.

Suppose that $C$ is a minimum cut such that $Cap(C) < 1$. Let $S$ be the set of all $x_l \in \mathcal{X}$ such that the edge $(s^*, x_l^0)$ belongs to $\Xi_C$. Then, by construction, any node of the form $(s^*, x_{l'}^0)$ for $x_{l'} \in S^C := \mathcal{X} \setminus S'$ does not belong to the cut.

Define $S' := \mathcal{X} \setminus S$ and note that by construction for any $x_{l'} \in S'$ and any $x_{l''} \notin \mathcal{X}$ such that $(x_{l'}, x_{l''}) \in R$, the edge $(x_{l''}^1, t^*)$ belongs to $C$. Otherwise, the path $(s^*, x_{l'}), (x_{l'}, x_{l''}), (x_{l''}, t^*)$ would connect $t^*$ and $s^*$.

But the set of elements $x_{l''} \in \mathcal{X}$ such that $(x_{l'}, x_{l''}) \notin R$ for some $x_{l'} \in S'$ is actually the set $L_{\geq^R}(S')^C$.

Therefore:

$$\begin{aligned}
&\sum_{c_e \in \Xi_C} c_e < 1 \\
\Longrightarrow &\sum_{e \in \mathcal{E}_R^{MaxFlow} \cap \{s^*\} \times S} c_e + \sum_{e \in \mathcal{E}_R^{MaxFlow} \cap L_{\geq^R}(S') \times \{t^*\}} c_e < 1 \\
\Longrightarrow &\sum_{e \in \mathcal{E}_R^{MaxFlow} \cap L_{\geq^R}(S') \times \{t^*\}} c_e < 1 - \sum_{e \in \mathcal{E}_R^{MaxFlow} \cap \{s^*\} \times S} c_e \\
\Longrightarrow &\sum_{e \in \mathcal{E}_R^{MaxFlow} \cap L_{\geq^R}(S') \times \{t^*\}} c_e < \sum_{e \in \mathcal{E}_R^{MaxFlow} \cap \{s^*\} \times S'} c_e \\
\Longrightarrow &\mathbb{P}[X \in L_{\geq^R}(S') | Z = z_1] < \mathbb{P}[X \in S' | Z = z_0]
\end{aligned}$$

The first implication follows from the previous discussion. The second from re-arranging terms. The third from $S' = S^C$ and the fourth from the definition of the capacities. The last equation is a violation of equation 12.

**Part 2:** $\Longrightarrow$: Suppose that a distribution over instrument-response types $p \in \Delta(\mathcal{T}_Z)$ exists that satisfies assumptions 1 and 3.

For $\emptyset \neq S \subset \mathcal{X}$, let $(\Lambda_S, \Lambda'_S)$ be a partition of $L_{\geq^R}(S)^C$ such that the following three conditions are satisfied:

1. $\emptyset \neq \Lambda'_S \subset S$.
2. For every $x_{l'} \in \Lambda'_S$ the set $L_{\geq^R}(x_{l'})^C \setminus \{x_{l'}\}$ is contained in $\Lambda_S$.
3. $L_{\geq^R}(S \setminus \Lambda'_S)^C \subset \Lambda_S$.

Take equation 77:

$$\begin{aligned}
\mathbb{P}[X \in S | Z = z_0] &= \sum_{x_{l'} \in L_{\geq^R}(S)^C} p(\{t^Z \in \mathcal{T}^Z : t^Z = (x_l, x_{l'}) \in S \times \{x_{l'}\}\}) \\
&= \sum_{x_{l'} \in \Lambda} p(\{t^Z \in \mathcal{T}^Z : t^Z = (x_l, x_{l'}) \in S \times \{x_{l'}\}\}) \\
&\qquad + \sum_{x_{l'} \in \Lambda'} p(\{t^Z \in \mathcal{T}^Z : t^Z = (x_l, x_{l'}) \in S \times \{x_{l'}\}\}) \\
&= \sum_{x_{l'} \in \Lambda} p(\{t^Z \in \mathcal{T}^Z : t^Z = (x_l, x_{l'}) \in S \times \{x_{l'}\}\}) \\
&\qquad + \sum_{x_{l'} \in \Lambda'} p(\{t^Z \in \mathcal{T}^Z : t^Z = (x_l, x_{l'}) \in (S \setminus \Lambda'_S) \times \{x_{l'}\}\}) \\
&\qquad + \sum_{x_{l'} \in \Lambda'} p(\{t^Z \in \mathcal{T}^Z : t^Z = (x_l, x_{l'}) \in \Lambda'_S \times \{x_{l'}\}\})
\end{aligned}$$

The second term from the last equation is equal to zero by condition 3. Condition 2 implies that the only possible value that $x_l$ can take in the last term is $x_{l'}$. Therefore,

the equation becomes:

$$\begin{aligned}\mathbb{P}[X \in S|Z = z_0] = &\sum_{x_{l'} \in \Lambda} p(\{t^Z \in \mathcal{T}^Z : t^Z = (x_l, x_{l'}) \in S \times \{x_{l'}\}\}) \\ &+ \sum_{x_{l'} \in \Lambda'} p(\{t^Z \in \mathcal{T}^Z : t^Z = (x_l, x_l) \text{ such that } x_l \in \Lambda'\}) \\ \leq &\mathbb{P}[X \in \Lambda|Z = z_1] + \sum_{x_l \in \Lambda'} p((x_l, x_l)) \\ \leq &\mathbb{P}[X \in \Lambda|Z = z_1] + \sum_{x_l \in \Lambda'} \Psi_{x_l}\end{aligned}$$

The first inequality follows from consistency of $p$ with observed probabilities and the fact that response types that realize treatments outside of $S$ when $Z = 0$ might also realize treatments in $\Lambda$ when $Z = 1$ along with the fact that $L_{\geq^R}(x_l)^C \setminus \{x_l\} \subset \Lambda$ for $x_l \in \Lambda'$. The second inequality follows from theorem 3.1.

$\Longleftarrow$

As for part 1, I prove the contrapositive statement. That is, I suppose that a cut with capacity less than 1 exists and show that it implies that inequality (13) from theorem (4.1) is not satisfied for some $S \subset \mathcal{X}$.

Now the capacities of edges of the form $(x_l^0, x_l^1)$ are potentially binding. For cuts that only contain edges connected to $s^*$ or $t^*$ the argument from part 1 still applies. Edges of the form $(x_l^0, x_{l'}^1)$ for $x_l \neq x_{l'}$ still have capacity 1 and therefore will not belong to a minimum cut that is strictly less than one.

Suppose that $C$ is a minimum cut with at least one element of the form $(x_l^0, x_l^1)$.

Let $S'$ be the set of elements in $x_l \in \mathcal{X}$ such that $(s^*, x_l^0) \in C$. Define $S = \mathcal{X} \setminus S'$. Note that for any $x_l \in S$ and any $x_{l'} \in L_{\geq^R}(x_l)^C$ one of the following sets of conditions must hold:

1. The following holds:
   - $(x_{l'}^1, t^*) \in C$
2. The following three conditions hold
   - $x_l = x_{l'} \in S \cap L_{\geq^R}(S)^C$
   - $(x_{l'}^0, x_{l'}^1) \in C$.
   - $(x_{l''}^1, t^*) \in C$ for all $x_{l''} \in L_{\geq^R}(x_{l'})^C \setminus \{x_{l'}\}$

The first condition prevents paths of the form $(s^*, x_l^0), (x_l^0, x_{l'}^1), (x_{l'}^1, t^*)$ as in part 1.

If condition 1 is not satisfied and $(x_{l'}^1, t^*) \notin C$, then, the path $(s^*, x_l^0), (x_l^0, x_{l'}^1), (x_{l'}^1, t^*)$ must be cut at the middle at edge $(x_l^0, x_{l'}^1)$. Because edges associated to $x_l \neq x_{l'}$ cannot belong to the cut, it must be that $x_l = x_{l'} \in S' \cap L_{\geq^R}(S')^C$. Furthermore, for every $x_{l''} \in L_{\geq^R}(x_l)^C \setminus \{x_l\}$, the path $(s^*, x_l^0), (x_l^0, x_{l''}^1), (x_{l''}^1, t^*)$ must be blocked at some point. By construction, this can only happen at $(x_{l''}^1, t^*)$, thus, $(x_{l''}^1, t^*) \in C$.

In other words, a path $(s^*, x_l^0), (x_l^0, x_{l'}^1), (x_{l'}^1, t^*)$ might be blocked by condition 1, if not, then, conditions 2 must hold.

Let $\Lambda'_S$ be the set $\Lambda'_S = \{x_l \in \mathcal{X} : (x_l^0, x_l^1) \in C\} \subset L_{\geq^R}(S)^C \cap S$. Set $\Lambda_S = L_{\geq^R}(S)^C \setminus \Lambda'$. By construction the sets $\Lambda'_S$ and $\Lambda_S$ partition $L_{\geq^R}(S)^C$.

Note that $\{x_{l'} \in \mathcal{X} : (x_{l'}, t^*) \in C\}$ must be a subset of $L_{\geq^R}(S)^C$, otherwise, $x_{l'}^1$ would not be reachable from $\{x_l^0 : x_l \in S\}$ and eliminating $(x_{l'}, t^*)$ from $C$ would attain a cut with strictly lower capacity. Thus, $\{x_{l'} \in \mathcal{X} : (x_{l'}, t^*) \in C\} \subset \Lambda_S$.

Similarly, note that if $x_{l'} \in \Lambda_S$ it must be that $x_l \in \{x_{l'} \in \mathcal{X} : (x_{l'}, t^*) \in C\}$, otherwise, a path through some $x_l \in S$ such that $L_{\geq^R}(x_l)^C \cap \Lambda_S \neq \emptyset$ to $x_{l'}$ would connect $s^*$ to $t^*$. Hence:

$$\{x_{l'} \in \mathcal{X} : (x_{l'}, t^*) \in C\} = \Lambda \tag{78}$$

Finally, suppose that for some $x_l \in S \setminus \Lambda'_S$ exists some $x_{l'} \in L_{\geq^R}(x_l)^C \cap \Lambda'_S \neq \emptyset$. By definition it must be that $x_{l'} \neq x_l$. Then, the path $(s^*, x_l), (x_l, x_{l'}), (x_{l'}, t^*)$ would connect $s^*$ to $t^*$. Hence, it must be that:

$$L_{\geq^R}(S \setminus \Lambda'_S)^C \subset \Lambda_S \tag{79}$$

Then:

- $\emptyset \neq \Lambda' \subset S$ (By conditions 2).
- For all $x_{l'} \in \Lambda'$, $L_{\geq^R}(x'_l)^C \setminus \{x_{l'}\} \subset \Lambda$ (By conditions 2 and equation (78)).
- $L_{\geq^R}(S \setminus \Lambda'_S)^C \subset \Lambda_S$ by expression (79).

Moreover, it must be that:

$$\begin{aligned}
&\sum_{c_e \in \Xi_C} c_e < 1 \\
\implies &\sum_{c \in C \cap \{s^*\} \times S'} c_e + \sum_{\{c \in C : c = (x_l^0, x_l^1) \ s.t. \ x_l \in \Lambda'\}} c_e + \sum_{\{c \in C : c = (x_l^1, t^*) \ s.t \ x_l \in \Lambda\}} c_e < 1 \\
\implies &\sum_{\{c \in C : c = (x_l^0, x_l^1) \ s.t. \ x_l \in \Lambda'\}} c_e + \sum_{\{c \in C : c = (x_l^1, t^*) \ s.t \ x_l \in \Lambda\}} c_e < 1 - \sum_{c \in C \cap \{s^*\} \times S'} c_e \\
\implies &\sum_{\{c \in C : c = (x_l^0, x_l^1) \ s.t. \ x_l \in \Lambda'\}} c_e + \mathbb{P}[X \in \Lambda | Z = 1] < \mathbb{P}[X \in S | Z = z_0]
\end{aligned}$$

Where the first implication follows from the previous discussion, the second from rearranging terms and the last from the definition of the capacities.

Setting $c_e((x_l^0, x_l^1)) = \Psi_{x_l}$ completes the proof.

□

### A.2.2 Proofs for subsection 4.2

To prove theorem 4.2 I begin by formally defining latent preference types. To this end, let $\mathcal{L}_s(\mathcal{X})$ denote the set of all strict preference relations over $\mathcal{X}$ and let $\mathcal{N}(x_l) = \{\succ \in \mathcal{L}_s(\mathcal{X}) : x_l \succ x_{l'} \text{ for all } x_{l'} \in \mathcal{X}\}$ denote the subset of strict preference relations that has $x_l$ as a maximizer.

**Definition A.2.4** (Latent preference types)**.** *A latent preference type $t^\succ \in \mathcal{T}_\succ$ is a function that maps instrument realizations to strict preference profiles: $t^\succ : \mathcal{Z} \longrightarrow \mathcal{L}_s(\mathcal{X})$.*

*Note that $t^\succ(z)$ is a preference relation. Therefore, the equation $x_l t^\succ(z) x_{l'}$ reads as "latent preference types $t^\succ$ strictly prefer $x_l$ over $x_{l'}$ when $Z = z$". To improve readibility, I denote $t^\succ(z)$ as $\succ^{t(z)}$ so that the previous equation becomes $x_l \succ^{t(z)} x_{l'}$.*

**Definition A.2.5** (Distribution over latent preference types)**.** *A distribution over latent preference types is a function $h : \mathcal{H} \longrightarrow [0, 1]$ that satisfies:*

$$\begin{aligned}
h(t^\succ) \geq 0 \text{ for all } t^\succ \in \mathcal{T}_\succ \\
\sum_{t^\succ \in \mathcal{T}_\succ} h(t^\succ) = 1
\end{aligned}$$

**Remark A.2.1.** *Let $h : \mathcal{T}_{\succ} \longrightarrow [0,1]$ be a distribution over latent preference types. The distribution over response types implied by $h$ is the function $p_h : \Omega \longrightarrow [0,1]$ defined as:*

$$p_h(x_l, x_{l'}) = \sum_{t^{\succ} \in N(x_l, x_{l'})} h(t^{\succ})$$

*Where $N(x_l, x_{l'}) = \{t^{\succ} \in \mathcal{T}_{\succ} | t^{\succ}(0) \in \mathcal{N}(x_l) \text{ and } t^{\succ}(1) \in \mathcal{N}(x_{l'})\}$.*

**Definition A.2.6.** *A distribution over latent preference types $h$ is $\geq^*$-Submonotone if $h(t^{\succ}) = 0$ for every $t^{\succ}$ that violates point 3 of assumption 4.2. This is, types $t^{\succ}$ such that $x_l \geq^* x_{l'}$ and $x_l \succ^{t(z_0)} x_{l'}$ but $x_{l'} \succ^{t(z_1)} x_l$ are not allowed.*

**Assumption A.2.1 (No $S$ always takers).** *Let $S \subset \mathcal{X}$ and $R \subset \mathcal{T}_Z$ a collection of instrument-response types . $R$ has no $S$ always takers if $(x_l, x_l) \in R$ for all $x_l \in S$.*

**Lemma A.2.1.** *Let $h$ be a distribution over latent preference types. Let $R$ be a collection of instrument-response types. Let $\geq^*$ be a binary relation (not necesarilly transitive or complete) over $\mathcal{X}$. Let $Ref_{\geq^*} = \{x_l \in \mathcal{X} | x_l \geq^* x_l\}$ denote the elements that belong to the reflexive part of $\geq^*$. Let $\geq^*_{Ref}$ and $\geq^*_{Irr}$ denote the reflexive and irreflexive part of $\geq^*$ respectively. Then:*

1. *$h$ satisfies $\geq^*_{Irr}$-Submonotonicity if and only if $h$ satisfies $\geq^*$-Submonotonicity.*
2. *If $\geq^*_{Irr} \setminus \geq^R \neq \emptyset$, then, a distribution over latent preference types $h$ exists such that $h$ satisfies $\geq^R$-Submonotonicity and $p_h(t^Z) = 0$ for all $t^Z \in R$ but $h$ is not $\geq^*-$Submonotone.*
3. *If $\geq^R \setminus \geq^*{}_{Irr} \neq \emptyset$, then, a distribution over latent latent preference types $h$ exists that is $\geq^*$-Submonotone but $p_h(t^Z) > 0$ for some $t^Z \in R$.*
4. *If $\geq^R{}_{Ref} \neq \emptyset$ a distribution over latent response types exists that is $\geq^R$-Submonotone but $p_h(t^Z) > 0$ for some $t^Z \in R$.*
5. *If $\geq^R$-Submonotonicity holds and no $R_{\geq^M}$ always takers holds, then, $p(t^Z) = 0$ for all $t^Z \in R$.*
6. *If a distribution over response types $p$ that is consistent with observed conditional treatment probabilities and $p(t^Z) = 0$ for all $t^Z \in R$ exists, then, a distribution $h$ over latent preference types exists such that $h$ is $\geq^R$-Submonotone and $p = p_h$.*

*This last equality implies that $h$ is consistent with the data and that if $p$ satisfies no $S$ always takers, so does $p_h$.*

*Proof of lemma A.2.1.*

**Part 1:**

Suppose $x_l \geq^* x_{l'}$ and $x_l \succ_1^{(0)} x_{l'}$. If $x_l = x_{l'}$, then $x_l \succ_1^{(1)} x_{l'}$ holds trivially, otherwise, $\geq^*_{Irr} -QM$ implies that $x_l \succ_1^{(1)} x_{l'}$. The converse is trivial since $\geq^*_{Irr} \subset \geq^*$.

**Part 2:**

Take the ordering induced by the indexing of the elements of $\mathcal{X}$. Let $\succ^{x_l,x_{l'},*}$ be the preference relation that ranks $x_l$ as the best alternative, $x_{l'}$ as the second best, and all remaining alternatives according to their index: $x_l \succ^{x_l,x_{l'},*} x_{l'} \succ^{x_l,x_{l'},*} x_1 \succ^{x_l,x_{l'},*} ... \succ^{x_l,x_{l'},*} x_{L-1}$.

Suppose that $x_l \neq x_{l'}$ and $(x_l, x_{l'}) \in \geq^*_{Irr} \setminus \geq^R$. Let $t^{\succ}$ be the unique latent preference type that satisfies $t^{\succ}(0) = \succ^{x_l,x_{l'},*}$ and $t^{\succ}(1) = \succ^{x_{l'},x_l,*}$. Define $h$ as:

$$\begin{aligned} h(t^{\succ}) &= 1 \\ h(t^{\succ'}) &= 0 \qquad \text{for all } t^{\succ'} \neq t^{\succ} \end{aligned}$$

Note that $(x_l, x_{l'}) \notin \geq^R$ implies that $h$ is $\geq^R$ $-$Submonotone, but $h$ is not $\geq^*$ $-$Submonotone (since preference reversals between $x_l$ and $x_{l'}$ occur with positive probability). Moreover, note that $p_h$ assigns zero probability to response types outside of $R$.

**Part 3:**

If $Ref_{\geq^R}$ is nonempty, a reasoning analogous to part 1 proves the claim. Otherwise, pick some $x_l \in Ref_{\geq^*}$ and define $h$ as follows:

$$\begin{aligned} h((\succ^{x_l,*}, \succ^{x_l,*})) &= 1 \\ h(t^Z) &= 0 \qquad \text{for all } t^Z \neq (\succ^{x_l,*}, \succ^{x_l,*}) \end{aligned}$$

$h$ is $\geq^*$-QM but $p_h$ assigns positive probability to the response type $(x_l, x_l) \notin R$.

**Part 4**

This part follows immediately from part 3 by setting $\geq^*_{Irr} = \emptyset$.

**Part 5:**

Let $R'$ be the set of response types ruled out by $\geq^R$-QM and no stayers at $Ref_{\geq^R}$. The goal is to show that $R \subset R'$.

Suppose that $(x_l, x_{l'}) \in R$. This implies that $x_l \geq^R x_{l'}$.

If $x_l \neq x_{l'}$, consider a decision maker $i$ such that $X(0) = x_l$, then, rationality implies that $x_l = X(0) \succ^{(0)} x_{l'}$. Therefore, by $\geq^R$-Submonotonicity, it must be that $x_l \succ^{(1)} x_{l'}$. This implies that $x_{l'}$ is not chosen by decision maker $i$ when $Z = 1$ a.s.. Therefore $(x_l, x_{l'}) \in R'$.

If $x_{l'} = x_l$, then, no stayers at $Ref_{\geq^R}$ implies that $(x_l, x_{l'}) \in R'$.

Therefore, $t^Z \in R \Longrightarrow t^Z \in R'$.

**Part 6:**

Let $t^{\succ}_{x_l,x_{l'},*}$ be the latent preference type such that $t^{\succ}_{x_l,x_{l'},*}(0) = \succ^{x_l,x_l{}',*}$ and $t^{\succ}_{x_l,x_{l'},*}(1) = \succ^{x_{l'},x_l,*}$. Let $t^{\succ} = t^{\succ}_{x_l,*}$ denote the latent preference type that always prefers $x_l$ and then all remaining alternatives in descending order. Define $h$ as follows:

$$h(t^{\succ}) = \begin{cases} p(x_l, x_{l'}) & if\ t^{\succ} = t^{\succ}_{x_l,x_{l'},*} \quad and x_l \neq x_{l'} \\ p(x_l, x_l) & if\ t^{\succ} = t^{\succ}_{x_l,*} \\ 0 & otherwise \end{cases}$$

Note that $h$ is $\geq^R$-QM and that for all $(x_l, x_l') \in R_Z$:

$$p_h((x_l, x_l')) = \sum_{t^{\succ} \in N(x_l,x_l')} h(t^{\succ}) = p(x_l, x_{l'})$$

□

Now I comment Lemma 1. The first part shows that only the irreflexive part of binary relations is relevant for submonotonicity. Part 2 shows that when a binary relation $\geq^*$ is stronger than the one induced by $R$, restricting response types to $R$ does not guarantee $\geq^*$-submonotonicity. Parts 3 and 4 show the converse, that imposing weaker than than $\geq^R$ dos not guarantee that response types outside of $R$ are ruled out. Part 5 shows that $R$ can be guaranteed by imposing $\geq^R$-submonotonicity and ruling out a specific set of always takers. 6 shows that $\geq^R$-submonotonicity is only falsified if the response type restriction associated to $R$ is falsified. Thus, in the binary instrument case $\geq^R$-submonotonicity is just an interpretation for response type restrictions but yields no additional identifying power.

Theorem 4.2 follows immediately from lemma 6 by noting there is a unique correspondence between $R$ and $\geq^R$ and therefore their cardinalities.

## A.3 Examples of shape restrictions encompassed by the framework

**Multivalued ordered monotonicity:** Ordered monotonicity captures the idea that treatment values are strictly ordered (e.g, by intensity) and that as the instrument increases, units may only switch to weakly higher treatment levels. That is, for any pair of indices $k' > k > 0$, response types such that $(t^Z)_k > (t^Z)_{k'}$ are ruled out. This restriction can be easily accommodated in the framework of assumption 2.2 by imposing that response types as the one described before occur with zero probability.

In their seminal paper about IV with heterogeneous treatment effects, Imbens and Angrist (1994) showed that if both the instrument and treatment are binary, and treatment is weakly monotonic in the instrument, the classical Wald estimand identifies a LATE for units that strictly increase their realized treatment when the instrument changes. This result has motivated the use of binary instruments and treatments in a wide range of empirical applications.

In a subsequent paper, Angrist and Imbens (1995) extended their analysis to multivalued ordered treatments. In this setting, no LATE is point identified, but the Wald estimand continues to identify a convex combination of LATEs among units that are induced to take strictly higher treatments.

Despite the interpretation provided by Angrist and Imbens (1995), it is common practice in applied work to binarize monotone multivalued treatments. That is, the researcher selects a threshold $\bar{x} \in \mathcal{X}$ and defines a binary treatment indicator $X^{Bin} = \mathbf{1}_{\{X \geq \bar{x}\}}$, thereby reducing the setting to the binary treatment case. A justification for this practice is that the resulting Wald estimand has a simpler interpretation.

However, this practice has two limitations, first, there are instances in which the binarized model is rejected by a specification test, even though the correctly specified multivalued model is not. Second, after binarization, the Wald estimand identifies a weighted sum of LATEs for certain response types, but there is no guarantee that the weights add up to one. As a result, the estimand cannot generally be interpreted as LATE or even as a convex combination of LATEs. An in-depth analysis of the second limitation is provided in Rose and Shem-Tov (2024).

To illustrate the first limitation, suppose that $\mathcal{X} = \{x_0, x_1, x_2\}$, $\mathcal{Y} = \{0, 1\}$ and $\mathcal{Z} = \{z_0, z_1\}$. Assume that half the population has the instrument-response type

$(x_0, x_1)$, one fourth $(x_1, x_2)$, and the last fourth $(x_2, x_2)$[23]. Suppose that potential outcomes are homogeneous so that $Y(x_0) = 0$, $Y(x_1) = 1$ and $Y(x_2) = 1$ for all $i \in \mathcal{I}$. Now suppose that a researcher sets $\bar{x} = x_2$. It is easy to see that units with instrument-response type $(x_0, x_1)$ will not change their treatment status with respect to the artificially defined treatment variable $X^{Bin}$ as $Z$ increases, but, their potential outcomes will, violating the exclusion restriction in the binarized model.

Furthermore, Kitagawa (2015) shows that if instrument validity holds in the binary case, the following equation must be satisfied:

$$\mathbb{P}[Y = 1, X^{Bin} = 0 | Z = 0] \geq \mathbb{P}[Y = 1, X^{Bin} = 0 | Z = 1]$$

In the example, the left hand side is equal to $\frac{1}{4}$ while the right hand side is $\frac{1}{2}$, thus, instrument validity would be rejected for the binarized model, even if it holds in the true multivalued model with the right (multivalued ordered monotone) instrument-response restriction.

Furthermore, as studied by Rose and Shem-Tov (2024), even if equation 80 is not violated, the Wald estimand obtained using $Z$ as an instrument for $X^{Bin}$ may be difficult to interpret. To see this problem, suppose that the instrument-response types $(x_0, x_1)$, $(x_1, x_2)$, and $(x_2, x_2)$ occur with equal probability $\frac{1}{3}$ in the population. In this case, the Wald estimand becomes:

$$W = \frac{\frac{1}{3}\mathbb{E}[Y(x_1) - Y(x_0) | i \in (x_0, x_1)] + \frac{1}{3}\mathbb{E}[Y(x_2) - Y(x_1) | i \in (x_1, x_2)]}{\frac{1}{3}} \tag{80}$$

While the numerator is a weighted sum of LATEs, the weights do not necessarily sum to one. Consequently, the estimand $W$ cannot be interpreted as LATE or a convex combination of LATEs. Moreover, if instead of $\frac{1}{3}$, the denominator, i.e. the population probability associated with instrument-response type $(x_1, x_2)$, was small, the magnitude of the Wald estimand could be arbitrarily large. Thus, even if binarizing a monotone multivalued treatment may simplify the model, it can produce estimands that are both biased and difficult to interpret.

These concerns provide a strong motivation for retaining a correctly specified monotone model when the treatment is multivalued.

Generalizing monotonicity to settings with multiple instruments is nontrivial. An-

[23]Recall that the instrument-response type $t = (x_l, x_{l'})$ represents units that realize treatment $x_l$ when $Z = z_0$ and $x_{l'}$ when $Z = z_1$.

grist and Imbens (1995) proposed one approach, which imposes binary monotonicity in one direction between every pair of treatment values. This definition was later criticized by Mogstad et al. (2021) who proposed an alternative extension. More recently, Goff (2020) introduced a tractable notion –vector monotonicity– which ensures identification of a series of LATEs.

Classical monotonicity and its three aforementioned generalizations can be expressed as restrictions that rule out specific response types, that is, they fall within assumption 2.2.

**Unordered monotonicity and its refinements:** Heckman and Pinto (2018) proposed an alternative notion of monotonicity for settings in which no natural ordering over treatment values exists. Intuitively, unordered monotonicity requires that, as the instrument changes from one value to another, a subset of treatments $X^I \subset \mathcal{X}$ are equally incentivized relative to the treatments in $\mathcal{X} \setminus \mathcal{X}^I$. This restriction not only implies that take-up of the incentivized treatments must weakly increase but also rules out unit-level switches within the set of promoted treatments $\mathcal{X}^I$, or its complement $\mathcal{X} \setminus \mathcal{X}^I$. This is another restriction on heterogeneity obtained from ruling out particular response types and thus, it can be accommodated within the framework of this paper.

Heckman and Pinto (2018) argue that the identifying power of unordered monotonicity can be strengthened by further restricting the set of admissible response types. For example, suppose that $x_0 \in \mathcal{X}^I$ represents a discounted car, and $x_1, x_2 \in \mathcal{X} \setminus \mathcal{X}^I$ represent, respectively, a similar full-priced car and a pick-up truck. One could impose the restriction that only individuals who would have chosen $x_1$ in the absence of the discount may switch to $x_0$ as a result of the discount, ruling out switches from $x_2$ to $x_0$. Unordered monotonicity augmented with this type of restriction remains within the framework of this paper.

**Other response type restrictions:** The identifying power of response type restrictions extends beyond ordered and unordered monotonicity. Goff (2024) shows that the ability of restrictions on treatment-response heterogeneity to point indetify some LATE without restricting the outcome space is underpined by the set of treatment-response types it allows. Leveraging this insight, Goff (2024) identifies a series of novel response-type restrictions that guarantee point identification of at least one LATE, and a procedure to find other such restrictions.

To advance the interpretability of IV models with multiple instruments and treatments, Lee and Salanié (2020) investigated identification when multiple instruments

are available and each instrument targets (i.e. incentivizes) adoption of one particular treatment more than the rest. In particular, Lee and Salanié (2020) provide a general framework and then –to achieve point or partial identification– introduce the notions of strict targeting, and one-to-one targeting. Multiple instances of these two assumptions fit into the framework studied in this paper.

In a closely related paper, Bai et al. (2024c) study the case in which every treatment value is uniquely incentivized by a single instrument. For this particular targeting structure they show that multiple LATEs are point identified and derive the sharp testable implications of the model. However, the assumption that each instrument uniquely targets one –and only one– treatment (without affecting any other comparison) might be overly restrictive in multiple empirical settings. The framework developed in this paper can be used to assess the validity of the complete model proposed by Bai et al. (2024c), as well as more plausible relaxations.

**Other linear restrictions on instrument-response types:** Rather than ruling out specific response types, a less restrictive approach to IV identification is to place bounds on the prevalence or relative frequency of certain behaviors or instrument-response types. A notable example is the "more compliers than defiers" condition of De Chaisemartin (2017). Another example is Richardson and Robins (2010), who derive bounds on the outcome distribution under assumptions on the fraction of always takers in the population.

Similar assumptions are also used as intermediate steps in sensitivity analysis methods that assess the robustness of IV estimands to relaxations of monotonicity (Huber (2014), Noack (2021), Yap (2025)). The tests from this paper can be used to falsify these types of restrictions on their own or combined with response-type restrictions.

## A.4 Examples of submonotonicity

**Ordered monotonicity:** Ordered monotonicity captures the notion that an implicit ranking (a total order) of alternatives exists, and, as the instrument increases, adoption of higher ranked alternatives is weakly incentivized. This assumption is implicit in most random utility models in which a mean structural utility function is common to all members of the population and unit-specific unobservables are additively separable. Thus, holding unobserved heterogeneity fixed, as observables change all decision makers agree on which alternatives yield higher utility levels and thus are being promoted.

Because subomotonicity allows for general binary relations, ordered monotonicity is a particular case.

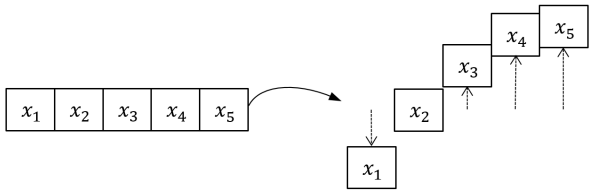



Figure 10: Ordered monotonicity

***Notes:*** *Visual representation of ordered monotonicity. The left of each figure represents alternatives before the intervention ($Z = z_0$). The right represents the effect of the intervention ($Z = z_1$). The higher an alternative is depicted on the left side, the more it is being incentivized. The black arrows represent the relative magnitude of the incentives.*

**Unordered monotonicity:** Unordered monotonicity seeks to abstain from imposing an ex-ante ordering of alternatives and instead simply acknowledging that instruments incentivize adoption of some treatments relative to the rest. This idea is formalized by assuming that a subset of treatments $S \subset \mathcal{X}$ exists such that all alternatives in $S$ are being promoted and no alternative in its complement $S' := \mathcal{X} \setminus S$ is.

While this assumption is general in spirit and in the multiple instruments setting unordered monotonicity is not implied nor implies ordered monotonicity, in the binary instrument case, it is a strictly more restrictive assumption. The reason is that unorderdered monotonicity not only requires that alternatives in $S$ are being promoted, but also that they are being promoted with exactly the same intensity so that switches within the set $S$ (and its complement) occur with zero probability.

The previous point is easy to see by noting that the binary relation associated to unordered monotonicity is defined as follows:

$$x_l \geq^R x_{l'} \iff \begin{cases} x_l, x_{l'} \in S & or \\ x_l, x_{l'} \in S' & or \\ x_l \in S, x_{l'} \in S' \end{cases} \tag{81}$$

This restriction induces a partial order with two indifference classes $S$ and $S'$, so that if $x_l, x_{l'} \in S$ then $x_l \geq^R x_{l'}$ and $x_{l'} \geq^R x_l$. The name partial order is misleading in this setting –it suggests weakening a total order–. The weakening is achieved by imposing that different treatments belong to the same equivalence class, that is, adding more ordered pairs to the binary relation which imposes additional restrictions on heterogeneity.

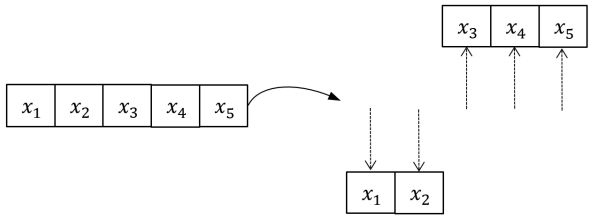



Figure 11: Unordered monotonicity

***Notes:*** *Visual representation of unordered monotonicity. The left of each figure represents alternatives before the intervention ($Z = z_0$). The right represents the effect of the intervention ($Z = z_1$). The higher an alternative is depicted, the more it is incentivized. The black arrows represent the relative magnitude of the incentives. The set $S = \{x_3, x_4, x_5\}$ represents the alternatives that are being incentivized relative to $S' = \{x_1, x_2\}$. Note that comparisons between alternatives within the same set are not altered.*

Furthermore, note that if $S$ has only one element $s$ (i.e. $|S| = 1$), then, an instrument that encourages adoption of $s$ (Lee and Salanié (2020)) or strictly targets $s$ (Bai et al. (2024c)) satisfies a particular case of unordered monotonicity. Targeting or encouragement are not implied nor imply unordered monotonicity in the multiple instruments case.

The sets $S$ and $S'$ could be further partitioned into three sets instead of two. For example, let $S = S_1 \cap S_2$ and $S' = S'_1 \cap S'_2$ and let alternatives in $S_1$ and $S'_1$ represent compact cars and alternatives in $S_2$ and $S'_2$ SUVs. Suppose that $S$ represents the set of all discounted vehicles (SUVs and compact) and $S'$ the set of vehicles without a discount. In this setting, one plausible restriction is that the discount on compact cars only affects decision makers who, absent the discount, would buy alternatives in $S'_1$; and the discount on SUVs only affects decision makers who would pick alternatives in

$S_2'$. Thus, switches from $S_1'$ to $S_2$ or $S_2'$ to $S_1$ are assumed to occur with zero probability. This restriction is no longer represented by a partial order, but it is derived from adding ordered pairs to the partial order described in equation 81, thus, it is a submonotone restriction associated to the extension of a partial order.

**Relaxations of unordered monotonicity:** The assumption that all alternatives are equally discounted might be too strong, instead, one could assume that there are $N$ discount bins and that comparisons within the same bin are not affected but comparisons across different bins are –always in favor of alternatives with larger discounts. For example, if $N = 3$ there could be a set of heavily discounted cars, a set of cars with payment facilities, and a set of cars that have to be payed upfront. This situation also corresponds to submonotonicity with respect to a partial order but now with three indifference classes.

Additional restrictions, as the one considered in the compact cars and SUVs examples could also be added across all three discount bins. The resulting restriction would remain within the submontonicity framework.

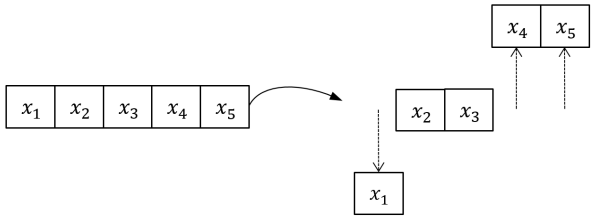



Figure 12: Submonotonicity with respect to a partial order with three indifference classes

***Notes:*** *Visual representation of submonotonicity with respect to a partial order with three indifference classes. The left of each figure represents alternatives before the intervention ($Z = z_0$). The right represents the effect of the intervention ($Z = z_1$). The higher an alternative is depicted, the more it is incentivized. The black arrows represent the relative magnitude of the incentives. The set $\{x_4, x_5\}$ represents the alternatives that are being incentivized the most (heavily discounted), the set $\{x_2, x_3\}$ the alternatives that for which payment facilities are offered (i.e. weakly incentivized). Alternative $x_1$ is not being incentivized in any way. Note that comparisons between alternatives within the same set are not altered and thus, their difference in height remains equal.*

**Relaxations of encouragement:** Now consider an intervention in which an instrument encourages adoption of certain treatments in $S \subset \mathcal{X}$, but there is a concern that the instrument might also affect utility from other treatments. For example, consider an an intervention aimed at increasing sports participation by providing information. It is clear that such an intervention could be used as an instrument for sports practice, but the intervention may also promote other activities that develop

motor or social skills such as community service or playing board-games. Hence, the restriction that the instrument is not affecting comparisons between treatments other than practicing sports might not be plausible and a weaker assumption might be preferred.

One such weaker assumption could be that alternatives $x_1$ and $x_2$ which represent sports are certainly being promoted relative to all others (since they are the main target of the intervention), but pairwise comparisons between other treatments might also affected and the direction of this effect is heterogeneous across the population. Thus, restrictions on the direction of the flows between non-sports activities cannot be credibly imposed. This situation is encoded by submonotonicity with respect to a quasi-order, i.e. a binary relation that is transitive but not complete.

If $x_5$ and $x_4$ represent different sports and $x_3$, $x_2$ and $x_1$ other activities, the specific submonotonicity assumption would be $x_l \geq^R x_{l'}$ for $l = 5, 4$ and $l' = 3, 2, 1$ but neither $x_l \geq^R x_{l'}$ nor $x_{l'} \geq^R x_l$ for $l, l' = 5, 4$ or $l, l' = 3, 2, 1$. I illustrate this in figure 13.

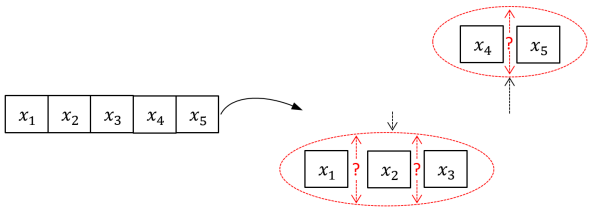



Figure 13: Submonotonicity with respect to a quasi-order - Sports example

***Notes:*** *Visual representation of submonotonicity with respect to a quasi-order. The left of each figure represents alternatives before the intervention ($Z = z_0$). The right represents the effect of the intervention ($Z = z_1$). The higher an alternative is depicted on the right side, the more it is incentivized. The set $\{x_4, x_5\}$ represents the alternatives that are being incentivized the most (sports), the set $\{x_1, x_2, x_3\}$ the alternatives that are not being incentivized. The black arrows represent the relative magnitude of the incentives but the red arrows represent relative uncertainty about this effect within the set of circled alternatives. That is, within the red elipses no restriction on how pairwise comparisons are affected is imposed.*

Note that relaxing the completeness requirement allows for more heterogeneity. In particular, it allows decision makers to change their relative ranking of the same pair of alternatives in different directions. This situation cannot be captured by a random utility model in which the effect of $Z$ is common to al decision makers, instead, it is allowing the effect of $Z$ to be determined by unobserved unit characteristics in the spirit of a random coefficients model.

Now consider an intervention aimed at preventing students from dropping college to take part-time jobs. Suppose this goal is achieved by providing additional classes

aimed at developing professional qualifications but also by providing information on the importance of college completion.

It is clear that such an intervention would promote staying in college relative to taking part-time jobs, however, the classes could help students find full-time jobs, even without the degree. On the other hand, the information could reduce the willingness of students to accept any type of job. If the latter effect is stronger for full-time jobs (because they are more demanding and thus increase the risk of not finishing college), some students might choose to leave a full-time job and take a part-time one. Thus, it might be desirable to only impose the restriction that college is being incentivized relative to part-time jobs, but nothing else.

This restriction can also be accomodated by submonotonicity by assuming that college degrees $x_4$ and $x_5$ are being incentivized relative to half-time jobs $x_3$ and $x_2$ but no restrictions involving full-time jobs $x_1$ are imposed. Formally, $x_l \geq^R x_{l'}$ for $l = 1, 2$ and $l' = 3, 4$.

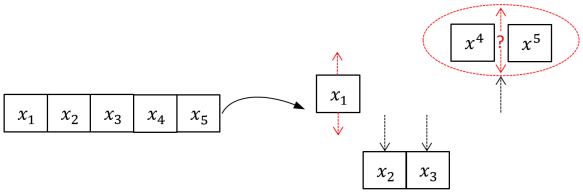



Figure 14: Submonotonicity with respect to a transitive quasi-order - College and full-time jobs example.

***Notes:*** *Visual representation of submonotonicity with respect to a quasi-order. The left of each figure represents alternatives before the intervention ($Z = 0$). The right represents the effect of the intervention ($Z = 1$). The higher an alternative is depicted, the more it is incentivized. The set $\{x_4, x_5\}$ represents the alternatives that are being incentivized the most (sports), the set $\{x_1, x_2, x_3\}$ the alternatives that are not being incentivized. The black arrows represent the relative magnitude of the incentives but the red arrows represent relative uncertainty about this effect relative to alternatives in the same class. Thus, within the red circle, no stance is taken on how pairwise comparisons are affected. Furthermore, no stance on whether $x_1$ is being promoted or not relative to $x_4$, $x_5$ or even $x_2$ and $x_3$ is taken.*

**Relaxing transitivity/encouraging experimentation:** The submonotonicity framework does not require transitivity of the binary relation $\geq^R$ –i.e. it allows for cycles in $\geq^R$. This observation raises the question of how can such a restriction be interpreted –specially in light of Goff (2024) who shows that some models associated to nontransitive binary relations can point identify LATEs for some specific response groups.

To answer this question, consider a three alternatives example and suppose that

$x_1 >^R x_2$, $x_2 >^R x_3$ and $x_3 >^R x_1$ (where $x_l >^R x_{l'}$ indicates $x_l \geq^R x_{l'}$ but not $x_{l'} \geq^R x_l$).

This restriction implies that:

- Any decision maker who choses $x_1$ when $Z = 0$ will now choose $x_1$ or $x_3$.
- Any decision maker who choses $x_2$ when $Z = 0$ will now choose $x_2$ or $x_1$.
- Any decision maker who choses $x_3$ when $Z = 0$ will now choose $x_3$ or $x_2$.

Furthermore, if the assumption that there are no always takers at $\mathcal{X}$ is added (i.e. all types of always takers are ruled out), then:

- Any decision maker who choses $x_1$ when $Z = 0$ will now choose $x_3$.
- Any decision maker who choses $x_2$ when $Z = 0$ will now choose $x_1$.
- Any decision maker who choses $x_3$ when $Z = 0$ will now choose $x_2$.

In the first case, choices when $Z = 0$ restrict potential choices when $Z = 1$. In the second, there is a one to one correspondence between one and the other. In both cases, if decision makers change their choice, they pick an alternative different than the one they would have chosen absent the instrument. The result is a situation where every alternative is promoted among subjects who do not choose it at the initial value of the instrument; and discouraged among those that already prefer it over the rest. The previous discussion suggests interpreting the effect of the instrument as promoting experimentation or variety.

Interpreting more general nontransitive restrictions is not straightforward, but, in all cases, submonotonicity induces restrictions on heterogeneity that can be expressed as a correspondence of choices when $Z = 0$, and thus, can be seen as restricting heterogeneity in terms of status-quo choices.

### A.4.1 Additional simulations for subsection 6.1: Full sample and K-fold cross-fitting estimators, and semi-deterministic classifiers.

I conduct a series of monte-carlo simulations to investigate the performance of classification-based full sample and K-fold cross fitting estimators for the overlap of two-densities. I explore simple estimators that replicate the method from 5, as well as an alternative estimator designed to have a standard normal distribution with a fixed variance-covariance matrix in the limit.

The key modification is to define a default pre-determined, arbitrary classification rule as well as a deterministic sequence $d_n \propto n^{-\tilde{\gamma}}$ that dominates the rate of uniform convergence for KDEs $r_n$. Then, the classifier differs from the pre-determined rule only when the estimated difference $\Delta_{n_1}(y, x_l)$ is larger than $d_n$.

The modified classifier is assymptotically optimal in the sense that it converges to a classifier that only follows the pre-determined rule in the contact set, where, as proposition 5.1 shows, the loss is invariant to classification decisions. If the sequence $d_n$ is sufficiently small so that an analogue of the margin condition from assumption 5.1.3 holds for $\tilde{\gamma}$ then, the estimator is still consistent for the overlap parameter, but, in the limit, classification decisions inside the contact set are deterministic.

For a sample of size $n$ and some fixed integer $\tilde{K} \geq 2$, a $\tilde{K}$-fold partition randomly distributes all observations into $\tilde{K}$ sub-samples or folds denoted $(J_\kappa)_{\kappa=1}^{\tilde{K}}$ of size $\left\lfloor n/\tilde{K} \right\rfloor$ each. Let, $J_\kappa^C$ denote the set of all observations except those in fold $\kappa$.

Let $\eta_{\kappa,n}^d(y, x_l)$ be the classifier trained in $J_\kappa^C$ according to the following semi-deterministic rule:

$$\eta_{\kappa,n}^d(y, x_l) := \begin{cases} \mathbf{1}_{\{\Delta_{\kappa,n}(y,x_l)\geq 0\}} & \text{if } |\Delta_{\kappa,n}(y, x_l)| \geq d_n \\ \mathbf{1}_{\{y\in D\}} & \text{otherwise} \end{cases} \tag{82}$$

Where $D$ is a fixed, pre-specified measurable subset of $\mathbb{R}^n$ and $\Delta_{\kappa,n}(y, x_l)$ is the difference between the kernel density estimators for $f_{Y,X|Z=z_1}$ and $f_{Y,X|Z=z_0}$ trained in $J_\kappa^C$. Note that $\eta_{\kappa,n}^d$ converges to the classifier $\eta^d$ that coincides with $\mathbf{1}_{\{\Delta_{\kappa,n}(y,x_l)>0\}}$ everywhere except in $\mathcal{C}_{x_l}$ where it is equal to $\mathbf{1}_{\{y\in D\}}$.

Let $L_\kappa(\eta_{\kappa,n}^d, x_l)$ denote the empirical loss of the classifier $\eta_{\kappa,n}^d$ evaluated in $J_\kappa$. Define the estimator $\Psi_{x_l,\kappa,n}^d := L_\kappa(\eta_{\kappa,n}^d, x_l)$.

Furthermore, define:

$$\boldsymbol{\theta}_n^T = \frac{1}{\tilde{K}} \sum_{\kappa=1}^{\tilde{K}} \boldsymbol{\theta}_{\kappa,n}^T. \tag{83}$$

The following simulations investigate the performance of $\boldsymbol{\theta}_n^T$ as an overlap estimator, as well as the performance of an estimator that uses the same entire sample both to estimate densities and the classifier, and to evaluate the loss of the classifier.

I find that both full-sample estimators and K-fold cross fitting estimators fail when the contact set is non-negligible. Bias is larger than for sample splitting estimators, and coverage rates fall below the nominal rate.

When the contact set is negligible, both approaches have low bias and their coverage rates in the simulations is close to the nominal rate.

Table 10: Bias, overlap coverage, and joint overlaps and CTPs coverage for semi-deterministic full sample and K-fold cross-fitting estimators. Gaussian kernel.

| | | | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $d_n = 0$ | | | | $d_n = (nh_n^{1.5})^{-1/2}$ | | | | $d_n = (nh_n^3)^{-1/2}$ | | | |
| $n = 500$ $\mathbb{E}n_{z_0} = 167$, $\mathbb{E}n_{z_1} = 333$ | | | SS | FS | 2-F | 8-F | SS | FS | 2-F | 8-F | SS | FS | 2-F | 8-F |
| $w = 0$ | $\mu_1 = 0$ | $\Psi_{x_0,n} - \Psi_{x_0}$ | 0.00 | -0.07 | 0.00 | 0.00 | 0.00 | -0.07 | 0.00 | 0.00 | 0.00 | -0.06 | 0.00 | 0.00 |
| | | $\Psi_{x_0} \in C_{0.95}$ | 0.93 | 0.55 | 0.91 | 0.87 | 0.93 | 0.58 | 0.91 | 0.87 | 0.94 | 0.67 | 0.92 | 0.91 |
| | | $\boldsymbol{\theta}_0 \in E_{0.95}$ | 0.89 | 0.37 | 0.86 | 0.77 | 0.89 | 0.37 | 0.86 | 0.77 | 0.59 | 0.50 | 0.82 | 0.76 |
| | $\mu_1 = 0.5$ | $\Psi_{x_0,n} - \Psi_{x_0}$ | 0.03 | -0.02 | 0.02 | 0.01 | 0.03 | -0.02 | 0.02 | 0.01 | 0.06 | -0.01 | 0.06 | 0.05 |
| | | $\Psi_{x_0} \in C_{0.95}$ | 0.88 | 0.94 | 0.86 | 0.88 | 0.88 | 0.95 | 0.86 | 0.88 | 0.81 | 0.98 | 0.52 | 0.54 |
| | | $\boldsymbol{\theta}_0 \in E_{0.95}$ | 0.73 | 0.94 | 0.72 | 0.79 | 0.73 | 0.92 | 0.72 | 0.79 | 0.42 | 0.94 | 0.16 | 0.22 |
| | $\mu_1 = \tau$ | $\Psi_{x_0,n} - \Psi_{x_0}$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.01 | 0.06 | 0.05 |
| | | $\Psi_{x_0} \in C_{0.95}$ | 0.94 | 0.95 | 0.94 | 0.94 | 0.94 | 0.96 | 0.94 | 0.94 | 0.71 | 0.96 | 0.26 | 0.30 |
| | | $\boldsymbol{\theta}_0 \in E_{0.95}$ | 0.92 | 0.94 | 0.92 | 0.92 | 0.92 | 0.94 | 0.92 | 0.92 | 0.56 | 0.93 | 0.10 | 0.10 |
| $w = 1$ | $\mu_1 = 0.5$ | $\Psi_{x_0,n} - \Psi_{x_0}$ | 0.02 | -0.03 | 0.02 | 0.01 | 0.02 | -0.03 | 0.02 | 0.01 | 0.03 | -0.03 | 0.03 | 0.03 |
| | | $\Psi_{x_0} \in C_{0.95}$ | 0.90 | 0.87 | 0.86 | 0.85 | 0.90 | 0.88 | 0.86 | 0.85 | 0.94 | 0.89 | 0.80 | 0.83 |
| | | $\boldsymbol{\theta}_0 \in E_{0.95}$ | 0.79 | 0.88 | 0.75 | 0.75 | 0.79 | 0.88 | 0.75 | 0.75 | 0.74 | 0.88 | 0.43 | 0.55 |
| | $\mu_1 = \tau$ | $\Psi_{x_0,n} - \Psi_{x_0}$ | 0.00 | -0.01 | 0.00 | 0.00 | 0.00 | -0.01 | 0.00 | 0.00 | 0.02 | -0.01 | 0.02 | 0.02 |
| | | $\Psi_{x_0} \in C_{0.95}$ | 0.94 | 0.92 | 0.94 | 0.91 | 0.94 | 0.93 | 0.94 | 0.91 | 0.93 | 0.94 | 0.85 | 0.87 |
| | | $\boldsymbol{\theta}_0 \in E_{0.95}$ | 0.92 | 0.90 | 0.91 | 0.90 | 0.92 | 0.92 | 0.91 | 0.90 | 0.91 | 0.94 | 0.78 | 0.77 |
| $n = 1000$ $\mathbb{E}n_{z_0} = 333$, $\mathbb{E}n_{z_1} = 666$ | | | SS | FS | 2-F | 8-F | SS | FS | 2-F | 8-F | SS | FS | 2-F | 8-F |
| $w = 0$ | $\mu_1 = 0$ | $\Psi_{x_0,n} - \Psi_{x_0}$ | 0.00 | -0.06 | 0.00 | 0.00 | 0.00 | -0.06 | 0.00 | 0.00 | 0.00 | -0.05 | 0.00 | 0.00 |
| | | $\Psi_{x_0} \in C_{0.95}$ | 0.94 | 0.44 | 0.89 | 0.87 | 0.94 | 0.50 | 0.89 | 0.87 | 0.94 | 0.57 | 0.96 | 0.91 |
| | | $\boldsymbol{\theta}_0 \in E_{0.95}$ | 0.93 | 0.21 | 0.86 | 0.78 | 0.93 | 0.23 | 0.86 | 0.78 | 0.63 | 0.32 | 0.88 | 0.72 |
| | $\mu_1 = 0.5$ | $\Psi_{x_0,n} - \Psi_{x_0}$ | 0.02 | -0.01 | 0.01 | 0.01 | 0.02 | -0.01 | 0.01 | 0.01 | 0.06 | 0.00 | 0.06 | 0.05 |
| | | $\Psi_{x_0} \in C_{0.95}$ | 0.88 | 0.95 | 0.87 | 0.90 | 0.88 | 0.96 | 0.87 | 0.90 | 0.61 | 0.98 | 0.23 | 0.32 |
| | | $\boldsymbol{\theta}_0 \in E_{0.95}$ | 0.79 | 0.94 | 0.76 | 0.85 | 0.79 | 0.95 | 0.76 | 0.85 | 0.28 | 0.94 | 0.03 | 0.05 |
| | $\mu_1 = \tau$ | $\Psi_{x_0,n} - \Psi_{x_0}$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.01 | 0.06 | 0.05 |
| | | $\Psi_{x_0} \in C_{0.95}$ | 0.95 | 0.95 | 0.94 | 0.95 | 0.95 | 0.95 | 0.94 | 0.95 | 0.45 | 0.95 | 0.04 | 0.05 |
| | | $\boldsymbol{\theta}_0 \in E_{0.95}$ | 0.92 | 0.94 | 0.92 | 0.94 | 0.92 | 0.94 | 0.92 | 0.94 | 0.29 | 0.91 | 0.00 | 0.00 |
| $w = 1$ | $\mu_1 = 0.5$ | $\Psi_{x_0,n} - \Psi_{x_0}$ | 0.01 | -0.02 | 0.01 | 0.00 | 0.01 | -0.02 | 0.01 | 0.00 | 0.02 | -0.02 | 0.02 | 0.02 |
| | | $\Psi_{x_0} \in C_{0.95}$ | 0.89 | 0.88 | 0.87 | 0.89 | 0.89 | 0.91 | 0.87 | 0.89 | 0.89 | 0.93 | 0.76 | 0.83 |
| | | $\boldsymbol{\theta}_0 \in E_{0.95}$ | 0.81 | 0.87 | 0.76 | 0.81 | 0.81 | 0.89 | 0.76 | 0.81 | 0.71 | 0.87 | 0.40 | 0.52 |
| | $\mu_1 = \tau$ | $\Psi_{x_0,n} - \Psi_{x_0}$ | 0.00 | -0.01 | 0.00 | 0.00 | 0.00 | -0.01 | 0.00 | 0.00 | 0.02 | 0.00 | 0.02 | 0.02 |
| | | $\Psi_{x_0} \in C_{0.95}$ | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 | 0.96 | 0.95 | 0.95 | 0.90 | 0.96 | 0.75 | 0.73 |
| | | $\boldsymbol{\theta}_0 \in E_{0.95}$ | 0.95 | 0.93 | 0.92 | 0.90 | 0.95 | 0.94 | 0.92 | 0.90 | 0.84 | 0.93 | 0.58 | 0.59 |
| $n = 2000$ $\mathbb{E}n_{z_0,x_l} = 666$, $\mathbb{E}n_{z_1,x_l} = 1333$ | | | SS | FS | 2-F | 8-F | SS | FS | 2-F | 8-F | SS | FS | 2-F | 8-F |
| $w = 0$ | $\mu_1 = 0$ | $\Psi_{x_0,n} - \Psi_{x_0}$ | 0.00 | -0.04 | 0.00 | 0.00 | 0.00 | -0.04 | 0.00 | 0.00 | 0.00 | -0.04 | 0.00 | 0.00 |
| | | $\Psi_{x_0} \in C_{0.95}$ | 0.94 | 0.41 | 0.90 | 0.88 | 0.94 | 0.42 | 0.90 | 0.88 | 0.96 | 0.52 | 0.95 | 0.92 |
| | | $\boldsymbol{\theta}_0 \in E_{0.95}$ | 0.94 | 0.10 | 0.89 | 0.78 | 0.94 | 0.13 | 0.89 | 0.78 | 0.64 | 0.21 | 0.86 | 0.79 |
| | $\mu_1 = 0.5$ | $\Psi_{x_0,n} - \Psi_{x_0}$ | 0.01 | -0.01 | 0.01 | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 | 0.05 | 0.00 | 0.05 | 0.04 |
| | | $\Psi_{x_0} \in C_{0.95}$ | 0.92 | 0.94 | 0.90 | 0.91 | 0.92 | 0.94 | 0.90 | 0.91 | 0.52 | 0.96 | 0.14 | 0.19 |
| | | $\boldsymbol{\theta}_0 \in E_{0.95}$ | 0.87 | 0.93 | 0.84 | 0.89 | 0.87 | 0.94 | 0.84 | 0.89 | 0.18 | 0.96 | 0.02 | 0.01 |
| | $\mu_1 = \tau$ | $\Psi_{x_0,n} - \Psi_{x_0}$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.01 | 0.05 | 0.05 |
| | | $\Psi_{x_0} \in C_{0.95}$ | 0.96 | 0.95 | 0.95 | 0.94 | 0.96 | 0.96 | 0.95 | 0.94 | 0.21 | 0.93 | 0.01 | 0.01 |
| | | $\boldsymbol{\theta}_0 \in E_{0.95}$ | 0.94 | 0.94 | 0.96 | 0.94 | 0.94 | 0.95 | 0.96 | 0.94 | 0.05 | 0.93 | 0.00 | 0.00 |
| $w = 1$ | $\mu_1 = 0.5$ | $\Psi_{x_0,n} - \Psi_{x_0}$ | 0.01 | -0.02 | 0.00 | 0.00 | 0.01 | -0.01 | 0.00 | 0.00 | 0.02 | -0.01 | 0.02 | 0.02 |
| | | $\Psi_{x_0} \in C_{0.95}$ | 0.92 | 0.90 | 0.90 | 0.91 | 0.92 | 0.91 | 0.90 | 0.91 | 0.88 | 0.92 | 0.70 | 0.77 |
| | | $\boldsymbol{\theta}_0 \in E_{0.95}$ | 0.86 | 0.90 | 0.83 | 0.85 | 0.86 | 0.91 | 0.83 | 0.85 | 0.70 | 0.92 | 0.35 | 0.44 |
| | $\mu_1 = \tau$ | $\Psi_{x_0,n} - \Psi_{x_0}$ | 0.00 | -0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.02 | 0.02 |
| | | $\Psi_{x_0} \in C_{0.95}$ | 0.95 | 0.94 | 0.94 | 0.93 | 0.95 | 0.96 | 0.94 | 0.93 | 0.82 | 0.95 | 0.52 | 0.50 |
| | | $\boldsymbol{\theta}_0 \in E_{0.95}$ | 0.95 | 0.95 | 0.94 | 0.93 | 0.95 | 0.97 | 0.94 | 0.93 | 0.67 | 0.96 | 0.25 | 0.22 |

***Notes:*** *This table shows the average bias ($\Psi_{x_l,0} - \Psi_{x_0}$), confidence interval coverage rate ($\Psi_{x_l,n} \in C_{0.95}$), and confidence ellipsoid coverage rates ($\boldsymbol{\theta}_0 \in E_{0.95}$) for overlap estimators obtained using Gaussian kernels with different combinations of tuning parameters. The sample size is represented by $n$ at the top of each panel along with $\mathbb{E}n_{z_k}$ which represents the expected number of observations per instrument value.*

*The parameters $w$ and $\mu_1$ define different data generating processes. $\mu_1$ determines the difference in means between the subdensities $f_{Y,X|Z=z_0}$ and $f_{Y,X|Z=z_1}$, and consequently, the point in which the subdensities cross. The parameter $w$ determines the width of the contact set at the point in which the subdensities cross. If $w$ is equal to zero, the contact set has zero measure. If $w > 0$ then, near the crossing points, a region in which the two densities are equal exist. When $\mu_1 = 0$ both subdensities are equal throughout the entire support of $Y$ and thus, the contact set corresponds to the entire real line.*

*SS indicates sample splitting estimators, FS full sample estimators and $K-F$ K-Fold cross-fitting estimators. The bandwidth rate is given by $h_n \propto n^{-1/5}$*