Sensitivity of Policy Relevant Treatment Parameters to Violations of Monotonicity

Luther Yap *

February 12, 2025

Abstract

This paper proposes a method to bound policy relevant treatment parameters (PRTP) when the monotonicity assumption that the instrumental variable affects individuals' treatment response in the same direction is weakened. The bounding framework uses the proportion of defiers relative to compliers as a sensitivity parameter, and yields an identified set that is an interval. The method is illustrated in an empirical application where the same-sex instrument was used to calculate the effect of having a third child on labor force participation. I find that bounds are informative only for small violations in monotonicity.

Keywords: Policy relevant treatment parameters, monotonicity

*Department of Economics, Princeton University. I thank Michal Kolesar and David Lee for helpful comments and suggestions.

1 Introduction

Since the seminal work of Heckman and Vytlacil (2005), there has been a large literature that is concerned with identification and inference of policy relevant treatment parameters (PRTP) in instrumental variable (IV) settings with heterogeneous treatment effects (TE). PRTP is a general class of objects that includes the local average treatment effect (LATE) and various TE in counterfactual environments. Existing methods that target the general class of PRTP rely on the monotonicity assumption that the instrument affects all individuals' treatment response in the same direction, which is usually imposed through an additively separable treatment selection equation (e.g., Mogstad et al. (2018)). However, monotonicity may not be realistic in many applications. Consider the Angrist and Evans (1998) study that was interested in the effect of having a third child on the mother's labor supply. They used an indicator for whether the first two kids are of the same sex as an instrument for the third child. Since parents have a preference for gender balance among their children, families with two boys or two girls are more likely to have a third child. But some parents may want two sons or two daughters, so they would violate monotonicity, which rules out families who would have a third child if their first two children are of the same sex, and would not have a third child if their first two children are of different sex. Further examples of monotonicity failure are considered in De Chaisemartin (2017). This observation raises the question of how much bounds on PRTP would change when monotonicity fails. This paper explicitly places a bound on the extent that monotonicity fails, which nests approaches that either impose or drop monotonicity as special cases.

The goal is to place bounds on PRTP while accommodating limited violations of mono-

tonicity. Sensitivity restrictions characterize these violations: I use a sensitivity parameter that places an upper bound on the proportion of defiers relative to compliers. To obtain bounds on PRTP, I adapt the setup and linear program in Mogstad et al. (2018) to accommodate defiers. PRTP can be written as linear combinations of conditional means of potential outcomes for subgroups defined by their treatment response to the instrument. Hence, with appropriate assumptions, the linear program can be retained. The baseline specification of the constraint set uses mean compatibility restrictions across conditional outcome distributions, but the method is amenable to additional restrictions researchers may wish to impose. This procedure yields an identified set that is an interval, and can be modified to incorporate covariates. Providing this tool for sensitivity analysis of PRTP is the main contribution of the paper.

As an application of the general theoretical results, I detail a particular type of PRTP — the treatment effect for compliers under a counterfactual policy environment, which I call the LATE*. In the Angrist and Evans (1998) study, the estimated effect of a third child on the mother's employment status from the IV regression is specific to the policy environment surrounding childcare in the dataset. Would we still have the same conclusion when the government gives a subsidy for childcare? What would the effect of a third child be for compliers in this counterfactual environment? These are questions answered by LATE*, which nests LATE as a special case (i.e., when there is no extrapolation). The LATE* is one way to think about external validity of a study's conclusions, which researchers are often interested in (e.g., Muralidharan et al. (2019); Ito et al. (2021)).[1]

[1]When calculating policy effects in counterfactual environments, parametric models of Brinch et al. (2017) and Kline and Walters (2019) are often used. However, these approaches are less useful when thinking of LATE* as a means to check external validity: since identification of heterogeneous treatment effects are often done without a parametric model, it seems desirable to avoid parametric models when evaluating the

In the counterfactual environment described, the treatment propensity for the entire population changes while the instrument values are the same. To obtain the LATE*, it suffices to characterize the mass of various treatment response groups in the original environment becoming compliers in the counterfactual environment. At a high level, (partial) identification of the LATE* is possible because the data places some restrictions on the means of potential outcomes, and objects of interest merely reweight these potential outcome means. If we are willing to put bounds on the fraction of people who respond to the instrument in the counterfactual environment relative to the original, meaningful bounds can be obtained. The same logic applies to other PRTP.

The procedure is implemented in the Angrist and Evans (1998) example. An instrument is used because it is believed that the OLS estimand is downward-biased: due to unobserved factors, women who are less likely to work are also those who are more likely to have a third kid. Hence, when the lower bound of the IV estimand reaches the OLS estimand, the bounds are no longer informative. I find that the bounds are informative only for small violations of monotonicity. Consider a counterfactual environment where a childcare subsidy is available. When the mass of defiers is more than 20% the mass of compliers, the lower bound for the LATE* falls from -0.103 under monotonicity to below the OLS benchmark of -0.134. Hence, the informativeness of the counterfactual estimates depends crucially on monotonicity.

This paper relates to several strands of literature. First, it is related to a literature on the failure of monotonicity in IV settings. Some papers that address violation of monotonicity include reinterpreting the estimand for the LATE (De Chaisemartin, 2017), using weaker monotonicity assumptions (Small et al., 2017; Heckman and Pinto, 2018; Kamat, 2018; Dahl

robustness of results.

et al., 2023) or alternative assumptions (Klein, 2010), and testing if it is indeed a concern (Kitagawa, 2015). Another common approach is the put bounds on the ATE or the LATE either using worst-case bounds or through some form of sensitivity analysis (Manski, 1989; Balke and Pearl, 1997; Horowitz and Manski, 2000; Noack, 2021; Kitagawa, 2021). There is also a literature that place bounds on further populations (e.g., compliers, defiers, never takers and always takers) (Richardson and Robins, 2010; Huber and Mellace, 2015; Huber et al., 2017; Ding and Lu, 2017). By targeting the PRTP, this paper not only covers bounds on these subpopulations, but also contributes bounds on extrapolated objects in counterfactual environments without monotonicity. Nonetheless, the approach in this paper does not have sharpness guarantees or closed-form solutions like in much of the existing literature.

Second, this paper is related to the literature on extrapolation and external validity in IV settings. In counterfactual environments, parametric models are often used (Brinch et al., 2017; Kline and Walters, 2019). Papers that target PRTP without a parametric model rely on a separable selection equation (Heckman and Vytlacil, 2005; Mogstad et al., 2018). The approach used in this paper neither uses a parametric model nor a separable selection equation — the latter cannot hold by construction when allowing for defiers. In light of the numerical equivalence between selection equations and the group primitives (Heckman and Vytlacil, 2005; Kline and Walters, 2019), this paper additionally contributes an example of how group primitives map to some nonseparable equation that permits extrapolation when monotonicity fails.

The rest of this paper discusses the proposed method and its applications. Section 2 explains the general framework in forming bounds for PRTP; Section 3 applies the framework to LATE*. Section 4 applies the procedure to the Angrist and Evans (1998) example. Section

5 concludes.

2 Framework for Identification without Monotonicity

2.1 Setting

We observe random variables (T, Z, Y), denoting treatment, instrument, and outcome respectively. We are interested in the effect of the endogenous T on Y in a counterfactual environment. Outcome Y can be discrete or continuous; instrument $Z \in \mathcal{Z} = \{0, 1, \cdots, k-1\}$ takes one of $k < \infty$ discrete values, and treatment $T \in \{0, 1\}$ is binary. Although the setup can be adapted to multivalued T, I focus on the binary case for simplicity. Let $T(z)$ denote the potential treatment when given instrument z, and let $Y(t)$ denote the potential outcome when given treatment t, which assumes that Y is not affected by Z directly. Let $T^*(z^*)$ denote the potential treatment when given instrument $z^* \in \mathcal{Z}^*$ in the counterfactual environment, where $\mathcal{Z}^*$ is the set of values that the instrument can take in the counterfactual environment. Without loss of generality, the instrument values are ordered such that $\Pr(T(z) = 1)$ is increasing in z.[2] Then, the observed T and Y are $Y = Y(T)$ and $T = T(Z)$.

Treatment response groups $g \in \mathcal{G}$ are characterized by the vector of potential treatments, i.e., $((T(z))_{z\in\mathcal{Z}}, (T^*(z^*))_{z^*\in\mathcal{Z}^*})$. $\mathcal{G}$ is the set of all possible combinations of $((T(z))_{z\in\mathcal{Z}}, (T^*(z^*))_{z^*\in\mathcal{Z}^*})$: with a binary treatment, k instrument values, and $\mathcal{Z}^* = \mathcal{Z}$, we have $|\mathcal{G}| = 2^{2k}$. Without extrapolation, the counterfactual environment is the original environment. Then, $\mathcal{Z}^* = \mathcal{Z}$

[2]There is a bijection from any set $\mathcal{Z}'$ with k discrete values to $\mathcal{Z}$ such that for any $z, z' \in \mathcal{Z}$ such that $z > z'$, $\Pr(T(z) = 1) \geq \Pr(T(z') = 1)$. Hence, beyond having k discrete values for the instrument, assumptions on $\mathcal{Z} \subset \mathbb{N}$ and the ordering of the values are without loss of generality.

and $T(z) = T^*(z)$, $\forall z \in \mathcal{Z}$. In general, the mass of each group in the population is

$$q_g := \Pr(g).$$

Let q denote a vector that stacks all q_g values that are nonzero. In applications, there may be groups with $q_g = 0$. Hence, the dimension of q, d_q, is defined as the number of groups with nonzero mass, so $d_q \leq |\mathcal{G}|$.

For example, consider an environment with binary treatment, $k = 2$ instrument values and $\mathcal{Z} = \mathcal{Z}^*$. Using terminology in the literature (e.g, Angrist et al. (1996)), the 4 response groups in the original environment are always-takers (A) with $T(0) = T(1) = 1$, compliers (C) with $T(0) = 0$ and $T(1) = 1$, defiers (D) with $T(0) = 1$ and $T(1) = 0$ and never-takers (N) with $T(0) = T(1) = 0$. Then, $|\{((T(z))_{z\in\mathcal{Z}}, (T^*(z^*))_{z^*\in\mathcal{Z}^*})\}| = 2^{2\times 2} = 16$. If we are not interested in the extrapolated environment, then $T(z) = T^*(z)$, so we only have $d_q = 4$ groups.

Define the conditional mean for each group as follows:

$$\mu_{gt} := E[Y(t)|g].$$

Similarly, let μ be the vector that stacks the μ_{gt} values, and let $d_\mu := dim(\mu)$ denote the dimension of μ. It is implicitly assumed that these μ_{gt} objects are well-defined. When treatment is binary, $d_\mu = 2d_q$.

Following Huber et al. (2017), it suffices to have mean independence of the potential outcomes across groups instead of full independence:

Assumption 1. *$E[Y(t)|g,z] = E[Y(t)|g]$ and $\Pr(g|z) = \Pr(g)$ for all g, z.*

These groups are the primitives of the setup. Random assignment of the instrument Z satisfies Assumption 1. In addition to Assumption 1, following Angrist and Imbens (1994), many papers also assume monotonicity, the assumption that the instrument weakly affects treatment in the same direction for all individuals.

Assumption 2. *For all $z_1, z_2 \in \mathcal{Z}$ either $\Pr(T(z_1) \geq T(z_2)) = 1$ or $\Pr(T(z_1) \leq T(z_2)) = 1$. For $z_1^*, z_2^* \in \mathcal{Z}^*$, either $\Pr(T^*(z_1^*) \geq T^*(z_2^*)) = 1$ or $\Pr(T^*(z_1^*) \leq T^*(z_2^*)) = 1$.*

Assumption 2 implies there are particular groups g with $q_g = 0$, which, in the environment without extrapolation, reduces number of treatment response types with nonzero mass from 2^k to $k+1$. This paper conducts sensitivity analysis for the failure of this assumption, so it relaxes Assumption 2. Since this assumption is a statement about the potential treatment response, sensitivity analysis involves careful consideration of the masses q_g of various groups.

The object of interest is the PRTP, defined as any estimand that can be written as:

$$\beta = \sum_{g,t} c_{gt}(q)\mu_{gt} = c(q)'\mu. \tag{1}$$

where $c_{gt}(q)$'s denote the weights on each of the μ_{gt}'s, and these coefficients can depend on q. The equality requires the object of interest to be linear in μ. $c(q)$ is the coefficient vector, with $c : [0,1]^{d_q} \to \mathbb{R}^{d_\mu}$ transforming the vector of proportions into weights on the conditional expectations. Once q is known, $c(q)$ is known. Objects of interest like the LATE and the average treatment effect (ATE) can be written in this form. For example, the ATE uses $c(q) = q \otimes (1,-1)'$, the average treatment effect on the treated (ATT) uses

$c(q) = (q_A, -q_A, q_C, -q_C, q_D, -q_D, 0, 0)'/(q_A+q_C+q_D)$, and the LATE* is explained in Section 3. This β can be viewed as a discretized version of the PRTP defined in Mogstad et al. (2018).

The relationship between β and the target object in Mogstad et al. (2018) warrants further discussion. Mogstad et al. (2018) assumed monotonicity, so treatment can be written as $T = 1[\tilde{\nu}(Z) \geq u]$, for unobserved $u \sim U[0,1]$. The primitives of their model are marginal treatment responses $E[Y(t) \mid u]$, and their target parameter integrates a weighted average of $E[Y(t) \mid u]$ over u. In the monotonic setting, u has a natural interpretation as a treatment propensity, where high values of u correspond to N, middle values to C, and and low values to A for a binary instrument. However, when monotonicity fails, the treatment equation becomes nonseparable with $T = 1[\nu(Z,u) \geq 0]$. Then, the interpretation of u is unclear unless a researcher has a particular $\nu(Z,u)$ in mind. Nonetheless, the groups remain well-defined in general. The unobserved u is meaningful in the target object insofar as it defines the groups that we are interested in. Hence, this paper uses the unobserved groups g to characterize conditional means, and characterizes the target object in terms of $E[Y(t) \mid g]$ instead of $E[Y(t) \mid u]$. The relationship between this group characterization and a nonseparable selection equation will be further clarified in Section 3.2 through an example.

2.2 Constraints on μ and q

The method places bounds on objects of interest by using the researcher's input for a sensitivity parameter. To explain this method, I first explain the constraints on μ implied by Assumption 1 in Section 2.2.1, where it is assumed that the vector q is known. Then, Section 2.2.2 shows how a single sensitivity parameter that affects q captures the extent that

monotonicity is violated.

2.2.1 Constraint Set for μ

$\mathcal{M}(q)$ denotes the set of μ that satisfies defined equality and inequality constraints. These constraints may depend on q, and may include ex ante restrictions and features of the data. The researcher can specify what these constraints are, but I require these constraints to be linear in μ and the set $\mathcal{M}(q)$ to be convex.

One example of $\mathcal{M}(q)$ is a set of mean compatibility constraints implied by Assumption 1. In $Y|T = t, Z = z$, the mean of the various structural μ_{gt} such that $T(z) = t$, weighted by their proportions, is equal to the reduced-form mean $E[Y|T = t, Z = z]$. Hence, where $p_{tz} := \Pr(T = t|Z = z)$, for all z, t, these constraints take the form:

$$\sum_{g:T(z)=t} q_g \mu_{gt} = p_{tz} E[Y|T = t, Z = z]. \tag{2}$$

Observe that (2) is a function of the q vector, so q parameterizes the constraint set $\mathcal{M}(q)$. Without ex ante restrictions, the set of μ that satisfies mean compatibility is in Equation (3). This set is denoted $\mathcal{M}_m(q)$ to avoid confusion with the general constraint set $\mathcal{M}(q)$:

$$\mathcal{M}_m(q) := \left\{ \mu \in \mathbb{R}^{d_\mu} : \sum_{g:T(z)=t} q_g \mu_{gt} = p_{tz} E[Y|T = t, Z = z] \qquad \forall z \in \mathcal{Z}, t \in \{0, 1\} \right\}. \tag{3}$$

The constraints in set $\mathcal{M}_m(q)$ do not exploit all distributional information, but nonetheless make the problem tractable, so $\mathcal{M}_m(q)$ can be used as a default. With binary outcomes, $\mu_{gt} \in [0, 1]$ should be used as a constraint. Without binary outcomes, we may consider

additional constraints implied by Assumption 1, such as the trimming bounds of Lee (2009). Additional restrictions that the researcher may impose include selection into treatment (e.g., Roy (1951)).

2.2.2 Sensitivity Parameter

To form a sensitivity parameter for violation of Assumption 2, I first define compliers and defiers. For $z > z'$, define sets of defiers and compliers respectively as:

$$S^d_{(z,z')} := \{g : T(z) < T(z')\}, \text{ and}$$
$$S^c_{(z,z')} := \{g : T(z) > T(z')\}.$$

Since $\Pr(T(z) = 1)$ is increasing in z, $\Pr(g \in S^d_{(z,z')}) \leq \Pr(g \in S^c_{(z,z')})$. Assumption 2 is equivalent to having no defiers, so the sensitivity parameter should control the proportion of defiers, which then affects the q vector. Hence, the sensitivity parameter λ imposes the restriction that, for all pairs (z, z'),

$$\sum_{g \in S^d_{(z,z')}} q_g \leq \lambda \sum_{g' \in S^c_{(z,z')}} q_{g'}. \tag{4}$$

I refer to the inequality restriction (4) imposed by λ as a "sensitivity restriction".[3] I also place an analogous sensitivity restriction on the counterfactual environment with $T^*(.)$. In particular, for $S^{d*}_{(z,z')} := \{g : T^*(z) < T^*(z')\}$ and $S^{c*}_{(z,z')} := \{g : T^*(z) > T^*(z')\}$, the sensitivity restriction is $\sum_{g \in S^{d*}_{(z,z')}} q_g \leq \lambda \sum_{g' \in S^{c*}_{(z,z')}} q_{g'}$. It is possible to have a different

[3]This sensitivity parameter was earlier proposed in Ding and Lu (2017) for the case with a binary instrument and binary treatment when targeting subpopulations in the sample.

sensitivity parameter for every pair (z, z') in nonbinary settings — this does not change the method, but increases the number of sensitivity parameters. To keep the exposition simple, I work with a single sensitivity parameter λ. When $\lambda = 0$, there is no pair of instrument values for which there are defiers.

With $\mathcal{Q}(\lambda)$ denoting the general constraint set, the proportion vector satisfies $q \in \mathcal{Q}(\lambda)$. As in the treatment of $\mathcal{M}(q)$, the researcher can specify additional restrictions, but I propose the minimal set of restrictions. Namely, the proportions chosen must be compatible with the observed p_{tz}. Assumption 1 implies $\forall t, z$,

$$\sum_{g:T(z)=t} q_g = p_{tz}. \tag{5}$$

The set $\mathcal{Q}(\lambda)$ may be empty for some choices of λ. Due to Proposition 1 of Noack (2021), there are bounds imposed on q_D by the data, so if λ is too small, the set will be empty. Notably, Noack (2021) assumes full independence rather than mean independence that is assumed in this paper, so if we assume full independence, tighter bounds on q_D can be obtained from her method.[4] The sensitivity restriction thus describes monotonicity violations that are not detectable by the data. Even if $q_D = 0$ is rejected by the data, the existing tests can construct a confidence interval for q_D that can feature as a restriction in $\mathcal{Q}(\lambda)$, which can still be used to bound the PRTP.

[4]Bounds on q_D are obtained from implications on the outcome distribution. With full independence, the entire outcome distribution can be used to obtain the bounds, but with mean independence, we can only use the conditional means of the outcome distribution.

The minimal constraint set satisfies (4) and (5), so it takes the form $\mathcal{Q}(\lambda) = \mathcal{Q}_m(\lambda)$:

$$\mathcal{Q}_m(\lambda) := \left\{ q \in [0,1]^{d_q} : \sum_{g \in S^d_{(z',z'')}} q_g \leq \lambda \sum_{g' \in S^c_{(z',z'')}} q_{g'}, \sum_{g \in S^{d*}_{(z',z'')}} q_g \leq \lambda \sum_{g' \in S^{c*}_{(z',z'')}} q_{g'}, \sum_{g:T(z)=t} q_g = p_{tz}, \forall (z',z''), t, z \right\}. \tag{6}$$

Observe that $\sum_g q_g = 1$ is implied by the condition that $\sum_{g:T(z)=t} q_g = p_{tz}, \forall t, z$. Following Mogstad et al. (2018), define our identified set for PRTP:

$$\mathcal{B}_\lambda = \{b \in \mathbb{R} : b = c(q)'\mu \text{ for some } \mu \in \mathcal{M}(q), q \in \mathcal{Q}(\lambda)\}. \tag{7}$$

More precisely, $\mathcal{B}_\lambda$ is the set identified by constraints in $\mathcal{M}$ and $\mathcal{Q}$.

Remark 1. Due to the generality of the framework, several extensions can be accommodated. First, we can extend the analysis to multivalued treatments. With $|\mathcal{T}|$ treatment values, we can analogously define $\mathcal{G}$ so that $|\mathcal{G}| = |\mathcal{T}|^{2|\mathcal{Z}|}$. The object of interest remains as a linear combination of group-specific average potential outcomes. Second, we can extend the analysis to multiple binary instruments. With b binary variables, $|\mathcal{Z}| = 2$, and we have $|\mathcal{G}| = 2^{2|\mathcal{Z}|^b}$ groups. Then, we may conduct sensitivity analysis with respect to partial monotonicity (Mogstad et al., 2021) or limited monotonicity (van't Hoff et al., 2023) by imposing inequality (4) only with respect to their affected groups rather than all pairs.

2.3 Theoretical Properties

This subsection presents the main identification result of the paper, that the identified set is an interval. The method for finding bounds on PRTP solves an optimization problem in light of the constraints on μ and q from the previous subsection. Since obtaining the upper

and lower bounds of the interval involves optimizing over μ and q, it is helpful to break the optimization problem into an inner problem that optimizes over μ given q and an outer problem that optimizes over q. Write the inner optimization problem as:

$$\underline{R}(q) := \min_{\mu \in \mathcal{M}(q)} c(q)'\mu, \text{ and} \qquad \overline{R}(q) := \max_{\mu \in \mathcal{M}(q)} c(q)'\mu. \tag{8}$$

These inner optimization problems are linear programs by assumption, given q. This rewriting is convenient because linear programs are computationally cheap. The linearity of the general program conditional on q is similar to the generic framework presented in Mogstad et al. (2018), which did not allow for monotonicity violations. Assumption 3 below provides sufficient conditions for the identified set to be an interval.

Assumption 3. *For a given $\lambda \in [0, 1)$, the following hold:*

(a) For all $g \in \mathcal{G}$, if $q_g > 0$, then μ_{gt} is well-defined and finite $\forall t \in \{0, 1\}$.

(b) Restrictions in $\mathcal{M}(q)$ can be written as a system of linear inequalities in μ such that $\mathcal{M}(q) = \{\mu : A(q)\mu \leq b(q)\}$ is continuous in $(A(q), b(q))$, and $\mathcal{M}(q)$ is convex.[5] *Hyperparameters $A(q)$ and $b(q)$ of the linear program are continuous in q.*

(c) $c(q)$ is continuous in q.

(d) $\mathcal{Q}(\lambda)$ is a nonempty convex set.

Theorem 1. *(Identified Set). Suppose Assumption 1 and 3 hold for some λ. Then, either $\mathcal{M}(q)$ is empty for all $q \in \mathcal{Q}(\lambda)$ and hence $\mathcal{B}_\lambda$ is empty, or the closure of $\mathcal{B}_\lambda$ is equal to the*

[5]The set $\mathcal{M}$ is continuous in (A, b) if it is lower and upper hemi-continuous in (A, b). In general, $\mathcal{M}$ is not lower hemi-continuous. For a counterexample, consider $t = (A, b)$ and $K(t) = \{x : Ax \leq b, x \geq 0\}$. The sequence $t_\nu = (A = \nu^{-1}, b = \nu^{-1})$ converges to $t^* = (0, 0)$. Observe that $K(t_\nu) = [0, 1]$. The point $2 \in K(t^*)$ cannot be reached by any sequence $\{x_\nu, \nu = 1, \cdots\}$ with $x_\nu \in [0, 1]$.

interval $[\underline{\beta}_\lambda, \overline{\beta}_\lambda]$*, where*

$$\underline{\beta}_\lambda = \min_{q \in \mathcal{Q}(\lambda)} \underline{R}(q), \text{ and} \qquad \overline{\beta}_\lambda = \max_{q \in \mathcal{Q}(\lambda)} \overline{R}(q). \tag{9}$$

The theorem claims that the identified set is an interval, so every point in the interval is achievable by some $\mu \in \mathcal{M}(q), q \in \mathcal{Q}(\lambda)$. This property is not immediately obvious when optimizing over (q, μ): when we optimize over q, the objective function is potentially nonconvex, since $c(q)$ is nonlinear in q. Continuity of functions and convexity of sets are hence required for the result. Proof details are in Appendix D. Notably, even if Assumption 3 fails, (9) still yields valid bounds, albeit conservative.

Generally, when using $\mathcal{M}_m(q)$ and $\mathcal{Q}_m(\lambda)$, the bounds are not sharp in that the (q, μ) pair that solves the problem need not be compatible with the data. The non-sharpness arises from two problems. The first problem is that not all $q \in \mathcal{Q}_m(\lambda)$ is compatible: for instance, it is known in the literature that there are tests for monotonicity (e.g., Richardson and Robins (2010); Kitagawa (2015); Huber et al. (2017); Noack (2021)), so $q_D = 0$ need not be compatible with the data. The second problem occurs because we have only used information on the means across distributions, and we have not yet exploited all distributional information. If outcomes are discrete, sharp bounds can be obtained by parameterizing the entire joint distribution of $(Y(0), Y(1), g)$, which is the approach taken by Balke and Pearl (1997). If the outcome is binary and all $q \in \mathcal{Q}_m$ are compatible, then we have sharp bounds, since $\mathcal{M}_m(q) \cup [0,1]^{d_\mu}$ contains all distributional information.

The sensitivity parameter also has a nice feature stated in Theorem 2, which reduces the number of inequality constraints.

Theorem 2. *Let* $z_0, z, z' \in \mathcal{Z}$. *If* $\sum_{g \in S^d_{(z_0,z_0+1)}} q_g \leq \lambda \sum_{g' \in S^c_{(z_0,z_0+1)}} q_{g'}$ *for all* $z_0 \in \mathcal{Z} \backslash \{k-1\}$, *then* $\sum_{g \in S^d_{(z,z')}} q_g \leq \lambda \sum_{g' \in S^c_{(z,z')}} q_{g'}$ *for any instrument value pair* (z, z').

This theorem implies that we do not need to consider all instrument pairs — it suffices to consider adjacent instrument pairs. For intuition, when there are no defiers at both the $(z, z+1)$ and $(z+1, z+2)$ margins, it must be that there are no defiers at the $(z, z+2)$ margin, because the defiers at the $(z, z+2)$ margin must switch at either margin. In light of this result, we only have to check $k-1$ instead of $\binom{k}{2}$ constraints.

Remark 2. The property in Theorem 2 is a feature of defining the sensitivity parameter in this way. If we had instead defined the sensitivity parameter as an upper bound on the proportion of defiers as done in Noack (2021), we no longer have this property. To see this, suppose we have three discrete instrument values $\{0, 1, 2\}$. Sensitivity parameter η is such that $q_{(1,0,0)} + q_{(1,0,1)} \leq \eta^*$ and $q_{(0,1,0)} + q_{(1,1,0)} \leq \eta^*$ at the $(0, 1)$ and $(1, 2)$ margin of the instrument respectively. In the worst case, we will have $q_{(1,0,0)} = \eta^*$ and $q_{(1,1,0)} = \eta^*$. Then, at the $(0, 2)$ margin, $q_{(1,0,0)} + q_{(1,1,0)} = 2\eta^*$, which is not bounded above by η^*.

Remark 3. Constructing the sensitivity restriction as $q_D/q_C \leq \lambda$ makes λ interpretable across applications. Suppose we have $q_D = 0.01$ — if $q_C = 0.5$, then the violation of monotonicity is relatively small; but if $q_C = 0.02$, the violation would be rather large. λ reflects the difference, despite having the same q_D. Nonetheless, if making an assumption on q_D directly instead of q_D/q_C is more interpretable in a particular application, a constraint of the form $q_D \leq \lambda_D$ can be used in $\mathcal{Q}$ without loss.

2.4 Implementation

To implement the procedure proposed in the paper, we can simply use the sample analog. We observe data (Y_i, T_i, Z_i) for $i = 1, \cdots, n$. An implementable algorithm is:

1. Estimate probability objects p_{tz} by $\hat{p}_{tz} = \frac{\sum_{i=1}^{n} 1[T_i=t, Z_i=z]}{\sum_{i=1}^{n} 1[Z_i=z]}$. Use sample analog $\hat{E}[Y|T = t, Z = z] = \frac{1}{n_{tz}} \sum_{i:T_i=t,Z_i=z} Y_i$ for $E[Y|T = t, Z = z]$, $n_{tz} := \sum_{i=1}^{n} 1[T_i = t, Z_i = z]$.

2. For given $q \in \mathcal{Q}(\lambda)$,

   (a) Plug in $\hat{E}[Y|T = t, Z = z]$ and q into (2).

   (b) Set up the objective function and solve the linear program in (8). Output the value of the objective function $R(q)$.

3. For given λ, optimize output of Step 2 over q in the outer loop as in (9) using the sample analog.

Denote the estimators obtained from the sample $(\hat{\underline{\beta}}_\lambda, \hat{\overline{\beta}}_\lambda)$ for the lower and upper bounds respectively for the problem in (9). These estimators can be shown to be consistent by applying the Glivenko-Cantelli theorem to iid data, for instance, and applying the continuous mapping theorem after proving continuity in the program. Inference can be done by the projection method, and details are in Appendix B. In empirical applications, the instrument may be valid only conditional on covariates, so Appendix C extends the procedure to incorporate covariates.

While the above procedure suffices for the numerical results in this paper, as Section 3.1 shows how Step 3 can be reduced to a one-dimensional optimization problem, Step 3 may be unwieldy in general as the dimension of q can be large. To address this concern,

Steps 2 and 3 can be combined into a bilinear program so we jointly optimize over (q, μ). Most objective functions considered can be written as linear fractionals of q, i.e., $c(q)'\mu = q'A\mu/d'q$, for some conformable matrix A and vector d, with $d'q > 0$ and linear constraints on (μ, q), say $Bq \leq b, C\mu \leq c$. Applying the Charnes-Cooper transformation by defining $t := 1/d'q, r := q/t$, the program is equivalent to optimizing $r'A\mu$ over (r, t, μ) such that $d'r = 1, Br \leq Bt, C\mu \leq c$. Then, standard algorithms for bilinear programs (Dutz et al., 2021; Shea, 2022) can be applied.

3 Identification of LATE*

The method in Section 2 is general, and allows partial identification of any combination of treatment response groups. Nonetheless, researchers often care about compliers. Hence, this section discusses and interprets LATE*, which is defined as the TE on compliers in counterfactual policy environments.[6]

For ease of exposition, I consider a binary instrument using the (A,C,D,N) notation as discussed in Section 2. The response groups in the counterfactual environment $\{A^*, C^*, D^*, N^*\}$ can be defined on $T^*(z)$ analogously. Using $G \in \{A, C, D, N\}$ and $G^* \in \{A^*, C^*, D^*, N^*\} =: \mathcal{G}_{cf}$ to denote response in the original and counterfactual environments respectively, $q_{GG^*} = \Pr(G, G^*)$ denotes the proportion who were G in the original environment and G^* in the new environment. Conditional probabilities are denoted $q_{G^*|G} := \Pr(G^*|G) = q_{GG^*}/(\sum_{H^* \in \mathcal{G}_{cf}} q_{GH^*})$.

[6]LATE in Angrist and Imbens (1994) is defined under monotonicity as the TE for the subpopulation who respond (i.e., change their treatment status) to the instrument, which is equivalent to the TE on compliers (TEC). In the presence of defiers, the TE on the marginal population (TEM) and TEC are no longer equivalent. Since LATE was defined on a subpopulation with a particular treatment response status, it is sensible to define it as the TEC when there are defiers present. Hence, I define LATE* in the rest of this paper as the TEC in the counterfactual environment. We could also instead calculate TEM*, but I focus on LATE* to be concrete.

Using the definition that the LATE* is the TE for the counterfactual compliers, and $\mu_{GG^*t} := E[Y(t) \mid G, G^*]$,

$$LATE^* = \frac{\sum_G q_{GC^*} \left(\mu_{GC^*1} - \mu_{GC^*0}\right)}{\sum_G q_{GC^*}}.$$

The LATE* is useful for several reasons. First, the counterfactual environment could differ in place or time. Since the Angrist and Evans (1998) used US data, if we believe that the Canadian population is similar to the US, and its only difference is that it has better childcare, then the LATE* is what the LATE in Canada would be. For extrapolation over time, the study used 1990 data, but the current policy environment has changed since then, so the LATE* tells us what the LATE is now. Second, the LATE* is as useful to the policy maker as the LATE. If LATE features in the policy function, then so must the LATE* once the policy is implemented because the environment would have changed. For example, if the policy maker wishes to give a $2000 subsidy in two tranches, once the first $1000 has been rolled out, the "LATE" would have changed, and we cannot expect the second $1000 to yield the same effect. This occurs because people no longer stick to their original groups. Such a setting is relevant when policy makers only have old studies or surveys available to inform current policy implementation. Third, the LATE* is useful in calibration. Parameter values in a model may be calibrated by using estimates from other studies. Then, the approach in this paper gives an explicit way of thinking about how the study at hand differs from the original study that the parameter value was calibrated from, and consequently the appropriate bounds on these values. Fourth, even though the LATE* is not point-identified, it is useful in policy choice when the social planner has a min/max objective function. The

policy-maker can then choose policy rules by using the worst-case bounds obtained. Finally, since LATE* identifies the TE for a subpopulation, it is useful for assessing the robustness of conclusions on TE.

Since the object of interest is the LATE*, when considering policy changes that do not change the potential outcomes and unobservables, it suffices to characterize the proportions of original groups becoming C^* in the counterfactual environment. Hence, the counterfactual policy environment is characterized by the four extrapolation parameters $q_{C^*|G}$, denoting the proportion of the original groups switching into our group C^* of interest. Using this setup, LATE and ATE are special cases of the LATE*: LATE is the LATE* without extrapolation, and ATE is the LATE* when everyone switches into C^*.[7] Nonetheless, in many counterfactual policies of interest, such as increasing the instrument strength or increasing treatment propensity, only $q_{C^*|N}$ and $q_{C^*|C}$ matter, as these counterfactual environments imply $q_{C^*|A} = 0$ and $q_{C^*|D} = 0$.[8] I provide two examples.

Example 1. (Changing Instrument Value). In Duflo and Saez (2003), people were randomly given a letter that gave them \$20 if they attended the meeting, but they could have been given \$30 instead. This counterfactual corresponds to changing the instrument value (Z), say from 1 to 2. Researchers were interested in the effect of the meeting (T) on taking up a pension plan (Y). Here, $T^*(0) = T(0)$. The counterfactual compliers are those with

[7] Observe that there is no gain in using sensitivity analysis for ATE, as observed by Kitagawa (2021), because the bounds are the widest when the proportion of defiers is the smallest.

[8] Recent literature that deal with counterfactual environments as in Carneiro et al. (2010), Carneiro et al. (2011) and Mogstad et al. (2018) consider three counterfactual policies. These policy counterfactuals are in the class considered by Heckman and Vytlacil (2005), which involves policies that do not affect the marginal treatment response of T on Y. Their policy counterfactuals include (i) Additive α change in propensity score with the same instrument value (ii) Proportional $1+\alpha$ change in propensity score with same instrument (iii) Additive α shift of the jth component of Z, so $Z^* = Z + \alpha e_j$ and $p^*(x,z) = p(x,z)$. Changing the value of the instrument corresponds to policy type (iii) and monotonically changing the probability of being treated corresponds to (i) and (ii), so I group the first two together.

$T^*(2) = 1, T^*(0) = 0$. Groups with $T(0) = 0$ are the original compliers and never-takers, so only C and N can become the counterfactual C^* group.

Example 2. (Changing Treatment Propensity). A policy may subsidize childcare in the Angrist and Evans (1998) context: regardless of a couple's gender preference, the probability of having a third child increases, i.e., $T^*(z) \geq T(z)$. Researchers were interested in the effect of a third child (T) on labor force participation (Y), and T is instrumented by first two kids having the same sex (Z). The counterfactual compliers are those with $T^*(1) = 1, T^*(0) = 0$. Since the policy weakly incentivizes treatment, individuals in C^* must have had $T(0) = 0$ in the original environment, which can only include the original C and N groups.

While we have not seen people respond to the counterfactual incentives, we have seen people respond to other incentives. If we put bounds on the fraction of people who respond to the counterfactual environment but not the original, we can make progress. To bound such fractions, some economic reasoning is required for how the environment maps to the fraction: in Duflo and Saez (2003), we require a mapping from the financial incentive to fraction of people changing their behavior; in Angrist and Evans (1998), we require a mapping from the subsidy amount to the fraction.

Remark 4. (Relation to extrapolation in Marginal Treatment Effects framework). Without monotonicity, $T = 1[\nu(Z, u) \geq 0]$ for some $\nu(.)$. The counterfactual policies map to (1) Change the value of the instrument so $T^* = 1[\nu(Z^*, u) \geq 0]$; and (2) Change the threshold for everyone so $T^* = 1[\nu(Z, u) \geq -\alpha]$, increasing treatment propensity. I defer details to Section 3.2.

The objective is hence $LATE^* = E[Y(1) - Y(0)|g \in \{CC^*, NC^*\}]$, which can be written

as a linear function of μ_{GG^*}. The sensitivity restrictions may be constructed analogously, where λ restricts the proportion in both the original and counterfactual environments. For instance, when increasing the treatment propensity, the defier restrictions are:

$$\begin{aligned} q_{DD^*} + q_{DA^*} &\leq \lambda(q_{CC^*} + q_{CA^*}), \text{ and} \\ q_{DD^*} + q_{ND^*} &\leq \lambda(q_{CC^*} + q_{NC^*}). \end{aligned} \tag{10}$$

This problem can then be written in the form of the linear program in Section 2, which uses an inner linear program $R(q)$ that is cheap, and an outer problem that optimizes over q. The implementation for the threshold crossing counterfactual is explained in the next subsection; the implementation for changing the instrument value is analogous, and is explained in Appendix A.1.

3.1 Treatment Propensity Implementation

To show how the framework of Section 2 applies, it suffices to specify the following: (i) the objective function (ii) what the groups g are (iii) linear restrictions for μ in the inner problem (iv) the constraint set for q in the outer optimization problem. Item (i) is LATE*, so the rest of this subsection explains the other items.

In our policy counterfactual, A will still be A^*. C can remain C^*, or they can become A^* when the policy is strong enough to shift their $Z = 0$ treatment to $T = 1$. The same argument applies to D. Finally, consider the N group. If the policy is weak, they would remain N^*. The policy may affect the outcome for only either $Z = 0$ or $Z = 1$, which changes their response behavior to D^* or C^*. The policy may also be strong enough to get

the N group to $T = 1$ regardless of the instrument. Then, N can change their behavior to N^*, C^*, D^*, or A^*.

Although there are 9 response types, if the researcher does not wish to impose restrictions on q_{GG^*} that affects the sensitivity inequalities, we can essentially deal with 6 response groups $(A, CA^*, CC^*, D, NC^*, NC'^*)$, where NC'^* denotes the set of groups that switch from N to anything but C^* in the counterfactual policy environment, and D is the cell that collects all types who were defiers in the original environment. To be precise, define the following objects when there are 9 treatment response groups:

$$LATE^* = \frac{q_{CC^*}(\mu_{CC^*1} - \mu_{CC^*0}) + q_{NC^*}(\mu_{NC^*1} - \mu_{NC^*0})}{q_{CC^*} + q_{NC^*}},$$

$$q = (q_A, q_{CA^*}, q_{CC^*}, q_{DA^*}, q_{DA}, q_{NA^*}, q_{NC^*}, q_{ND^*}, q_{NN^*})',$$

$$\overline{R}^{TC}(q) := \max_{\mu \in \mathcal{M}_m^{TC}(q)} LATE^*, \text{ and}$$

$$\mathcal{M}_m^{TC}(q) := \left\{ \mu \in [0,1]^{18} : \sum_{g:T(z)=t} q_g \mu_{gt} = p_{tz} E[Y|T=t, Z=z] \qquad \forall z \in \{0,1\}, t \in \{0,1\} \right\}.$$

When there are 6 treatment response groups,

$$\tilde{R}(\tilde{q}) := \max_{\tilde{\mu} \in \tilde{\mathcal{M}}_m(\tilde{q})} LATE^*,$$

$$\tilde{q} := (q_A, q_{CC^*}, q_{CC'^*}, q_D, q_{NC^*}, q_{NC'^*})',$$

$$\tilde{\mu} := (\mu_{A1}, \mu_{A0}, \mu_{CC^*1}, \mu_{CC^*0}, \mu_{CC'^*1}, \mu_{CC'^*0}, \mu_{D1}, \mu_{D0}, \mu_{NC^*1}, \mu_{NC^*0}, \mu_{NC'^*1}, \mu_{NC'^*0})', \text{ and}$$

$$\tilde{\mathcal{M}}_m(\tilde{q}) = \left\{ \mu \in [0,1]^{12} : \sum_{g:T(z)=t} \tilde{q}_g \mu_{gt} = p_{tz} E[Y|T=t, Z=z] \qquad \forall z \in \{0,1\}, t \in \{0,1\} \right\}.$$

Proposition 1. *Consider* $q = (q_A, q_{CA^*}, q_{CC^*}, q_{DA^*}, q_{DA}, q_{NA^*}, q_{NC^*}, q_{ND^*}, q_{NN^*})'$. *If* $q_{CC'^*} =$

q_{CA^*}, $q_D = q_{DA^*} + q_{DA}$, *and* $q_{NC'^*} = q_{NA^*} + q_{ND^*} + q_{NN^*}$, *then* $\overline{R}^{TC}(q) = \tilde{R}(\tilde{q})$.

Proposition 1 tells us that the bound for our object of interest does not change when we solve the 6 response group problem instead of the 9 response group problem, as long as we use the minimal constraint set $\mathcal{M}_m^{TC}(q)$ for μ. The proof proceeds by using the observation that $LATE^*$ is a function of $(q_{CC^*}, q_{NC^*}, \mu_{CC^*1}, \mu_{CC^*0}, \mu_{NC^*1}, \mu_{NC^*0})$. Then, it remains to argue that both optimization problems place the same restrictions on those parameters.

Finally, we can consider the constraint set on q. There are two restrictions in the form of (5); two restrictions based on the chosen $q_{C^*|C}, q_{C^*|N}$ as $q_{C^*|G} = q_{GC^*}/(\sum_{H^* \in \mathcal{G}_{cf}} q_{GH^*})$; and probabilities must sum to one. In addition to the five linear equality restrictions, sensitivity restrictions (10) must be satisfied. By using the linear equality restrictions, we only need to optimize over a single parameter in the outer problem with q. To see this result, there are 5 linearly independent restrictions involving q, and we can also write $q_D = q_D$ as a trivial relationship. Hence, using a system of 6 equations and 6 unknowns in q, for a given environment, once we know q_D, we know the rest of the q vector. Details are in Appendix A.2. Consequently, bounds on LATE* can be obtained by solving a cheap linear program in μ in the inner loop with a one-dimensional optimization over q_D in the outer loop.

The next subsection gives examples of selection equations that justify treatment response groups. It can be skipped without loss of continuity.

3.2 Example of Selection Equations

In counterfactual environments, we could augment $\nu(Z, u)$ in Mogstad et al. (2018) to account for defiers, but it is difficult to do so without more structure on how defiers feature in the

selection equation. Since characterizing the counterfactual environment based on groups is new, it is instructive to consider how this approach relates to selection equations. In particular, I show how selection equations under monotonicity with a binary instrument maps to groups in the counterfactual environment. I then use that intuition to explain what happens with a nonseparable selection equation. To begin, I consider the case without defiers, so the selection equation is given by $T = 1[\tilde{\nu}(Z) \geq u]$, where $u \sim U[0,1]$. Since Z is binary, $\tilde{\nu}(Z)$ can only take two values, and the environment is illustrated in Figure 1.

In Figure 1, panel (i) illustrates the original environment, so low values of u are always-takers, those with middle values of u are compliers and those with high values of u are never-takers. Since u is uniformly distributed, $q_A = \tilde{\nu}(0), q_C = \tilde{\nu}(1) - \tilde{\nu}(0), q_N = 1 - \tilde{\nu}(1)$. In panel (ii), we have a counterfactual environment where the threshold is shifted by α such that $T^* = 1[\tilde{\nu}(Z) + \alpha \geq u]$. Consequently, the A^*, C^*, N^* groups are defined by the new cutoffs at $\tilde{\nu}(0) + \alpha$ and $\tilde{\nu}(1) + \alpha$. Panel (iii) combines the groups from panels (i) and (ii): for instance, the CA^* group are observations with $u \in [\tilde{\nu}(0), \tilde{\nu}(0) + \alpha]$, as they would have been compliers in the original environment, but always-takers in the new environment. With monotonicity, α has a natural interpretation in that propensity for treatment is increased by α. With the existing illustration, there is no NA^* group, because α is small. When α is large enough, we will have a scenario like panel (iv), where, by doing a similar analysis as before, an NA^* group exists, but we no longer have a CC^* group. A corollary is that, under monotonicity, we can only either have NA^* or CC^*, but not both. In the empirical application, I have a relatively small α, so I have the CC^* group.

When monotonicity fails, we have a nonseparable selection equation $T = 1[\nu(Z, u) \geq 0]$, $u \sim U[0,1]$. One possible $\nu(.)$ function that can generate a nontrivial proportion of defiers

(though not unique or interpretable) is as follows:

$$\nu(Z,u) = 1\left[u \leq \frac{2}{3}\right] sin\left(3u\pi - \frac{3}{2}Zu\pi\right) + 1\left[u > \frac{2}{3}\right] sin\left(6u\pi - 6\pi - Z\left(3u\pi - 3\pi\right)\right). \quad (11)$$

The $\nu(Z,u)$ is application-specific. The goal here is not to argue for the empirical relevance of any particular $\nu(Z,u)$, but to show that there exists such a function that rationalizes the group formulation. This function is more clearly illustrated in Figure 2 in panel (i). The solid nu0 line plots $\nu(0,u)$ while the dashed nu1 line plots $\nu(1,u)$. For $u < 1/3$, both the solid and dashed lines are above 0, so they form the A group. For $u \in [1/3, 2/3]$, only the dashed line is above zero, so they would be treated when $Z = 1$ and untreated when $Z = 0$, so they are the C group. By doing the same analysis, $u \in [2/3, 5/6]$ are the defiers and $u \in [5/6, 1]$ are the never-takers. Panel (ii) illustrates the counterfactual environment where $T^* = 1[\nu(Z,u) + \alpha \geq 0]$, which shifts the ν function up by $\alpha = 0.5$, but the shape remains unchanged. In this non-monotonic environment, α is less interpretable. By looking at the regions where the dashed and solid lines are above or below 0, we can work out the new A^*, C^*, D^*, N^* groups. Panel (iii) combines the old and new groups from the previous panels to illustrate the region of u values that form the 9 treatment response groups. Unlike the separable case, it is possible to generate all 9 groups simultaneously.

If the researcher has a selection function ν in mind, such as (11), then it is possible to analytically derive the intercepts of the relevant curves and hence the q vector. With q known, bounds can be obtained conveniently using the linear program. Instead of estimating $\nu(Z,u)$ or imposing additional assumptions on ν, the approach in this paper transparently makes assumptions on the q vector by using $q_{C^*|C}, q_{C^*|N}$ as extrapolation parameters.

4 Empirical Application

In the Angrist and Evans (1998) problem, we are interested in the effect of a third child (T) on women's labor force participation (Y), and the instrument is whether the first two kids are of the same sex (Z). All variables are binary. Defiers are parents who have a preference for either two boys or two girls. Following Angrist and Evans (1998), I focus on the 1990 IPUMS data for mothers.[9] This empirical application illustrates how sensitivity analysis bridges the two extremes of monotonicity and worst-case bounds for LATE and LATE*, giving bounds at intermediate values of λ. The bounds vary continuously with λ, and are sensitive to failures of monotonicity. As a benchmark, De Chaisemartin (2017) argues that 5% of defiers is a conservative upper bound, which translates to $\lambda = 0.44$. Further, $q_D = 0$ is in the confidence interval constructed by Noack (2021).

We have $n = 380007$ observations and the proportions are given by $\widehat{\Pr}(Z = 1) = 0.504$, $\widehat{\Pr}(T = 1|Z = 1) = 0.402$ and $\widehat{\Pr}(T = 1|Z = 0) = 0.339$. Hence, the first stage is 0.063. Suppose we are interested in a counterfactual environment where there is a childcare subsidy that has a marginal effect on the probability of a third child. Here, $LATE^* = E[Y(1)-Y(0)|C^*]$ is the TE for people who used to be C and remain C^*, and people who were N but become C^* when there is a childcare subsidy. The units in the CC^* group have very strong preference for gender balance, and are hence unmoved by the subsidy, and the units in the NC^* group may be interpreted as those with weak preference for gender balance, but need a sufficient financial incentive to have a third child. Since existing papers (e.g., Carneiro et al. (2010)) calculate counterfactual effects at the margin i.e., $q_{C^*|C} \to 1, q_{C^*|N} \to 0$, I use

[9]The baseline implementation follows their Table 5 where no additional covariates were included. The implementation with covariates follows their Table 8(2).

$q_{C^*|C} = 0.99, q_{C^*|N} = 0.01$ to mimic their approach. This environment can also be interpreted as a 1% change in the relevant proportions.

Figure 3 presents the main result for sensitivity analysis. I impose the condition that $-0.3 \leq \mu_{g1} - \mu_{g0} \leq 0$ for all g, which is reasonable when the researcher believes that the TE for all groups is negative, and the data informs us how negative this TE is. With the OLS benchmark of -0.134, TE of -0.3 (which is more than twice the OLS benchmark) is a conservative a priori lower bound. Since the intersection of convex sets is convex, the additional a priori restriction of $\mu_{g1} - \mu_{g0} \in [-0.3, 0]$ satisfies the conditions of Theorem 1.[10] Only estimated bounds are presented, and issues on inference are omitted.

The OLS estimate of -0.134 is a benchmark for how informative the bounds are. Instruments are used in this context because we believe that OLS is downward biased: there are unobservable characteristics where people who are more likely to have a third child are also those who are less likely to work. Since IV is used to correct this downward bias, when the lower bound of the identified set hits the OLS estimate, the procedure is no longer informative about correcting the downward bias.

The curve labeled LATE is the original policy environment (i.e., no extrapolation). At $\lambda = 0$, there is point identification, resulting in the original LATE of -0.083. It is evident here that, even without extrapolating, bounds can be very wide (and uninformative) when monotonicity does not hold, but sensitivity analysis allows us to obtain the intermediate points. The lower bound of the LATE is above OLS for $\lambda \leq 0.2$, but it becomes uninformative for $\lambda \geq 0.25$. Hence, we can conclude that the LATE bounds are informative only for small values of λ. In the special case where the minimal constraint set $\mathcal{M}_m$ is imposed, the LATE

[10] $\mu_{g1} - \mu_{g0} \in [-0.3, 0]$ is the intersection of half planes, which is convex.

bounds are linear in λ, a result in Noack (2021).

The counterfactual environment labeled LATE* with $q_{A^*|C} = 0.01, q_{C^*|N} = 0.01$ incentivizes both groups into treatment. When q_{NC^*} is nonzero, the worst-case bounds of $\{-0.3, 0\}$ are imposed for $\mu_{NC^*1} - \mu_{NC^*0}$, and the bounds are no longer linear in λ. When monotonicity holds, the identified set is $[-0.103, -0.0737]$, which is informative; the bounds become uninformative for $\lambda \geq 0.2$. When we relax the sensitivity parameter, the upper bound eventually gets close to the trivial upper bound of 0. The numerical bounds on LATE* depend on the extrapolated environment: if we had extrapolated more, the LATE* at $\lambda = 0$ can be much wider than LATE at $\lambda = 0.1$. Hence, the bounds are informative only for small violations of monotonicity, and a counterfactual environment that differs locally.

The curve in Figure 4 uses the same set of covariates as in Angrist and Evans (1998). Implementing the procedure in Section C yields the curve in Figure 4. The result is qualitatively similar to Figure 3, but the magnitudes differ when controls are included. When there is no extrapolation and monotonicity holds, we point identify the TSLS estimand from the original study. As we extrapolate the environment and allow λ to increase, we obtain bounds on the LATE* that widen. The λ required before the result is uninformative is also higher than the setting without covariates.

5 Conclusion

This paper shows how policy relevant treatment parameters, including LATE and LATE*, can be partially identified with a sensitivity parameter that controls the extent monotonicity fails. Identification uses assumptions on proportions of the population that have a particular

response to the instrument instead of assumptions on the outcome function. This paper impacts empirical practice by providing a novel tool: sensitivity analysis of PRTP to failures of monotonicity, even for various treatment effects in extrapolated environments and when covariates are present. Depending on the empirical application, it may be more sensible to construct some structural model on selection $\nu(Z, u)$ (e.g., Chan et al. (2022)) instead of parameterizing the problem based on $E[Y(d)|g]$. Having a structural model is application-specific and left for future work.

References

Andrews, D. W. and Soares, G. (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica*, 78(1):119–157.

Angrist, J. D. and Evans, W. N. (1998). Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review*, pages 450–477.

Angrist, J. D. and Imbens, G. W. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176.

Blandhol, C., Bonney, J., Mogstad, M., and Torgovitsky, A. (2022). When is tsls actually late? *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2022-16).

Brinch, C. N., Mogstad, M., and Wiswall, M. (2017). Beyond late with a discrete instrument. *Journal of Political Economy*, 125(4):985–1039.

Carneiro, P., Heckman, J. J., and Vytlacil, E. (2010). Evaluating marginal policy changes and the average effect of treatment for individuals at the margin. *Econometrica*, 78(1):377–394.

Carneiro, P., Heckman, J. J., and Vytlacil, E. J. (2011). Estimating marginal returns to education. *American Economic Review*, 101(6):2754–81.

Chan, D. C., Gentzkow, M., and Yu, C. (2022). Selection with variation in diagnostic skill: Evidence from radiologists. *The Quarterly Journal of Economics*, 137(2):729–783.

Dahl, C. M., Huber, M., and Mellace, G. (2023). It is never too late: a new look at local average treatment effects with or without defiers. *The Econometrics Journal*, 26(3):378–404.

De Chaisemartin, C. (2017). Tolerating defiance? local average treatment effects without monotonicity. *Quantitative Economics*, 8(2):367–396.

Ding, P. and Lu, J. (2017). Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):757–777.

Duflo, E. and Saez, E. (2003). The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment. *The Quarterly Journal of Economics*, 118(3):815–842.

Dutz, D., Huitfeldt, I., Lacouture, S., Mogstad, M., Torgovitsky, A., and Van Dijk, W. (2021). Selection in surveys: Using randomized incentives to detect and account for nonresponse bias. Technical report, National Bureau of Economic Research.

Gneezy, U. and Rustichini, A. (2000). Pay enough or don't pay at all. *The Quarterly Journal of Economics*, 115(3):791–810.

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

Heckman, J. J. and Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation 1. *Econometrica*, 73(3):669–738.

Horowitz, J. L. and Manski, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association*, 95(449):77–84.

Huber, M., Laffers, L., and Mellace, G. (2017). Sharp iv bounds on average treatment effects on the treated and other populations under endogeneity and noncompliance. *Journal of Applied Econometrics*, 32(1):56–79.

Huber, M. and Mellace, G. (2015). Testing instrument validity for late identification based on inequality moment constraints. *Review of Economics and Statistics*, 97(2):398–411.

Ito, K., Ida, T., and Tanaka, M. (2021). Selection on welfare gains: Experimental evidence from electricity plan choice. Technical report, National Bureau of Economic Research.

Kamat, V. (2018). On the identifying content of instrument monotonicity. *arXiv preprint arXiv:1807.01661*.

Kitagawa, T. (2015). A test for instrument validity. *Econometrica*, 83(5):2043–2063.

Kitagawa, T. (2021). The identification region of the potential outcome distributions under instrument independence. *Journal of Econometrics*.

Klein, T. J. (2010). Heterogeneous treatment effects: Instrumental variables without monotonicity? *Journal of Econometrics*, 155(2):99–116.

Kline, P. and Walters, C. R. (2019). On heckits, late, and numerical equivalence. *Econometrica*, 87(2):677–696.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, 76(3):1071–1102.

Manski, C. F. (1989). Anatomy of the selection problem. *Journal of Human Resources*, pages 343–360.

Mogstad, M., Santos, A., and Torgovitsky, A. (2018). Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica*, 86(5):1589–1619.

Mogstad, M., Torgovitsky, A., and Walters, C. R. (2021). The causal interpretation of two-stage least squares with multiple instrumental variables. *American Economic Review*, 111(11):3663–3698.

Muralidharan, K., Singh, A., and Ganimian, A. J. (2019). Disrupting education? experimental evidence on technology-aided instruction in india. *American Economic Review*, 109(4):1426–60.

Noack, C. (2021). Sensitivity of late estimates to violations of the monotonicity assumption. *arXiv preprint arXiv:2106.06421*.

Richardson, T. S. and Robins, J. M. (2010). Analysis of the binary instrumental variable model. *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, 25:415–444.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, 3(2):135–146.

Shea, J. (2022). Testing for racial bias in police traffic searches. *University of Illinois, Champaign Urbana, USA*.

Small, D. S., Tan, Z., Ramsahai, R. R., Lorch, S. A., and Brookhart, M. A. (2017). Instrumental variable estimation with a stochastic monotonicity assumption. *Statistical Science*, 32(4):561–579.

van't Hoff, N., Lewbel, A., and Mellace, G. (2023). *Limited Monotonicity and the Combined Compliers LATE*. University of Southern Denmark, Faculty of Business and Social Sciences.

Wets, R. J.-B. (1985). On the continuity of the value of a linear program and of related polyhedral-valued multifunctions. In *Mathematical Programming Essays in Honor of George B. Dantzig Part I*, pages 14–29. Springer.

A Details on LATE*

A.1 Changing Instrument Value

Suppose we have a counterfactual instrument value Z^*. For illustration, I extrapolate the instrument rightward: in the original study, we have $Z \in \{0, 1\}$, but now we have $Z^* = 2$. The reasoning is similar if we wish to interpolate the instrument, or extrapolate leftward.[11] For every original group $G \in \{A, C, D, N\}$, individuals can have two possible responses at $Z^* = 2$, resulting in 8 treatment response groups, given by $(T(0), T(1), T^*(2))$.

LATE* is the TE for the compliers in the counterfactual environment. Since LATE is defined on an instrument pair, we have to consider which instrument pair the researcher is referring to. In the right-extrapolation exercise, one of the instrument values is $Z = 2$ that we do not have data for, so the LATE* can be defined either at the $(0, 2)$ pair or the $(1, 2)$ pair. In the Duflo and Saez (2003) running example, we ask what the LATE of the experiment would have been if we had given people \$30 instead of \$20. This corresponds to the $(0, 2)$ pair, because the control group still did not receive any financial incentive, and the treatment group simply received a larger incentive. Hence, $T^*(0) = T(0)$. In the right-extrapolation setup, compliers are those who switch their treatment status from 0 to 1 at the 0-2 instrument margin. This would be groups (0,0,1) and (0,1,1). Thus, $LATE^* = E[Y(1) - Y(0) | g \in \{(0, 0, 1), (0, 1, 1)\}]$. We can write this using the $c(q)'\mu$ notation

[11] Interpolation in Duflo and Saez (2003) with a \$20 incentive would ask what the LATE is if \$10 had been offered.

for the objective function:[12]

$$LATE^* = E[Y(1) - Y(0)|g \in \{(0,0,1),(0,1,1)\}] = \sum_{g,t} c_{gt}(q)\mu_{gt} = E[Y(1) - Y(0)|C^*].$$

The four observed distributions will now each be a mixture of four extrapolated groups. Namely, in $T = 0, Z = 0$, we originally had N and C. When extrapolating rightward, the original N consists of $(0,0,0) = NN^*$ and $(0,0,1) = NC^*$ while the original C consists of $(0,1,1) = CC^*$ and $(0,1,0) = CN^*$. Groups such as NA^* cannot exist in this environment. Hence, the distribution $Y|T = 0, Z = 0$ now contains a mixture of four groups (0,0,0), (0,0,1), (0,1,1) and (0,1,0). Given q, $\mathcal{M}(q)$ is well defined by mean compatibility as before, specialized to our binary context, and using only the information from the original environment that we have data for:

$$\mathcal{M}_m^{Ex}(q) := \left\{ \mu \in [0,1]^{16} : \sum_{g:T(z)=t} q_g \mu_{gt} = p_{tz} E[Y|T = t, Z = z] \qquad \forall z \in \{0,1\}, t \in \{0,1\} \right\}. \tag{A.1}$$

It remains to consider what $\mathcal{Q}(\lambda)$ is with extrapolation. For the extrapolation parameter, observe that $q_{C^*|A} = q_{C^*|D} = 0$ by construction, so we only have to consider $q_{C^*|C}$ and $q_{C^*|N}$. To make the extrapolated environment comparable with and without monotonicity, we can set $q_{N^*|C} = 0$, so $q_{C^*|C} = 1$. This rules out the $(0,1,0)$ group, which has defiers at the 1-2

[12] The interpretation for the LATE* is the TE of the meeting for people who are somewhat sensitive to financial incentives: this group includes (0,0,1) who are less sensitive to the incentive than the original compliers who are (0,1,1). The coefficient takes the form:

$$c_{gt}(q) = \begin{cases} \frac{(-1)^{1-t} q_{(0,0,1)}}{q_{(0,0,1)} + q_{(0,1,1)}} & \text{if } g = (0,0,1) \\ \frac{(-1)^{1-t} q_{(0,1,1)}}{q_{(0,0,1)} + q_{(0,1,1)}} & \text{if } g = (0,1,1) \\ 0 & \text{otherwise} \end{cases}.$$

margin. Hence, the only extrapolation parameter is $q_{C^*|N} = \Pr((0,0,1)|N)$. At the 0-1 and 1-2 instrument margins, sensitivity restrictions are:

$$\begin{aligned} q_{(1,0,0)} + q_{(1,0,1)} &\leq \lambda(q_{(0,1,1)} + q_{(0,1,0)}), \text{ and} \\ q_{(0,1,0)} + q_{(1,1,0)} &\leq \lambda(q_{(0,0,1)} + q_{(1,0,1)}). \end{aligned} \tag{A.2}$$

Hence, the constraint set is:

$$\mathcal{Q}_m^{Ex}(\lambda; q_{C^*|N}) = \left\{ q : \text{Eq. (5) and (A.2)}, \frac{q_{(0,0,1)}}{q_{(0,0,1)} + q_{(0,0,0)}} = q_{C^*|N}, q_{(0,1,0)} = 0 \right\}. \tag{A.3}$$

Corollary 1. *Suppose μ_{gt} is finite for all g, t. Then, using $\mathcal{M}(q) = \mathcal{M}_m^*(q)$ and $\mathcal{Q}(\lambda) = \mathcal{Q}_m^{Ex}(\lambda)$, the identified set for the LATE* is an interval.*

Extrapolation is characterized by q, so the analysis here does not depend on the value of Z. Regardless of whether the counterfactual Z is 1.1 or 100, the same argument from extrapolating rightward applies. Instead, the approach parameterizes the extent of extrapolation by the q vector. Namely, $Z = 1.1$ is an environment that is very similar to the original policy, so we expect $q_{C^*|N}$ close to zero. In contrast, with $Z = 100$, or a very different propensity score, it is analogous to a large extrapolation with $q_{C^*|N}$ close to 1. For instance, this could be a monetary incentive, so having a large incentive would move all N into taking up treatment.

Remark 5. When LATE* is defined as the TE on some subpopulation, we can use the same method to obtain TE on other subpopulations that are potentially more interesting. For instance, (1,0,1) is the group that are defiers at the \$0-\$20 margin, and compliers at

the $20-$30 margin in Duflo and Saez (2003). Behavioral studies on fund raisers in Gneezy and Rustichini (2000) show such behavior exist, where giving a bit of financial incentive disincentivises intrinsic effort, but offering a large financial incentive increases their effort. LATE* answers: For people with such behavioral responses, what is their take-up rate of a pension plan?

Remark 6. (Overpowering Experiments). Having a large incentive, say $100 in the Duflo and Saez (2003) experiment, can incentivize many people into treatment (the meeting). But this also includes people who go just for the money rather than because they are interested in the pension plan. If the incentive were $5 instead, the LATE of information on taking up the pension plan is likely larger, since this excludes the people who are not interested in the plan. Exercise in extrapolation places bounds on what the results of the experiment would have been if it had been designed differently.

A.2 Unified Econometric Approach

This section explains how the multi-dimensional optimization in the outer loop over the vector q in the two different counterfactual environments can be simplified into a one-dimensional optimization problem in the outer loop. The inner loop is then a function of this one-dimensional parameter, and solves a linear program. Hence, estimation of the bounds is tractable. The treatment propensity counterfactual is explained in Section 3.1.

There is an analogous result in the counterfactual that changes the instrument value.

For right extrapolation, with $\mathcal{M}_m^{Ex}(q)$ as defined in (A.1), define:

$$\overline{R}^{Ex}(q) = \max_{\mu \in \mathcal{M}_m^{Ex}(q)} LATE^* = \max_{\mu \in \mathcal{M}_m^{Ex}(q)} \frac{q_{(0,0,1)}(\mu_{(0,0,1),1} - \mu_{(0,0,1),0}) + q_{(0,1,1)}(\mu_{(0,1,1),1} - \mu_{(0,1,1),0})}{q_{(0,0,1)} + q_{(0,1,1)}}.$$

Lemma 1. *Consider* $q = (q_{(0,0,0)}, q_{(0,0,1)}, q_{(0,1,0)}, q_{(0,1,1)}, q_{(1,0,0)}, q_{(1,0,1)}, q_{(1,1,0)}, q_{(1,1,1)})$. *If* $q_A = q_{(1,1,0)} + q_{(1,1,1)}$, $q_{CC^*} = q_{(0,1,1)}$, $q_{CC'^*} = q_{(0,1,0)}$, $q_D = q_{(1,0,0)} + q_{(1,0,1)}$, $q_{NC^*} = q_{(0,0,1)}$, *and* $q_{NC'^*} = q_{(0,0,0)}$, *then* $\overline{R}^{Ex}(q) = \tilde{R}(\tilde{q})$.

With Proposition 1 and Lemma 1 telling us that the inner loop of the two counterfactual programs can be solved using a 6-parameter problem, the main result of this section is:

Theorem 3. *With scalar* q_D, *there exists an invertible matrix* J *and vector* $v(q_D)$ *that are known functions of* $(q_D, p_{tz}, q_{C^*|G})$ *such that:*

$$\max_{q \in \mathcal{Q}_m^{TC}(\lambda)} \overline{R}^{TC}(q) = \max_{q_D \in \mathcal{Q}_d^{TC}(\lambda)} \tilde{R}(J^{-1}v(q_D)), \text{ and} \tag{A.4}$$

$$\max_{q \in \mathcal{Q}_m^{Ex}(\lambda)} \overline{R}^{Ex}(q) = \max_{q_D \in \mathcal{Q}_d^{Ex}(\lambda)} \tilde{R}(J^{-1}v(q_D)), \tag{A.5}$$

where

$$\mathcal{Q}_d^{TC}(\lambda) = \left\{ q_D \in [0,1] : q_D \leq \frac{\lambda(p_{00} + p_{11} - 1)}{1 - \lambda} \right\}, \text{ and} \tag{A.6}$$

$$\mathcal{Q}_d^{Ex}(\lambda) = \mathcal{Q}_d^{TC}(\lambda) \cap \left\{ q_D : \frac{1 - (p_{00} + p_{11} - q_D)(1 - q_{C^*|C}) - 2q_{C^*|C}}{-2 + q_{C^*|C}} \leq \lambda \left(q_D + \frac{-1 + p_{11} + q_{C^*|N} + q_D}{-2 + q_{C^*|N}} \right) \right\}. \tag{A.7}$$

The result for the minimum is analogous.

The upshot of Theorem 3 is that bounds on the object of interest such as $\max_{q \in \mathcal{Q}_m^{TC}(\lambda)} \overline{R}^{TC}(q)$ can be obtained by solving a one-dimensional optimization problem in q_D instead of a multi-

dimensional problem. Observe that we are using the same $\tilde{R}$, J, and $v(q_D)$ in both problems, so the inner problem is econometrically identical. Further, $\tilde{R}(\tilde{q})$ is a linear program in $\tilde{\mu}$, so it can be solved efficiently. To prove this result, first use the previous two lemmas to obtain equivalence in the inner program. Then, observe that there are 5 linearly independent equality constraints in the $\tilde{q}$ problem, so once q_D is known, $\tilde{q} = J^{-1}v(q_D)$ is known. Their expressions are provided in the proof in Appendix D. The remaining constraint set for q_D comes from sensitivity restrictions that have been set up differently.

B Inference

Using Theorem 3, there are six groups when using $\mathcal{M}_m$ and $\mathcal{Q}_m$: $(A, CA^*, CC^*, D, NC^*, NC'^*)$. Proportion restrictions on p_{tz} yield:

$$\begin{aligned} E[(q_{CC^*} + q_{CA^*} + q_{NC^*} + q_{NC^{*\prime}} - (1-T))(1-Z)] = 0, \text{ and} \\ E[(q_{CC^*} + q_{CA^*} + q_A - T)Z] = 0. \end{aligned} \tag{B.8}$$

Mean compatibility constraints are:

$$\begin{aligned} E\left[\left(\frac{q_{NC^*}\mu_{NC^*0} + q_{NC'^*}\mu_{NC'^*0} + q_{CC^*}\mu_{CC^*0} + q_{CA^*}\mu_{CA^*0}}{q_{NC^*} + q_{NC'^*} + q_{CC^*} + q_{CA^*}} - Y\right)(1-T)(1-Z)\right] = 0, \\ E\left[\left(\frac{q_A\mu_{A1} + q_D\mu_{D1}}{q_A + q_D} - Y\right)T(1-Z)\right] = 0, \\ E\left[\left(\frac{q_{NC^*}\mu_{NC^*0} + q_{NC'^*}\mu_{NC'^*0} + q_D\mu_{D0}}{q_{NC^*} + q_{NC'^*} + q_D} - Y\right)(1-T)Z\right] = 0, \text{ and} \\ E\left[\left(\frac{q_A\mu_{A1} + q_{CA^*}\mu_{CA^*1} + q_{CC^*}\mu_{CC^*1}}{q_A + q_{CA^*} + q_{CC^*}} - Y\right)TZ\right] = 0. \end{aligned} \tag{B.9}$$

Finally, there are inequality constraints imposed by a binary outcome, and further restrictions on the q's imposed by the sensitivity parameter:

$$\begin{gathered} 0 \leq \mu_{gt} \leq 1, \qquad 0 \leq q_g \leq 1, \qquad q_D \leq \lambda(q_{CC^*} + q_{CA^*}), \qquad \sum_g q_g = 1, \\ \frac{q_{CC^*}}{q_{CC^*} + q_{CA^*}} = q_{C^*|C}, \text{ and} \qquad \frac{q_{NC^*}}{q_{NC^*} + q_{NC'^*}} = q_{C^*|N}. \end{gathered} \tag{B.10}$$

In general, with moment equalities and inequalities, algorithms such as Andrews and Soares (2010) can be applied. In this application, uncertainty from the data only features in moment equalities of (B.8) and (B.9), so I proceed only with moment equalities.

Parameters are denoted $\theta := (q', \mu')'$. Let $m(\theta) = 0$ denote the moment conditions of (B.8) and (B.9), where $m(\theta)$ is the vector of expectations, and let $\hat{m}(\theta)$ be the sample analog. Under standard CLT assumptions, $\sqrt{n}(\hat{m}(\theta) - m(\theta)) \xrightarrow{d} N(0, \Omega)$, where Ω is the variance covariance matrix for the moment conditions. Since $m(\theta) = 0$, $T(\theta) := n\hat{m}(\theta)'\Omega^{-1}\hat{m}(\theta) \xrightarrow{d} \chi^2_6$ for the test statistic $T(\theta)$. The χ^2 distribution has 6 degrees of freedom because there are 6 moment conditions. We do not reject θ if $T(\theta) \leq \chi^2_6(1-\alpha) =: c_\alpha$ for a size α test, where c_α denotes the critical value. Since Ω can be consistently estimated, plug in the sample analog $\hat{\Omega}$ to use feasible test statistic $\hat{T}(\theta) := n\hat{m}(\theta)'\hat{\Omega}^{-1}\hat{m}(\theta) \xrightarrow{d} \chi^2_6$ for inference.

Finally, to calculate the upper bound for the confidence interval, solve the following problem:

$$\max_{\theta := (q', \mu')'} c(q)'\mu \qquad s.t. \qquad \hat{T}(\theta) \leq c_\alpha, \text{ and } \theta \text{ satisfies Eq. (B.10).} \tag{B.11}$$

Calculating the lower bound is analogous. This problem corresponds to having a partially identified θ that is in confidence set C_θ, and we are interested in the confidence set (CS) of $g(\theta) = c(q)'\mu$, and in particular the extremum of the CS of $g(\theta)$. The procedure described

here is identical to the projection method described in Dufour (1997) Section 5.2 for obtaining a CS for $g(\theta)$. These optimization problems can be implemented using canned packages.

C Extension to Incorporate Covariates

In many situations, the instrument is valid only conditional on covariates, and hence researchers may wish to incorporate covariates into their model. Covariates W feature in the model through Assumption 1, which would be: $E[Y(t)|g,z,W] = E[Y(t)|g,W]$ and $\Pr(g|z,W) = \Pr(g \mid W)$ for all g,z,W. There are at least two ways that they can be incorporated. One way mimics Noack (2021, appendix A3): we can run the aforementioned procedure at every covariate level w, then reweigh the bounds by the covariate masses. While this procedure yields more restrictions and hence tighter bounds, it is computationally intensive, requires the researcher to make an assumption on defier bounds and extrapolation parameters for every covariate value, and does not nest the two-stage least squares (TSLS) estimand in general. It is also cumbersome when W is continuous. Without further assumptions, this is the only procedure available to the best of my knowledge.

Instead, I propose a second approach for the LATE*. With covariates and without extrapolation, researchers run the TSLS regression as a standard practice. Hence, a goal of the procedure is to nest TSLS with covariates as a special case without extrapolation and when monotonicity holds, and I provide conditions under which such a procedure is reasonable. This procedure allows some dependence of μ_{gt} and q_g on W, and augments the existing linear program.

Without extrapolation and with monotonicity, parametric assumptions are already re-

quired to interpret TSLS with heterogeneous treatment effects when there are covariates (e.g., Blandhol et al. (2022)). In particular, theory has developed around interpreting TSLS as some weighted average of LATE's (i.e., weighted average of treatment effect of compliers at different covariate values), but it is often not obvious why that particular weighting is the most interesting. To circumvent the issue of which weighted average of LATE's should be targeted, I consider the environment where the treatment effect for compliers is the same at all covariate values, motivating the assumption below.

To be clear on notation, linear regressions are run with a constant, and W does not include the intercept term. T and Z are binary. Assume the following:

Assumption 4. *(a) For* $g \in \{CA^*, CC^*, NC^*, D\}$, $q_g = Pr(g|W_1) = Pr(g|W_2)$ *for all* W_1, W_2, *while for* $g \in \{NC'^*\}$, $q_g = \alpha_g^{int} + \alpha_g' W$. q_A *can depend on* W *flexibly.* $q \in [0,1]^{d_q}$.

(b) $\mu_{Dt}(W) = \eta_{Dt} + \xi_D(W)$; *for* $g \in \{CA^*, CC^*, NC^*\}$, $\mu_{gt}(W) = \eta_{gt} + \xi_g' W$; *for* $g \in \{NC'^*\}$ $\mu_{gt}(W) = \eta_{gt} + \xi_{gt}' W$. *Finally,* $\mu_{At}(W)$ *can depend on* W *flexibly.*

(c) $E[Z|W] = \tilde{\xi}^{int} + \tilde{\xi}' W \in [0,1]$.

There are three parts to the assumption. Part (a) makes restrictions on the q vector; part (b) makes restriction on the μ vector; part (c) ensures that linear projections are interpretable as conditional expectations.

In Assumption 4(a), we cannot have q_{CA^*}, q_{CC^*}, q_D depend on W so that the TSLS estimand does not depend on W. When we are interested in the LATE*, we additionally cannot have q_{NC^*} depend on W so that the target object LATE* does not depend on W. $q_{NC'^*}$ is linear in W so that the conditional expectation of $Y|Z=0, T=0, W$ is quadratic in W.

Other parametric forms may be possible, but the expression of the conditional expectation has to match accordingly. q_A is allowed to depend on W flexibly, as it is differenced out in the procedure.

The TE for $g \in \{CA^*, CC^*, NC^*, D\}$ must be constant for all covariate values so that the LATE* and the TSLS estimand do not depend on W. This requirement is denoted in Assumption 4(b) as having the same ξ_g for treated and untreated potential outcomes so that the treatment effect $\eta_{g1} - \eta_{g0}$ does not depend on W. Having the same TE is required even without extrapolation and with monotonicity so that TSLS identifies the unique LATE. The functional form in $\mu(W)$ is required in this paper's framework so that we can match coefficients and obtain a linear program. $\xi_D(W)$ can depend flexibly on W because it is not used in coefficient matching. In contrast, for $g \in \{CA^*, CC^*, NC^*\}$, $\mu_{gt}(W)$ is linear in W so that the conditional expectation of $Y|Z = 0, T = 0, W$ is quadratic in W. For $g \in \{NC'^*\}$, $\mu_{gt}(W)$ allows ξ_{gt} to vary by potential treatments, and its linearity is required for coefficient matching. No restriction is required for $\mu_{At}(W)$.

Once q_D and the extrapolation parameters $q_{C^*|C}$ and $q_{C^*|N}$ are fixed, the rest of the q vector and α's are point-identified. Details are in Appendix D.3. With α and q point identified, and $p_{00}(W) := Pr[T = 0|Z = 0, W]$, the assumption implies:

$$p_{00}(W) = \alpha^{int}_{NC'^*} + \alpha'_{NC'^*} W + q_{NC^*} + q_{CC^*} + q_{CA^*},$$

and hence

$$p_{00}(W)E\left[Y|Z=0,T=0,W\right]=\alpha_{NC'*}^{int}\eta_{NC'*0}+q_{NC^*}\eta_{NC^*0}+q_{CC^*}\eta_{CC^*0}+q_{CA^*}\eta_{CA^*0}$$
$$+\left(\eta_{NC'*0}\alpha_{NC'*}'+\alpha_{NC'*}^{int}\xi_{NC'*0}'+q_{NC^*}\xi_{NC^*}'+q_{CC^*}\xi_{CC^*}'+q_{CA^*}\xi_{CA^*}'\right)W+\alpha_{NC'*}'W\xi_{NC'*0}'W. \tag{C.12}$$

The object of interest can then be written as:

$$LATE^*=\frac{1}{q_{CC^*}+q_{NC^*}}\left(q_{CC^*}\eta_{CC^*1}+q_{NC^*}\eta_{NC^*1}-q_{CC^*}\eta_{CC^*0}-q_{NC^*}\eta_{NC^*0}\right).$$

Then, the proposed algorithm for finding LATE* uses the following steps (S):

S1. Run TSLS regression with the full set of controls W to obtain the TSLS estimand β.

S2. Calculate the sample analogs of q and α based on the identification argument of Appendix D.3 to construct $p_{00}(W)$. Using the partition on $T=0,Z=0$, run the regression of $p_{00}(W)Y$ on 1, W, and $(\alpha_{NC'*}'W)W$. Denote the intercept as γ_0.

S3. Set up the linear program, whose objective is LATE*, optimizing over parameters η and appropriate a priori linear restrictions. Additionally, use the following linear restrictions:

(a) $\beta\left(q_{CA^*}+q_{CC^*}-q_D\right)=q_{CC^*}\left(\eta_{CC^*1}-\eta_{CC^*0}\right)+q_{CA^*}\left(\eta_{CA^*1}-\eta_{CA^*0}\right)+q_D\left(\eta_{D0}-\eta_{D1}\right)$, and

(b) $\gamma_0=\alpha_{NC'*}^{int}\eta_{NC'*0}+q_{NC^*}\eta_{NC^*0}+q_{CC^*}\eta_{CC^*0}+q_{CA^*}\eta_{CA^*0}$.

To see how this procedure is reasonable, first observe that S1 and S2 merely calculates objects used in S3, so it suffices to motivate S3. When there is no extrapolation, the TSLS

estimand without covariates is given by $[q_C(\eta_{C1} - \eta_{C0}) + q_D(\eta_{D0} - \eta_{D1})]/(q_C - q_D)$. Since the assumptions are constructed such that the TSLS estimand does not depend on W, S3(a) uses an analogous expression for the TSLS that accommodates the counterfactual environment. S3(b) is motivated by (C.12). By regressing the left-hand side on W and a quadratic term, the intercept term γ_0 must match the structural objects described.

Due to the constraint in S3(a), the proposed algorithm collapses exactly to TSLS without extrapolation and under monotonicity. We also use covariate information through S3(b).

D Proof of Results

D.1 Proofs for Section 2

Let b denote the target object, so for a given q, the set of feasible values in the inner problem is:

$$\mathcal{B}(q) = \{b \in \mathbb{R} : b = c(q)'\mu \text{ for some } \mu \in \mathcal{M}(q)\}. \tag{D.1}$$

Lemma 2. *Under Assumption 1, suppose that $\mathcal{M}(q)$ is convex for some fixed q. Then, either $\mathcal{M}(q)$ is empty and hence $\mathcal{B}(q)$ is empty, or the closure of $\mathcal{B}(q)$ is equal to the interval $[\underline{R}(q), \overline{R}(q)]$, defined in (8). Further, if $\mathcal{M}(q)$ can be written as a system of linear inequalities in μ, both optimization problems are linear programs.*

Proof of Lemma 2. Convex $\mathcal{M}(q)$ is either empty or nonempty. If $\mathcal{M}(q)$ is empty, then by definition $\mathcal{B}(q) = \emptyset$. Next, consider a nonempty $\mathcal{M}(q)$. Since a linear mapping of a convex set also yields a convex set, and $c(q)'\mu$ is a linear map of μ, it follows that $\mathcal{B}(q)$ is a convex set. Thus, any $b \in [\underline{R}(q), \overline{R}(q)]$ must also be in $\mathcal{B}(q)$. Proving that optimization problems are

indeed linear programs is straightforward from its construction. The constraints are linear in μ and the objective function is a linear function of μ. $\square$

Proof of Theorem 1. Proof for an empty $\mathcal{B}_\lambda$ is identical to the proof in Lemma 2. Only consider nonempty $\mathcal{M}(q)$. The objective is to show that any $b \in [\underline{\beta}_\lambda, \overline{\beta}_\lambda]$ is achievable for some $q \in \mathcal{Q}(\lambda)$. For this, I first show first show that $\overline{R}(q)$ and $\underline{R}(q)$ are continuous in q. Continuity of these objects can then be used to complete the argument.

Apply Theorem 2 from Wets (1985) that the objective value of a linear program is continuous in its hyperparameters. The sufficient condition for the theorem is that the feasible set in both the primal and dual linear programs are continuous in the hyperparameters. For the dual problem, it is assumed that $\mathcal{M}(q)$ is bounded by Assumption 3(a), so Corollary 11 from Wets (1985) implies that the feasible set of the dual problem is continuous in the hyperparameters. Turning to the primal problem, continuity in the hyperparameters is given by Assumption 3(b). The conditions for the Wets (1985) theorem is hence satisfied. Then, using Theorem 2 from Wets (1985), and the fact that the composition of continuous functions is continuous, with $c(q)$ continuous in q due to Assumption 3(c), $\underline{R}(q)$ and $\overline{R}(q)$ are continuous in q.

It remains to show that any $b \in [\underline{\beta}_\lambda, \overline{\beta}_\lambda]$ is achievable for some $q \in \mathcal{Q}(\lambda)$. Pick a point $q^0 \in \mathcal{Q}(\lambda)$ such that $\mathcal{M}(q^0)$ is nonempty. This is guaranteed to exist because because we work in the environment where $\exists q \in \mathcal{Q}(\lambda)$ s.t. $\mathcal{M}(q)$ is nonempty. Using Lemma 2, any $r \in [\underline{R}(q^0), \overline{R}(q^0)]$ can be satisfied by some $\mu \in \mathcal{M}(q^0)$. Since an analogous argument can be made for $[\underline{\beta}_\lambda, \underline{R}(q^0)]$, it suffices to show that for all $b \in [\overline{R}(q^0), \overline{\beta}_\lambda]$, there exists some $q \in \mathcal{Q}(\lambda)$ such that $b = \overline{R}(q)$. If $\overline{R}(q^0) = \overline{\beta}_\lambda$, the desired conclusion is immediate, so I focus

on $\overline{R}(q^0) < \overline{\beta}_\lambda$.

Let $\overline{q}$ be the q that achieves $\overline{\beta}_\lambda$ i.e., $\overline{\beta}_\lambda = \overline{R}(\overline{q})$. With slight abuse of notation, let $[q^0, \overline{q}]$ denote the set of convex combinations on $\mathbb{R}^{d_q}$ between q^0 and $\overline{q}$, so it is a convex set. Since by Assumption 3(d) $\mathcal{Q}(\lambda)$ is convex, any $q \in [q^0, \overline{q}]$ must also lie in $\mathcal{Q}(\lambda)$ and is hence feasible. Since convex sets are connected, $[q^0, \overline{q}]$ is connected. Using the fact that the image of a connected set is connected for a continuous mapping, $\overline{R}([q^0, \overline{q}])$ is a connected set. Since $\overline{R}(q^0)$ and $\overline{R}(\overline{q})$ are both feasible, and $\overline{R}(.) \in \mathbb{R}$, $[\overline{R}(q^0), \overline{R}(\overline{q})] \subseteq \overline{R}([q^0, \overline{q}])$. Hence, $\exists q \in [q^0, \overline{q}] \subseteq \mathcal{Q}(\lambda)$ such that $\overline{R}(q) \in [\overline{R}(q^0), \overline{R}(\overline{q})]$.

□

The following lemma is used to prove Theorem 2.

Lemma 3. *Suppose that for any $z, z' \in \mathcal{Z} \subset \mathbb{N}$, $z > z'$ implies $\Pr(T(z) = 1) \geq \Pr(T(z') = 1)$. For any $n_1, n_2 \in \mathbb{Z}_+$ with $n_2 > n_1$ and $z + n_2 \in \mathcal{Z}$, if $\sum_{g \in S^d_{(z,z+n_1)}} q_g \leq \lambda \sum_{g' \in S^c_{(z,+n_1)}} q_{g'}$ and $\sum_{g \in S^d_{(z+n_1,z+n_2)}} q_g \leq \lambda \sum_{g' \in S^c_{(z+n_1,z+n_2)}} q_{g'}$, then $\sum_{g \in S^d_{(z,z+n_2)}} q_g \leq \lambda \sum_{g' \in S^c_{(z,+n_2)}} q_{g'}$.*

Proof of Lemma 3. The defiers at the $(z, z + n_2)$ margin switch exactly once: either at $(z, z + n_1)$ or $(z + n_1, z + n_2)$. Individuals who switch twice are either always takers or never takers when looking at the $(z, z + n_2)$ margin. It also means that they will be compliers at either one of the two margins and defiers at the other margin. This implies

$$S_2 := S^d_{(z,z+n_2)} \subset S^d_{(z,z+n_1)} \cup S^d_{(z+n_1,z+n_2)} =: S_1.$$

To be precise, S_2 consists of defiers who switch exactly once, and $S_1 \backslash S_2$ consists of defiers who switch twice, resulting in their being compliers at one margin.

Let $q(.)$ be the probability measure on sets. By assumption, $q(S^d_{(z,z+n_1)}) \leq \lambda q(S^c_{(z,z+n_1)})$ and $q(S^d_{(z+n_1,z+n_2)}) \leq \lambda q(S^c_{(z+1,z+n_2)})$. Due to binary treatment, the sets $S^d_{(z,z+n_1)}, S^d_{(z+n_1,z+n_2)}$ are disjoint. Similarly, the sets $S^c_{(z,z+n_1)}, S^c_{(z+n_1,z+n_2)}$ are also disjoint. Summing the inequalities,

$$q(S_1) = q(S^d_{(z,z+n_1)}) + q(S^d_{(z+n_1,z+n_2)}) \leq \lambda(q(S^c_{(z,z+n_1)}) + q(S^c_{(z+n_1,z+n_2)})).$$

Consider the set $S^c_{(z,z+n_1)} \cup S^c_{(z+n_1,z+n_2)}$. This set consists of compliers at the $(z, z+n_2)$ margin (which implies $S^c_{(z,z+n_2)}$ is a subset), and $S_1 \backslash S_2$. Namely, $S^c_{(z,z+n_1)} \cup S^c_{(z+n_1,z+n_2)} = (S_1 \backslash S_2) \cup S^c_{(z,z+n_2)}$. Observe that $S^c_{(z,z+n_2)}$ is the set of compliers who switch their treatment status exactly once in the correct direction. Then, the summed inequality is:

$$\begin{aligned} q(S_2) + (q(S_1) - q(S_2)) &\leq \lambda(q(S^c_{(z,z+n_2)}) + q(S_1) - q(S_2)) \\ \Rightarrow q(S_2) &\leq \lambda q(S^c_{(z,z+n_2)}) - (1-\lambda)(q(S_1) - q(S_2)) \\ \Rightarrow q(S^d_{(z,z+n_2)}) &\leq \lambda q(S^c_{(z,z+n_2)}). \end{aligned}$$

□

Proof of Theorem 2. The condition of Lemma 3 is satisfied due to how $\mathcal{Z}$ is defined. For $z' > z$, we can write $z' = z + l$ with $l > 0$. Thus, it is sufficient to show that $\sum_{g \in S^d_{(z,z+l)}} q_g \leq \lambda \sum_{g' \in S^c_{(z,z+l)}} q_{g'}$ for any $l \in \mathbb{Z}_+$.

Prove by induction. Apply Lemma 3, using $n_1 = 1, n_2 = 2$. Since $\sum_{g \in S^d_{(z,z+1)}} q_g \leq \lambda \sum_{g' \in S^c_{(z,z+1)}} q_{g'}$ and $\sum_{g \in S^d_{(z+1,z+2)}} q_g \leq \lambda \sum_{g' \in S^c_{(z+1,z+2)}} q_{g'}$, obtain $\sum_{g \in S^d_{(z,z+2)}} q_g \leq \lambda \sum_{g' \in S^c_{(z,z+2)}} q_{g'}$. Suppose $\sum_{g \in S^d_{(z,z+l)}} q_g \leq \lambda \sum_{g' \in S^c_{(z,z+l)}} q_{g'}$, and we want to show $\sum_{g \in S^d_{(z,z+l+1)}} q_g \leq \lambda \sum_{g' \in S^c_{(z,z+l+1)}} q_{g'}$

due to adjacency. Apply Lemma 3 with $n_1 = l, n_2 = l+1$ to obtain the result. □

Proof of Proposition 1. The objective is to show $\max_{\mu \in \mathcal{M}_m^{TC}(q)} LATE^* = \max_{\tilde{\mu} \in \tilde{\mathcal{M}}_m(\tilde{q})} LATE^*$, where $LATE^*$ is a function of $(q_{CC^*}, q_{NC^*}, \mu_{CC^*1}, \mu_{CC^*0}, \mu_{NC^*1}, \mu_{NC^*0})$.

Let $h(\mu) = \mu_4 := (\mu_{CC^*1}, \mu_{CC^*0}, \mu_{NC^*1}, \mu_{NC^*0})'$ denote the function that extracts the subvector μ_4 from a higher-dimensional vector $\mu \in \mathbb{R}^{18}$. Then, since $LATE^*$ only contains μ_4, $\max_{\mu \in \mathcal{M}_m^{TC}(q)} LATE^* = \max_{\mu_4 \in \mathcal{M}_4^{TC}(q)} LATE^*$, where

$$\mathcal{M}_4^{TC}(q) = \left\{ \mu_4 : \mu_4 = h(\mu), \mu \in \mathcal{M}_m^{TC}(q) \right\}.$$

Let $\tilde{h}(.)$ similarly extract μ_4 from $\tilde{\mu} \in \mathbb{R}^{12}$. Then, $\max_{\mu \in \tilde{\mathcal{M}}_m(\tilde{q})} LATE^* = \max_{\mu_4 \in \tilde{\mathcal{M}}_4(\tilde{q})} LATE^*$:

$$\tilde{\mathcal{M}}_4(\tilde{q}) = \left\{ \mu_4 : \mu_4 = \tilde{h}(\mu), \mu \in \tilde{\mathcal{M}}_m(\tilde{q}) \right\}.$$

Hence, it is sufficient to show that $\tilde{\mathcal{M}}_4(\tilde{q}) = \mathcal{M}_4^{TC}(q)$ to obtain the result. Do change of variables for $\mathcal{M}_m^{TC}(q)$, with the given substitution for q. Then, the respective μ's can be redefined:

$$\begin{aligned}
\mu_{NC'^*t} &= \frac{1}{q_{NC'^*}}[q_{NA^*}\mu_{NA^*t} + q_{ND^*}\mu_{ND^*t} + q_{NN^*}\mu_{NN^*t}], \\
\mu_{Dt} &= \frac{1}{q_D}[q_{DA^*}\mu_{DA^*t} + q_{DD^*}\mu_{DD^*t}], \text{ and} \\
\mu_{CC'^*t} &= \mu_{CA^*t}.
\end{aligned}$$

Then, equality constraints characterized by $\sum_{G:T(z)=t} q_G \mu_{Gt} = p_{tz} E[Y|T = t, Z = z]$ are identical to those of $\tilde{\mathcal{M}}_m(\tilde{q})$. Since the counterfactual μ's are weighted averages of the

original μ's, the counterfactual μ's in $\tilde{\mu}$ must also lie in $[0,1]$, so $\mathcal{M}_4^{TC}(q) \subseteq \tilde{\mathcal{M}}_4(\tilde{q})$. Then, it is sufficient to show $\tilde{\mathcal{M}}_4(\tilde{q}) \backslash \mathcal{M}_4^{TC}(q) = \emptyset$. The set $\tilde{\mathcal{M}}_4(\tilde{q}) \backslash \mathcal{M}_4^{TC}(q)$ contains values of μ_4 where the μ's in $\tilde{\mu}$ are in $[0,1]$, but the individual components that construct the averages, such as μ_{DD^*t} need not be in $[0,1]$. However, restrictions on μ_4 only occur through the averages in the equality constraints, in addition to $\mu_4 \in [0,1]^4$. Thus, since the averages in $\tilde{\mathcal{M}}_4(\tilde{q})$ and in $\mathcal{M}_4^{TC}(q)$ face the same constraints, μ_4 face the same constraints in both sets. Hence, $\tilde{\mathcal{M}}_4(\tilde{q}) \backslash \mathcal{M}_4^{TC}(q) = \emptyset$, which then implies $\tilde{\mathcal{M}}_4(\tilde{q}) = \mathcal{M}_4^{TC}(q)$. □

D.2 Proofs for Appendix A

Proof of Corollary 1. The condition satisfies Assumption 3(a). It is sufficient to check other conditions of Assumption 3, then apply Theorem 1. Continuity of $c(q)$ is immediate. $\mathcal{M}_m^{Ex}(q)$ is convex because it is intersection of linear subspaces. To see that $\mathcal{Q}(\lambda) = \mathcal{Q}_m^{Ex}(\lambda)$ satisfies convexity, take any two elements $q^0, q^1 \in \mathcal{Q}_m^{Ex}(\lambda)$ with $q^0 \neq q^1$. Form convex combination $q^* = \alpha q^0 + (1-\alpha) q^1$, with $\alpha \in (0,1)$. Taking the weighted sums of the constraints on q^0 and q^1, $q^* \in \mathcal{Q}_m^{Ex}(\lambda)$ is immediate. □

Proof of Lemma 1. By redefining groups as stated, the proof is analogous to Proposition 1. □

Proof of Theorem 3. By defining $\tilde{q}$ appropriately and applying Proposition 1 and Lemma 1, $\overline{R}^{TC}(q) = \tilde{R}(\tilde{q})$ and $\overline{R}^{Ex}(q) = \tilde{R}(\tilde{q})$.

Then, consider equality restrictions in $\mathcal{Q}_m^{TC}(\lambda)$ and $\mathcal{Q}_m^{Ex}(\lambda)$. In both constraint sets, there are 5 linearly independent restrictions, with 2 from p_{11} and p_{00}, 1 from the fact that probabilities sum to 1, and 2 from the extrapolation parameters. We can also write $q_D = q_D$

as a trivial relationship. Writing these 6 equations in matrix form, we have $J\tilde{q} = v(q_D)$, where

$$J = \begin{bmatrix} 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & q_{C^*|C} - 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & q_{C^*|N} - 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}, \text{ and}$$

$$v(q_D) = (p_{00}, p_{11}, 1, q_{C^*|C}, q_{C^*|N}, q_D)'.$$

Note that $det(J) = 4 - 2q_{C^*|C} - 2q_{C^*|N} + q_{C^*|C}q_{C^*|N} = (2 - q_{C^*|N})(2 - q_{C^*|C}) \neq 0$. Since J is invertible, $\tilde{q} = J^{-1}v(q_D)$.

It remains to consider the inequality restrictions imposed by sensitivity parameters. In $\mathcal{M}_m^{Ex}(q)$, the specific choice of q_{DD^*} and q_{ND^*} makes restrictions on q_{CC^*} and q_{NC^*}. Since q_{CC^*} and q_{NC^*} are arguments in the optimization problem, the optimum is found by using the least restrictive setting for q_{CC^*} and q_{NC^*}. Due to Proposition 1, setting $q_{DD^*} = q_{ND^*} = 0$ is innocuous. Then, $q_{DD^*} + q_{ND^*} \leq \lambda(q_{CC^*} + q_{NC^*})$ is automatically satisfied. The only relevant sensitivity restriction is $q_D = q_{DD^*} + q_{DA^*} \leq \lambda(q_{CC^*} + q_{CA^*})$. Using the substitution in Proposition 1, and $\tilde{q} = J^{-1}v(q_D)$, the constraint set results. This constraint set will give us the optimum, because the objective value has to perform weakly better with one fewer constraint.

Turning now to $\mathcal{Q}_m^{Ex}(\lambda)$, use the least restrictive values as before. One can set $q_{(1,0,0)} = 0$

so $q_D = q_{(1,0,1)}$, and upper bound at the 0-2 margin is the largest possible, while the inner problem does not change. Then, the constraint at the 0-1 margin will be $q_D \leq \frac{\lambda(p_{00}+p_{11}-1)}{1-\lambda}$ by using the relevant substitutions and $\tilde{q} = J^{-1}v(q_D)$. With the relevant substitutions, and setting $q_{(1,1,0)} = 0$ (to create the most flexible constraint), the constraint at the 1-2 margin is $q_{CC'^*} \leq \lambda(q_{NC^*} + q_D)$. Finally, substitute $\tilde{q} = J^{-1}v(q_D)$ to obtain the required inequality. □

D.3 Derivations for Appendix C

$$\begin{aligned}
p_{00}(W)E\left[Y|Z=0,T=0,W\right] &= \left(\alpha^{int}_{NC'^*} + \alpha'_{NC'^*}W\right)\left(\eta_{NC'^*0} + \xi'_{NC'^*0}W\right) + q_{NC^*}\left(\eta_{NC^*0} + \xi'_{NC^*}W\right) \\
&\quad q_{CC^*}\left(\eta_{CC^*0} + \xi'_{CC^*}W\right) + q_{CA^*}\left(\eta_{CA^*0} + \xi'_{CA^*}W\right) \\
&= \alpha^{int}_{NC'^*}\eta_{NC'^*0} + q_{NC^*}\eta_{NC^*0} + q_{CC^*}\eta_{CC^*0} + q_{CA^*}\eta_{CA^*0} \\
&\quad + \left(\eta_{NC'^*0}\alpha'_{NC'^*} + \alpha^{int}_{NC'^*}\xi'_{NC'^*0} + q_{NC^*}\xi'_{NC^*} + q_{CC^*}\xi'_{CC^*} + q_{CA^*}\xi'_{CA^*}\right)W \\
&\quad + \alpha'_{NC'^*}W\xi'_{NC'^*0}W
\end{aligned}$$

First Stage Identification. Let the first-stage regression be $T = \pi Z + \theta_0 + \theta'W + v$. Then, the first stage estimand is:

$$\begin{aligned}
\pi &= \frac{E\left[T\left(Z - E^*\left[Z|W\right]\right)\right]}{E\left[\left(Z - E^*\left[Z|W\right]\right)^2\right]} \\
&= \frac{E\left[E\left[Z|W\right]\left(1 - E\left[Z|W\right]\right)\left(E\left[T|Z=1,W\right] - E\left[T|Z=0,W\right]\right)\right]}{E\left[E\left[Z|W\right]\left(1 - E\left[Z|W\right]\right)\right]}.
\end{aligned}$$

Using Assumption 4(a),

$$E[T|Z=1,W]-E[T|Z=0,W]=q_A(W)+q_{CA^*}+q_{CC^*}-q_D-q_A(W)$$
$$=q_{CA^*}+q_{CC^*}-q_D=\pi.$$

For a given q_D and an extrapolation parameter $q_{C^*|C}$, since π is identified, q_{CA^*}, q_{CC^*}, q_D are all identified. Since $E[T|Z=0,W]-q_D=q_A(W)=\alpha_A^{int}+\alpha_A'W$, by regressing $T-q_D$ in the partition with $Z=0$ on W, α_A^{int} and α_A are identified. Conversely,

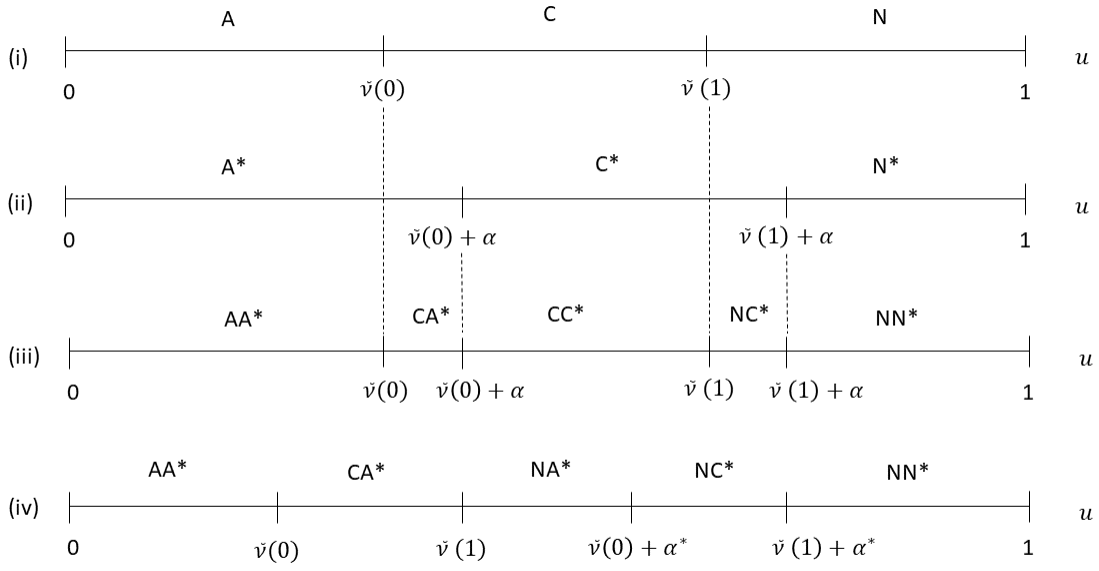
$$1-E[T|Z=1,W]=q_{NC^*}+q_{NC'^*}(W)+q_D$$
$$1-E[T|Z=1,W]-q_{NC^*}-q_D, \text{ and } =q_{NC'^*}(W)=\alpha_{NC'^*}^{int}+\alpha_{NC'^*}'W.$$

Observe that $1-E[T|Z=1]=q_{NC^*}+E[q_{NC'^*}(W)]+q_D$. Since q_D and $q_{C^*|N}$ are known, q_{NC^*} is identified. Then, by regressing $1-T-q_{NC^*}-q_D$ in the partition with $Z=1$ on W, $\alpha_{NC'^*}^{int}$ and $\alpha_{NC'^*}$ are identified.

TSLS Estimand. Due to Assumption 4(c), the TSLS estimand is given by

$$\beta=\frac{E[Y(Z-E^*[Z|W])]}{E[T(Z-E^*[Z|W])]}$$
$$=\frac{E[E[Z|W](1-E[Z|W])(E[Y|Z=1,W]-E[Y|Z=0,W])]}{E[E[Z|W](1-E[Z|W])(E[T|Z=1,W]-E[T|Z=0,W])]}.$$

Figure 1: Separable Selection Equation



Then, due to Assumptions 4(a) and 4(b),

$$\begin{aligned} E[Y|Z=1,W]-E[Y|Z=0,W] &= q_{CC^*}\left(\mu_{CC^*1}(W)-\mu_{CC^*0}(W)\right)+q_{CA^*}\left(\mu_{CA^*1}(W)-\mu_{CA^*0}(W)\right) \\ &\quad + q_D\left(\mu_{D0}(W)-\mu_{D1}(W)\right) \\ &= q_{CC^*}\left(\eta_{CC^*1}-\eta_{CC^*0}\right)+q_{CA^*}\left(\eta_{CA^*1}-\eta_{CA^*0}\right)+q_D\left(\eta_{D0}-\eta_{D1}\right), \text{ and} \\ \beta &= \frac{q_{CC^*}\left(\eta_{CC^*1}-\eta_{CC^*0}\right)+q_{CA^*}\left(\eta_{CA^*1}-\eta_{CA^*0}\right)+q_D\left(\eta_{D0}-\eta_{D1}\right)}{q_{CA^*}+q_{CC^*}-q_D}. \end{aligned}$$

E Figures

Figure 2: Nonseparable Selection Equation

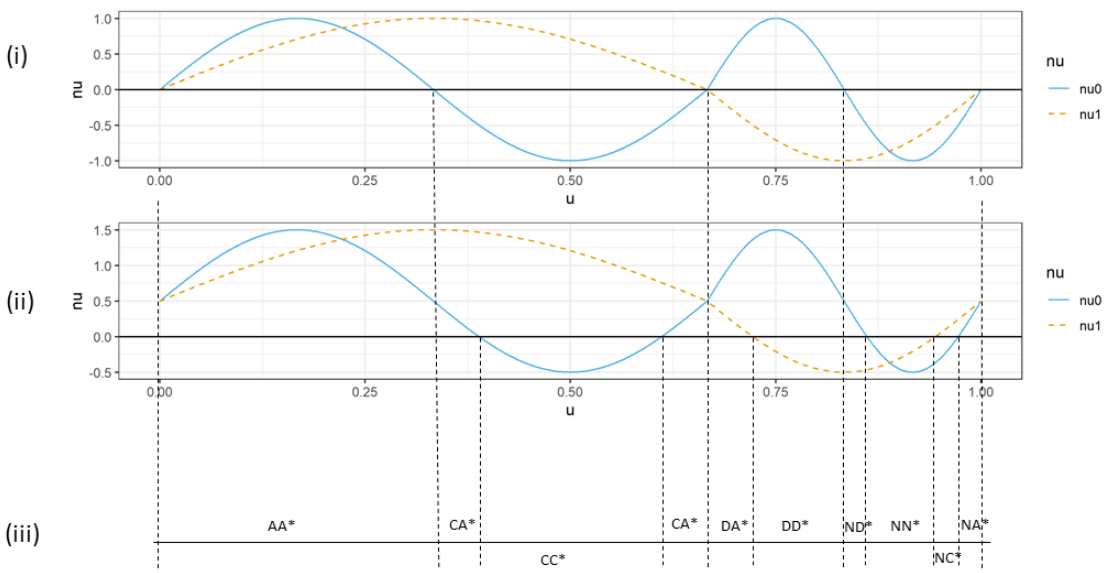


Figure 3: Plot of $LATE^* = E[Y(1) - Y(0)|C^*]$ bounds against λ without covariates. Impose $-0.3 \leq \mu_{g1} - \mu_{g0} \leq 0$ for all g. LATE has $q_{C^*|C} = 1, q_{C^*|N} = 0$ so there is no extrapolation; LATE* has $q_{C^*|C} = 0.99, q_{C^*|N} = 0.01$. The red horizontal line is the OLS benchmark of -0.134.

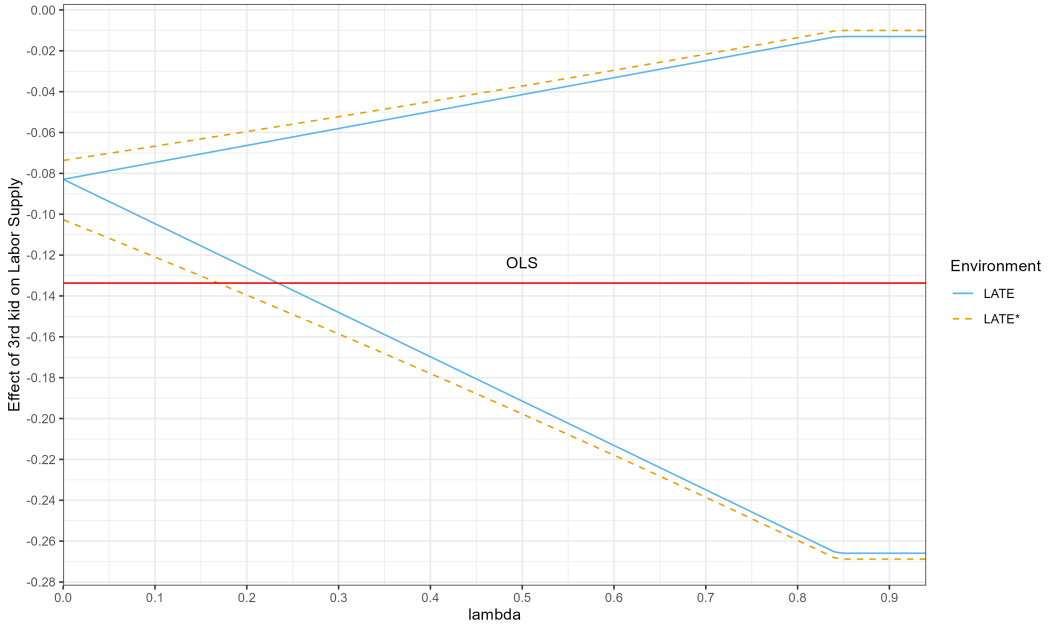


Figure 4: Plot of $LATE^* = E[Y(1) - Y(0)|C^*]$ bounds against λ with covariates. Impose $-0.3 \leq \mu_{g1} - \mu_{g0} \leq 0$ for all g. LATE has $q_{C^*|C} = 1, q_{C^*|N} = 0$ so there is no extrapolation; LATE* has $q_{C^*|C} = 0.99, q_{C^*|N} = 0.01$. The red horizontal line is the OLS benchmark of -0.164. Covariates include age of mother, age at first birth, gender of the first two kids, and race indicators for white, black, and hispanic.