Sensitivity Analysis of Policy Relevant Treatment Effects to Failures of Monotonicity

Luther Yap *

January 26, 2022

Abstract

Researchers and policy makers are often interested in the external validity of a study's conclusions. An object of interest that addresses concerns of external validity is the policy-relevant treatment effect (PRTE), the average treatment effect of people who respond to the instrument in a different policy environment. This paper provides a method for bounding PRTE without functional forms or monotonicity, and gives a novel interpretation for PRTE without monotonicity as treatment effects of various subpopulations. The bounding framework uses the proportion of defiers relative to compliers as a sensitivity parameter and formulates the problem as a linear program. This method can thus be used to assess the sensitivity of various treatment effects to violations of the monotonicity assumption when using instrumental variables to make inferences on causal effects for various subpopulations. Bounds are sharp for binary outcomes and can be consistently estimated. I illustrate this method with an empirical application where defiers are present. The framework can be extended to multivalued instruments, and to parameterizing the problem using distributions to obtain quantile objects and sharp bounds. The method has many applications in applied work, including constructing nonparametric bounds on causal objects in counterfactual environments, allowing the researcher to flexibly incorporate restrictions, and testing model specification.

Keywords: Instrumental variables, treatment effects, local average treatment effect, LATE, policy relevant treatment effect, PRTE, partial identification, monotonicity, sensitivity analysis

*Department of Economics, Princeton University. Email: lyap@princeton.edu.

1 Introduction

Since the seminal work of Angrist and Imbens (1994) and Angrist et al. (1996), which interpreted an instrumental variables (IV) estimand as a local average treatment effect (LATE) in a model of unobserved heterogeneity in treatment effects, much attention has been paid to the fact that in various contexts, there are various objects of interest beyond the LATE (see, for instance, Heckman and Vytlacil (2005)). Several empirical studies have also paid attention to issues on external validity of the study's conclusions. One object of interest for external validity is the policy-relevant treatment effect (PRTE), the average treatment effect of people who respond to the instrument (i.e, LATE) in a different policy environment.[1]

The PRTE is a benchmark for external validity and robustness of one's conclusion that is under-appreciated in literature. Consider the Angrist and Evans (1998) study that considers the effect of having a third child on the mother's labor supply. They use an indicator for whether the first two kids are of the same sex as an instrument for the third child since parents generally have a preference for gender balance among their children. Suppose the conclusion from the IV regression is that having a third child decreases the mother's employment probability by 0.083 for some sub-population. This value is specific to the policy environment surrounding childcare in the dataset. Would we still have the same conclusion when the government gives a $5000 subsidy for childcare? What would the effect of a third child be for compliers in the new environment? Another example is from Duflo and Saez (2003), who are interested in the effect of a meeting (information treatment) on the take-up rate of a pension plan. If we had given people $10 instead of $5 to attend the meeting, how different would the conclusion on the effect of the meeting be? More generally, with minimal assumptions, what can we say about external validity? What would the result of the experiment be if it had been designed differently? How robust are the results to a different policy environment? These are questions that are answered by the PRTE.[2]

Existing methods to obtain the PRTE relies on monotonic response to the instrument. Namely, procedures like Mogstad et al. (2018) rely on the marginal treatment effect (MTE) framework (Vytlacil (2002); Heckman and Vytlacil (2005)). MTE assumes an additively separable treatment selection equation, which is equivalent to monotonicity, the assumption that the instrument acts in the same direction for all individuals.[3] However, monotonicity may not be a realistic assumption in many applications. For instance, in Angrist and Evans (1998), parents have a preference for gender balance among their children, so 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: under monotonicity, parents would have a third child if their first two children feature a balanced gender portfolio, while not having a third child if their first two children are of the same gender.[4] Consequently, this paper proposes a method to place bounds on PRTE without

[1] A different policy environment could refer to a different instrument value, if the instrument is itself a policy: we have data when the instrument takes value $Z \in \{0, 1\}$, but we might want to make statements about environments when $Z = 2$. Another counterfactual policy might increase the propensity of treatment for everyone.

[2] Empirical examples that explicitly mention PRTE include Muralidharan et al. (2019) who are interested in the effect of an education program when implemented for 90 days when the experiment only ran for 86 days, Ito et al. (2021) who are interested in welfare effects under a counterfactual take-up incentive regime, and Carneiro et al. (2011) who are interested in the LATE for counterfactual education policies.

[3] With heterogeneous treatment effects, the result in Angrist et al. (1996) is that if monotonicity holds, two-stage least squares estimand (TSLS) is interpretable as the average treatment effect of people who change their treatment status in response to the instrument (i.e., compliers).

[4] Other examples of monotonicity violation are provided in Appendix C.

monotonicity by leveraging a linear program similar to Mogstad et al. (2018), but does not require the MTE framework.

The goal of this paper is to provide an intuitive, and computationally tractable method to bound the PRTE. This is achieved by using a sensitivity parameter that places an upper bound on the proportion of defiers (i.e., people who respond to the instrument in the opposite direction relative to the majority) relative to compliers (i.e., people who respond to the instrument in the same direction as the majority). This allows us to assess how bounds with defiers compare to those without defiers, thereby assessing the sensitivity of the conclusions to monotonicity violations. I contribute to existing literature by (1) Providing a tool for sensitivity analysis of monotonicity violations for various objects of interest through defier restrictions, and showing that the identified set is convex. This is especially useful for PRTE without functional form assumptions, which to the best of my knowledge does not yet have a viable method for sensitivity analysis. (2) Interpreting objects of interest in counterfactual policy environments without monotonicity. To the best of my knowledge, the PRTE literature relies on the MTE framework, so discussion and interpretation is based on a monotonic environment. The analogous PRTE objects without monotonicity and their interpretation remains unaddressed.

To achieve this, I parameterize the problem using conditional means of potential outcomes for various subgroups characterized by their response to the instrument so that bounds on objects of interest can be written as linear combinations of these parameters. The basic setup uses a binary instrument so that defier proportion is immediately interpretable, but the method can be extended. The constraint set flexible, but the baseline specification is characterized by trimming bounds suggested in Lee (2009) within each conditional outcome distribution, and observed mean restrictions across conditional outcome distributions. Then, the problem can be written as a linear program with a convex identified set.

At a high level, (partial) identification without functional forms is possible in a potential outcomes model because the data already places some restrictions on potential outcomes, and objects of interest merely reweights these potential outcomes. Consider a stylized example where we are interested in the expected outcome of some subpopulation C when we use an instrument value $Z = 2$ that we do not have observations for, denoted $E[Y|C, Z = 2]$. We have data only for instrument values $Z \in \{0, 1\}$. We have binary treatment $T \in \{0, 1\}$, so subpopulation C has two mean potential outcomes μ_{C1} and μ_{C0}, and every individual has two potential outcomes Y_{1i} and Y_{0i}. Existing data gives us restrictions on these mean potential outcomes: for instance, under monotonicity, the two-stage least squares estimand is $TSLS = \mu_{C1} - \mu_{C0}$, when C denotes compliers. We do not know how various individuals in C will respond to $Z = 2$, but regardless of their response, their outcome must always be either Y_{1i} or Y_{0i}. Divide C into two groups: Co1 denotes those who do not take up treatment under $Z = 2$ and Co2 denotes those who take up treatment. Suppose proportion $q_{Co1|C}$ of C are in the first group. Then, the object of interest can be written as $E[Y|C, Z = 2] = q_{Co1|C}\mu_{Co1,0} + (1 - q_{Co1|C})\mu_{Co2,1}$, where μ denotes the mean potential outcomes. Bounds on this object can be obtained using the fact that $\mu_{C1} = q_{Co1|C}\mu_{Co1,1} + (1 - q_{Co1|C})\mu_{Co2,1}$ (in fact, the distribution of potential outcomes for C is a mixture of analogous distributions of Co1 and Co2), with an analogous expression for μ_{C0}, and that restrictions on μ_{C1} and μ_{C0} still apply in the counterfactual environment. The input in this exercise is hence a single parameter - the mixing proportion $q_{Co1|C}$. This assumption on mixing proportion allows identification without making assumptions on the outcome space. Without monotonicity, we simply add defiers as a subpopulation in the mixture, and the same intuition applies to obtain bounds.

The approach presented in this paper is more general than necessary for the purposes discussed. The exposition uses the minimal assumptions approach and focuses on the PRTE. First, my framework is sufficiently flexible that the researcher can obtain bounds completely nonparametrically, or impose parametric assumptions, or any intermediate approach by setting up the constraint set appropriately.[5] This follows Mogstad et al. (2018), who set up the problem as a linear program with a flexible constraint set that is amenable to the assumptions that the researcher is willing to make, However, since identification of heterogeneous treatment effects are often done without functional form assumptions, it is rather peculiar that functional form assumptions are imposed when we want to evaluate the robustness of the results.[6] PRTE is hence discussed without functional form assumptions in this paper. Second, the framework is general enough to nest other objects of interest (e.g., LATE), but there is existing literature that considers sensitivity analysis to LATE (Noack, 2021), and it is known that the worst case bounds for the average treatment effect (ATE) are obtained when monotonicity holds (Kitagawa, 2021). Sensitivity analysis for PRTE is novel in literature and nontrivial.

1.1 Literature Review

This paper speaks to three broad strands of literature: (1) Responses to monotonicity violations (2) Partial Identification and Sensitivity Analysis (3) PRTE.

One response to potential monotonicity violations is to find testable implications for monotonicity. If we do not reject the null of monotonicity, then we can proceed with causal inference as before. There are several papers that do this, including Kitagawa (2015), Mourifié and Wan (2017) and Huber and Mellace (2015). While this makes the claim of monotonicity falsifiable, it is still not obvious what to do when the null of monotonicity is rejected. Attempts to relax monotonicity include using probabilistic monotonicity in DiNardo and Lee (2011) for random experiments with imperfect compliance and average monotonicity in Frandsen et al. (2019) for the judges design. Semenova (2020) has a form of monotonicity conditional on covariates. But even these weaker assumptions may be violated when there is systematic sorting by some unobserved variable. Another approach is to use the compliers-defiers assumption in De Chaisemartin (2017) which allows interpretation of IV estimates as the treatment effect on a subset of compliers. This only allows point identification of a very small subset of the population. Heckman and Pinto (2018) suggest the unordered monotonicity condition, which allows for a particular class of defiers.

Instead of relaxing the assumption, there is other literature that aims to put bounds on (partially identify) objects of interest or conduct sensitivity analysis to its violations. Balke and Pearl (1997) has a linear program to obtain bounds on the ATE in the population, generalized from binary outcomes to a continuous outcome space in Kitagawa (2021). Manski (1989) and Horowitz and Manski (2000) have bounds generated from the worst-case outcome values. By parameterizing the proportion of defiers as a sensitivity parameter, we can get narrower bounds. Noack (2021) has a method to conduct sensitivity analysis of the LATE with respect to two parameters: proportion of defiers and treatment effect heteroegeneity between compliers and defiers. Beyond this literature,

[5] This nests the parametric approach of Brinch et al. (2017) and Kline and Walters (2019) and the Manski (1989) minimal assumption bounds, where the relevant missing values are imputed by the largest and smallest possible values of the outcome.

[6] Functional form assumptions are reasonable when evaluating the policy impact, since that involves extrapolation to answer a particular policy question.

there is also broader interest in sensitivity analysis, as in Masten and Poirier (2018), Masten and Poirier (2020) and Armstrong and Kolesár (2021). The method in this paper yields Balke and Pearl (1997) for the ATE when the outcome is binary. My approach is more general than Noack (2021) in that it works even if we do not consider the LATE, which is what she focuses on. But in other ways, it is more restrictive, because I only use defier proportion as a sensitivity parameter, whereas she uses both defier proportion and heterogeneity in outcomes, and she has sharp bounds in closed form. Consequently, my bounds will be the same as Noack (2021) for LATE under binary outcomes and we use the worst-case heterogeneity. Nonetheless, my main contribution is still to address PRTE instead of LATE. [7]

Thus far, bounds on PRTE are dependent on the MTE framework in Vytlacil (2002) and Heckman and Vytlacil (2005). The MTE framework is used in applied work including Dobbie et al. (2018) and Kowalski (2018). Mogstad et al. (2018) (henceforth MST) use an additively separable selection equation in a linear program to obtain bounds on PRTE, which assumes monotonicity, whereas my method treats monotonicity violations as a sensitivity parameter. MST's constraint set involves moment conditions in the form of mean equality restrictions only, so a theoretical point here is that we can include trimming bounds to get tighter bounds, while preserving the linear program. My bounds will be identical to MST when monotonicity holds, and the constraint set only consists of proportion equalities and mean restrictions for a discrete instrument.

1.2 Outline

The rest of this paper discusses the proposed method, its applications, and its extensions. Section 2 explains the general framework in forming bounds; Section 3 applies it to PRTE; Section 4 outlines how the method can be implemented; Section 5 applies the method to the gender preference problem of Angrist and Evans (1998); Section 6 suggests some extensions; and Section 7 concludes.

2 General Framework for Sensitivity Analysis to Monotonicity Violations

This section presents the generic framework for sensitivity analysis for various treatment effects. Subsections 2.1 and 2.2 generalizes the setup in Mogstad et al. (2018) (MST) to allow for a generic space of treatment response groups. When monotonicity holds, the setup reduces exactly to MST. Subsection 2.3 proposes a sensitivity parameter for the setting, which is the proportion of defiers relative to compliers, novel in this literature.

2.1 Setting

We observe variables (T, Z, Y), denoting treatment, instrument, and outcome respectively. We are interested in the effect of endogenous T on Y. Outcome Y is unrestricted. For ease of exposition,

[7] It is not obvious why we would still be interested in the LATE when monotonicity is violated. Under monotonicity, the LATE is simply what we get when we run TSLS, and that happens to have a nice interpretation.

suppose treatment T and instrument Z are both binary. The method generalizes for any discrete T and discrete Z, which is discussed in Section 6. Assume there are no covariates.[8]

Index treatment response groups by G. In the basic binary setup, this gives four response groups. For each observation in the data, we can observe the instrument $Z \in \{0, 1\}$ and the treatment $T \in \{0, 1\}$. For each combination, the individuals can be always-takers (A), compliers (C), defiers (D) or never-takers (N).

Table 1: Classification

	$Z = 0$	$Z = 1$
$T = 0$	N,C	N,D
$T = 1$	A,D	A,C

I denote conditional means of potential outcomes for various treatment response groups as μ_{gt}, and their respective proportions q_g. The binary case has groups $G \in \{A, C, D, N\}$ and treatment $t \in \{0, 1\}$.

$$\mu_{gt} := m_t(g) = E[Y_t|G = g] \qquad q_g := Pr(g)$$

Remark 1. *The $m_t(g)$ notation maps directly to the MTE framework of Heckman and Vytlacil (2005), where marginal treatment responses are denoted $m_t(u)$, where u is the unobservable that affects the treatment status. Namely, $T = 1[\nu(Z) - u \geq 0]$, and one can innocuously normalize $u \sim U[0, 1]$ such that low values of u get treated. The u hence indexes response groups, reflecting how the individual responds to various values of $\nu(Z)$.*

Let Y_{tz} denote potential outcomes with treatment t and instrument z, and T_z denote treatment given instrument z.

Assumption 1. *The following properties hold:*

(a) (Independence). We have $(Y_{00}, Y_{01}, Y_{10}, Y_{11}, T_0, T_1) \perp Z$.

(b) (Exclusion). For all $t \in \{0, 1\}$, $Y_{t0} = Y_{t1} = Y_t$.

(c) (Finite Second Moment). $E[Y_t^2] < \infty$ for $t \in \{0, 1\}$

(d) Observations are drawn independently from the same population.

The first stage coefficient is $FS := P[T = 1|Z = 1] - P[T = 1|Z = 0]$, and the reduced form coefficient is $RF := E[Y|Z = 1] - E[Y|Z = 0]$. Consequently, the two-stage-least-squares (TSLS) estimand is $TSLS = RF/FS$. This is equivalently known as the Wald estimator. The standard result from Angrist and Imbens (1994) is that if Assumption 1 holds, then

$$FS = q_C - q_D$$
$$TSLS = \frac{RF}{FS} = \frac{q_C(\mu_{C1} - \mu_{C0}) - q_D(\mu_{D1} - \mu_{D0})}{q_C - q_D}$$

[8] If there are covariates, then we can run the procedure for each covariate value, then take the weighted average, an approach similar to Noack (2021).

Assumption 2. *(Strict Monotonicity, SM). Where T_{iz} is an indicator of treatment for individual i conditional on assignment z, either $T_{i1} \geq T_{i0}$ $\forall i$ or $T_{i1} \leq T_{i0}$ $\forall i$.*

Without defiers (i.e. $q_D = 0$), it is straightforward to see that $TSLS = \mu_{C1} - \mu_{C0}$, which is the treatment effect on compliers (often referred to as the local average treatment effect (LATE)), and that no-defiers is identical to (SM). But when there are defiers, $TSLS$ can no longer be interpreted in the same way. The rest of the paper maintains Assumption 1, but relaxes Assumption 2.

2.2 Method Description

Objects of interest, including PRTE, can generically be written as:

$$\begin{aligned} \beta &= \sum_t \int_g m_t(g)\omega_t(g)d\nu(g) \\ &= \sum_g q_g(m_0(g)\omega_0(g) + m_1(g)\omega_1(g)) \\ &= c(q)'\mu \end{aligned} \tag{1}$$

The first line presents the object of interest in the most general form for any discrete T and space of response groups. This maps directly to the MTE formulation in MST. The second equation specializes the object to our expository setting of binary T, and a discrete and finite number of response groups G. The final equation shows how the second line can be expressed as the dot product of a coefficient vector c dependent on the proportions and a vector of conditional means μ.

The setup can be specialized further to the binary context. Suppose ex ante that the researcher is willing to make an assumption on the proportion of defiers (i.e., q_D is known). Then, the proportions of all other groups are immediately identified by the data. Namely, given q_D, and observing $p_{tz} := Pr(T = t|Z = z)$ with Assumption 1, all other proportions are given by:

$$\begin{aligned} q_A &= 1 - p_{00} - q_D \\ q_N &= 1 - p_{11} - q_D \\ q_C &= p_{00} + p_{11} - 1 + q_D \end{aligned} \tag{2}$$

Denote the vector of proportions as $q := (q_A, q_C, q_D, q_N)$, and the vector of conditional means:

$$\mu := (\mu_{A1}, \mu_{A0}, \mu_{C1}, \mu_{C0}, \mu_{D1}, \mu_{D0}, \mu_{N1}, \mu_{N0})'$$

Our objects of interest will take the general form of $c(q)'\mu$, where $c : [0,1]^4 \to \mathbb{R}^8$ is a function that maps the q vector into coefficients on μ. In general, we would be interested in some combination of treatment effects (TE) $\mu_{G1} - \mu_{G0}$ for various groups. For instance, the average treatment effect (ATE) weights the TE for each group by their relevant proportion, so $c_{ATE}(q) = (q_A, -q_A, q_C, -q_C, q_D, -q_D, q_N, -q_N)'$. More examples are provided in Table 2, and their details are in Appendix D. The PRTE is more involved, and hence discussed in the next section.

The μ object is constrained by the observable distribution in the data, and I use $\mathcal{S}$ to denote the set

Table 2: Example of Coefficient Vectors on parameter vector $\mu \in \mathbb{R}^8$

Object of Interest	$c(q)'$
Treatment Effect on Compliers (TEC)	$(0,0,1,-1,0,0,0,0)$
Treatment Effect on Defiers (TED)	$(0,0,0,0,1,-1,0,0)$
Treatment Effect on Marginal Population (TEM)	$(0,0,q_C,-q_C,q_D,-q_D,0,0)/(q_C+q_D)$
Average Treatment Effect (ATE)	$(q_A,-q_A,q_C,-q_C,q_D,-q_D,q_N,-q_N)$
Treatment Effect On Treated (TOT)	$(q_A,-q_A,q_C,-q_C,q_D,-q_D,0,0)/(q_C+q_A+q_D)$

of μ that satisfies defined equality and inequality constraints from data. The researcher can specify what these constraints are, but we require the equality constraints to take the form in Equation 3 and inequality constraints to preserve convexity of μ. Indexing the moments by s,

$$\sum_g q_g(\mu_{g0}\omega_{0s}(g) + \mu_{g1}\omega_{1s}(g)) = \Gamma_s \tag{3}$$

From Assumption 1, the weakest constraints we can impose are trimming bounds and mean restrictions.

First consider trimming bounds.[9] We observe four outcome distributions $Y|T,Z$. To illustrate this method, consider cell $(T=0,Z=1)$, which contains N and D, and we observe the Y_0 distribution. Since we know q_D, to obtain the lower bound on μ_{N0}, we trim away the top mass of the distribution that is commensurate with q_D, and take the mean of the remaining distribution. Similarly, to obtain the upper bound of μ_{N0}, trim the bottom q_D/p_{01} mass and take the means of what is left. Denote the upper and lower trimming bounds as $y_{GT}^{L(Z)}$ and $y_{GT}^{U(Z)}$ for group G with instrument Z and treatment T.

$$\begin{aligned} y_{GT}^{L(Z)} &:= E[Y|T,Z,Y \le y_{q_G}] \\ y_{GT}^{U(Z)} &:= E[Y|T,Z,Y \ge y_{1-q_G}] \\ y_q &:= F^{-1}(q) \text{ with } F \text{ the cdf of } Y, \text{ conditional on } T,Z \end{aligned} \tag{4}$$

Since there are 4 distributions, and 2 groups in each distribution, we have 8 such pairs of bounds listed below in Inequality 5.

$$\begin{aligned} y_{A1}^{L(1)} \le \mu_{A1} \le y_{A1}^{U(1)} \qquad & y_{A1}^{L(0)} \le \mu_{A1} \le y_{A1}^{U(0)} \\ y_{N0}^{L(1)} \le \mu_{N0} \le y_{N0}^{U(1)} \qquad & y_{N0}^{L(0)} \le \mu_{N0} \le y_{N0}^{U(0)} \\ y_{C1}^{L(1)} \le \mu_{C1} \le y_{C1}^{U(1)} \qquad & y_{C0}^{L(0)} \le \mu_{C0} \le y_{C0}^{U(0)} \\ y_{D1}^{L(0)} \le \mu_{D1} \le y_{D1}^{U(0)} \qquad & y_{D0}^{L(1)} \le \mu_{D0} \le y_{D0}^{U(1)} \end{aligned} \tag{5}$$

Another set of constraints is the observed means across the four distributions $Y|T,Z$: the weighted mixture of the conditional means of the two groups should match the observed mean. Then, we

[9] Subsequent work like Semenova (2020) call these the Lee Bounds, but I use "trimming" bounds as they are more descriptive of the method, and remains faithful to the term used in Lee (2009).

would define $\mathcal{S} := \{\mu : \mu \text{ satisfies Equation 6 and Inequality 5}\}$

$$\begin{aligned} q_N\mu_{N0} + q_C\mu_{C0} &= (q_N + q_C)E[Y|T=0, Z=0] \\ q_A\mu_{A1} + q_D\mu_{D1} &= (q_A + q_D)E[Y|T=1, Z=0] \\ q_N\mu_{N0} + q_D\mu_{D0} &= (q_N + q_D)E[Y|T=0, Z=1] \\ q_A\mu_{A1} + q_C\mu_{C1} &= (q_A + q_C)E[Y|T=1, Z=1] \end{aligned} \tag{6}$$

Remark 2. *The two sets of constraints provide different information. Trimming bounds prevent $\mu_{C0} \to -\infty$ and $\mu_{N0} \to \infty$ while maintaining the observed mean. This is important when ex ante bounds are not imposed. The observed mean ensures that values chosen are consistent across different cells: for instance, the μ_{A1} used in $(T = 1, Z = 0)$ must be the same μ_{A1} used in $(T = 1, Z = 1)$.*

One may wish to place ex ante restrictions on μ such that $\mu \in \mathcal{M}$. For instance, if the Roy (1951) model is applicable in a particular context such that $\mu_{A1} \geq \mu_{A0}$ and $\mu_{N0} \geq \mu_{N1}$, these constraints can be added into the constraint set. For complete agnosticism, $\mathcal{M} = \mathbb{R}^8$. Alternatively, when we have binary outcomes, $\mathcal{M} = [0, 1]^8$. Denote the set of μ's that satisfies the above constraints as $\mathcal{M}_\mathcal{S}$.

$$\mathcal{M}_\mathcal{S} := \mathcal{M} \cap \mathcal{S}$$

Denote our target object as b, so the identified set can be written as:

$$\mathcal{B} = \{b \in \mathbb{R} : b = c(q)'\mu \text{ for some } \mu \in \mathcal{M}_\mathcal{S}\}$$

The generic form of our optimization problem will be a convex program. The objective value is then the bound on our objects of interest. If the constraint set can be written as a system of linear inequalities, then we will obtain linear programs.

Lemma 1. *(Linearity and Convexity). Suppose that $\mathcal{M}$ is convex and q fixed. Then, either $\mathcal{M}_\mathcal{S}$ is empty and hence $\mathcal{B}$ is empty, or the closure of $\mathcal{B}$ is equal to the interval $[\underline{\beta}, \overline{\beta}]$. The objects of interest can be written as*

$$\underline{\beta} := \min_{\mu \in \mathcal{M}_\mathcal{S}} c(q)'\mu \qquad \overline{\beta} := \max_{\mu \in \mathcal{M}_\mathcal{S}} c(q)'\mu \tag{7}$$

Further, if $\mathcal{M}$ can be written as a system of linear inequalities in μ, both optimization problems are linear programs.

All proofs are in the appendix. Lemma 1 tells us that the set identified by moment conditions is convex for a fixed q, so it is sufficient to compute the upper and lower bound, which is a fact we need for sensitivity analysis later. This method did not require monotonicity as q is flexible. This method is generalizable to having more subgroups of interest, as long as we can write them in the form of the setup above.

Remark 3. *There are two special cases. (i) Equality $\underline{\beta} = \overline{\beta}$ holds when $q_A = q_N = 0$. The four distributions point identify $(\mu_{C1}, \mu_{C0}, \mu_{D1}, \mu_{D0})$. Namely, $\mu_{C0} = y_{C0}^{L(0)} = y_{C0}^{U(0)} = E[Y|T =$*

$0, Z = 0]$. *From Equation 2, this can only occur when* $p_{00} = p_{11}$, *and we set* $q_D = 1 - p_{00}$. *(ii) When monotonicity holds (i.e.,* q_D*=0),* $(\mu_{A1}, \mu_{C1}, \mu_{C0}, \mu_{N0})$ *are point-identified from the system in Equation 6. Setting* $q_D = 0$, q *is point-identified,* $\mu_{A1} = E[Y|T = 1, Z = 0]$, *and* $\mu_{N0} = E[Y|T = 0, Z = 1]$. *By substitution,*

$$\mu_{C1} = \frac{q_A + q_C}{q_C} E[Y|T = 1, Z = 1] - \frac{q_A}{q_C} E[Y|T = 1, Z = 0]$$
$$\mu_{C0} = \frac{q_N + q_C}{q_C} E[Y|T = 0, Z = 0] - \frac{q_N}{q_C} E[Y|T = 0, Z = 1]$$

Then, $\mu_{C1} - \mu_{C0}$ *is point-identified, and it will be identical to TSLS.*

Remark 4. *(Testable Implications). If the set* $\mathcal{M}_\mathcal{S}$ *is empty, then the model is misspecified: when* $\mathcal{M}$ *is unrestricted, it means that we cannot find* μ *that satisfies the constraints, which came from Assumption 1. This falsifies Assumption 1. A notable feature is that the setup here tests for exclusion and independence directly, whereas many existing tests do a joint test of those assumptions with monotonicity (or additive separability). This framework can also be used for other tests, including testing for selection bias.*[10]

2.3 Sensitivity Parameter

In the binary case, since monotonicity is equivalent to the absence of defiers, a logical sensitivity parameter controls the proportion of defiers. Consider $q_D/q_C \leq \lambda$, with λ being the sensitivity parameter, telling us the proportion of defiers we allow relative to the mass of compliers. By construction, $\lambda \in [0, 1]$ as $q_D \leq q_C$, so $\lambda = 0$ is where monotonicity holds, and $\lambda = 1$ places no restrictions on defiers.[11] This setup makes sensitivity comparable 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. Consequently, I define the set $\mathcal{Q}(\lambda)$ that is allowable for q to be:

$$\mathcal{Q}(\lambda) = \{q \in [0, 1]^4 : q_D \leq \lambda q_C \text{ and } q \text{ satisfies Equation 2}\}$$

Further, define:

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

This formulation allows us to compute a linear program in an inner loop while optimizing over q in the outer loop. With a well-behaved $\mathcal{M}_\mathcal{S}$, the sensitivity region is convex, the main result of this paper.

Theorem 1. *Suppose that* $\mathcal{M}$ *is convex and can be written as a system of linear inequalities in* μ. *Let* $c(q)$ *be continuous in* q. *Then, either* $\mathcal{M}_\mathcal{S}$ *is empty and hence* $\mathcal{B}(\lambda)$ *is empty, or the closure of* $\mathcal{B}(\lambda)$ *is equal to the interval* $[\underline{\beta}_\lambda, \overline{\beta}_\lambda]$, *where*

$$\underline{\beta}_\lambda := \min_{q \in \mathcal{Q}(\lambda)} \min_{\mu \in \mathcal{M}_\mathcal{S}} c(q)'\mu \qquad \overline{\beta}_\lambda := \max_{q \in \mathcal{Q}(\lambda)} \max_{\mu \in \mathcal{M}_\mathcal{S}} c(q)'\mu \tag{8}$$

[10] Test for selection bias corresponds to the quantity $E[Y_0|T = 1] - E[Y_0|T = 0]$. (see Section 4.3 of Mogstad et al. (2018)) This can be done by writing the objective function appropriately.

[11] Unless the instrument is irrelevant, the inequality cannot be binding at $\lambda = 1$. Since first stage is $FS = q_C - q_D$, using $\bar{q}_D$ to denote the maximum proportion of defiers, the largest λ that can yield a binding constraint is $\frac{\bar{q}_D}{FS + \bar{q}_D}$.

This notion of a sensitivity parameter generalizes to a non-binary setup. Defiers are always characterized for a pair of instrument values (z, z'). Thus, for every pair p of instrument values, we have a mass of defiers. Then, $\mathcal{Q}(\lambda_p) = \{q : \int_{g \in De(p)} d\nu(g) \leq \lambda_p \int_{g \in Co(p)} d\nu(g)\}$. To summarize this as a single sensitivity parameter, we can use λ such that $\lambda_p \leq \lambda$ for all p. Here, λ is the upper bound on defiers over all pairs of instrument values. When $\lambda = 0$, there does not exist any pair of instrument values for which there are defiers. This will be illustrated in Section 3.

In general, while bounds can be written in a linear program, they are not sharp. However, if we are interested in the means of a subgroup and we only have information on $Y|T, Z$ for a given (T, Z), the trimming bounds will be sharp, a result in Lee (2009). The problem hence lies in the fact that we only used information on the means across distributions, and we have not yet exploited all distributional information. To make sharp bounds, we should instead use restrictions on the entire distribution, which is an issue considered in Section 6. Nonetheless, if our outcome space is binary, the bounds obtained from the procedure and the sensitivity region will be sharp. This is formalized in Proposition 1.

Proposition 1. *If $Y \in \{0, 1\}$, bounds obtained in Equation 7 and Equation 8 are sharp.*

The proof of this proposition relies on using the fact that bounds from Balke and Pearl (1997) are sharp (since they exploit all distributional information), and that the parameterization in the method described can be rewritten as their problem and be mapped directly to their constraints. Sharpness in the binary case is what gives us equivalence to Balke and Pearl (1997) and Noack (2021) in special cases. The non-sharpness of the general program is similar to the generic framework presented in Mogstad et al. (2018) (henceforth MST), which did not allow for monotonicity violations. Hence, if we are interested in sensitivity analysis of PRTE in MST to monotonicity failure, the method is appropriate.

Remark 5. *A natural question to ask is how we might know what these q_D are. For sensitivity analysis, λ is simply an assumption that is made. I nonetheless provide two suggestions for how information on q_D can be obtained. (i) Use complementary studies: in Angrist and Evans (1998), we might use some social survey on gender preferences to get some idea of proportion of parents who prefer two girls or two boys. (ii) Use covariates: run the first-stage regression for every subset of covariates W. Since monotonicity requires all first stage coefficients (FS_W) to lie on either side of 0, the mass of observations that have FS_W lie on the opposite side of zero compared to the majority then constitute the defiers. However, in practice, if we do observe that FS_W lies on both sides of zero, the sensible thing to do is to condition on W to remove these violations. Getting q_D from the covariate method would be useful only when, in that particular context, the researcher is interested in the specification without these covariates.*

3 Policy Relevant Treatment Effects

This section discusses the coefficient vector $c(q)$ in the objective function for policy relevant treatment effects (PRTE), which is understood as the LATE in counterfacutal policy environments. The two subsections discuss PRTE where monotonicity fails, which is novel in literature. I loosely use "extrapolation" to refer to extending conclusions from a study to populations or environments that we do not have data for.

Recent literature in PRTE as in Carneiro et al. (2011) and MST 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 $Z^* = Z$ and $p(x,z) = p(x,z) + \alpha$ (ii) Proportional $1+\alpha$ change in propensity score with same instrument, so $Z^* = Z, p^*(x,z) = (1+\alpha)p(x,z)$ (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 changing the probability of being treated corresponds to (i) and (ii), so I group the first two together.

With monotonicity violations, we work with a nonseparable threshold-crossing selection model of the form $T = 1[\nu(Z,u) \geq 0]$, where u is the individual unobservable. We should think about the ν function theoretically because it is not identified. The counterfactual policies are thus:

1. Change the value of the instrument so $T^* = 1[\nu(Z^*,u) \geq 0]$ e.g., in an experiment where the instrument is randomly giving people incentives to take a particular action, and we gave people \$5 in the experiment, plan to give \$10 when the policy rolls out.
2. Change the threshold for everyone so $T_* = 1[\nu(Z,u) \geq -\alpha]$ e.g., subsidize childcare in the Angrist and Evans (1998) context: regardless of a couple's gender preference, the probability of having a third child increases.

For these two policy counterfactuals, I first discuss the mechanics, then discuss the target objects. Since the PRTE is the LATE in the counterfactual environment, this translates to identifying the treatment effect for a subgroup of the population. This is useful for assessing the robustness of conclusions on treatment effects.

3.1 Changing Instrument Value

Suppose we have a new instrument value Z^* such that, at this value, treatment is given by: $T^* = 1[\nu(Z^*,u) \geq 0]$. For illustration, I extrapolate the instrument rightward. The reasoning is similar if we wish to interpolate the instrument, or extrapolate leftward. In the original study, we have $Z \in \{0,1\}$, but now we have $Z^* = 2$. For every pre-existing group $G \in \{A,C,D,N\}$, individuals can have two possible responses at $Z^* = 2$. The extrapolated group is considered a defier as long as there is one pair of instrument values that go against the instrument. I summarize the classification in Table 3.

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, since N consists of Nt and Co1, and C consists of Co2 and De1, the distribution $Y|T=0,Z=0$ now contains a mixture of four groups Nt, Co1, Co2 and De1. The reasoning is analogous for other (T,Z) pairs, so the mixture of groups is as follows:

- T=0, Z=0: N,C: Nt, Co1, Co2, De1
- T=1, Z=0: D,A,: De4, At, De2, De3

Table 3: Last three columns refer to the original groups before we extrapolate rightward, leftward, or interpolate it.

Extrapolated Group	Response $T(Z)$	Right	Left	Interpolate
At	(1,1,1)	A	A	A
Nt	(0,0,0)	N	N	N
Co1	(0,0,1)	N	C	C
Co2	(0,1,1)	C	A	C
De1	(0,1,0)	C	D	N
De2	(1,0,0)	D	N	D
De3	(1,0,1)	D	C	A
De4	(1,1,0)	A	D	D

- T=0, Z=1: D,N: De2, De3, Nt, Co1
- T=1, Z=1: C,A: De4, At, De1, Co2

Then, potential outcome vector is now $\mu \in \mathbb{R}^{16}$ instead of $\mathbb{R}^8$. The mean condition analogous to Equation 6 can now be written as such:

$$\begin{aligned}
q_{Nt}\mu_{Nt,0} + q_{Co1}\mu_{Co1,0} + q_{Co2}\mu_{Co2,0} + q_{De1}\mu_{De1,0} &= p_{00}E[Y|T=0,Z=0] \\
q_{De4}\mu_{De4,1} + q_{At1}\mu_{At,1} + q_{De2}\mu_{De2,1} + q_{De3}\mu_{De3,1} &= p_{10}E[Y|T=1,Z=0] \\
q_{De2}\mu_{De2,0} + q_{De3}\mu_{De3,0} + q_{Nt}\mu_{Nt,0} + q_{Co1}\mu_{Co1,0} &= p_{01}E[Y|T=0,Z=1] \\
q_{De4}\mu_{De4,1} + q_{At}\mu_{At,1} + q_{De1}\mu_{De1,1} + q_{Co2}\mu_{Co2,1} &= p_{11}E[Y|T=1,Z=1]
\end{aligned} \tag{9}$$

Since PRTE is the $LATE^*$ in the counterfactual policy environment, it translates to the treatment effect for the "compliers" (TEC^*) in the counterfactual environment. 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 Co1 and Co2. Thus, $LATE^* = E[Y_1 - Y_0|Co1, Co2]$. However, it is not obvious why we would still be interested in compliers when monotonicity fails, since the treatment effect on compliers is simply what we get when using TSLS. Depending on the context, one may be interested in the TE for various subgroups of the population, where these subgroups are partitioned according to their response to the instrument.

Suppose we are nonetheless interested in $LATE^* = E[Y_1 - Y_0|Co1, Co2]$. The groups Co1 and Co2 come from the old C and N groups. Then, it is sufficient to specify two extrapolation parameters $Pr(Co1|N)$ and $Pr(Co2|C)$ and our sensitivity parameter λ.[12] We have 3 pairs of instrument values, so the inequalities defined by sensitivity parameters $\lambda_{01}, \lambda_{02}, \lambda_{12} \leq \lambda$ are:

$$\begin{aligned}
q_{De2} + q_{De3} &\leq \lambda_{01}(q_{Co2} + q_{De1}) \leq \lambda(q_{Co2} + q_{De1}) \\
q_{De2} + q_{De4} &\leq \lambda_{02}(q_{Co1} + q_{Co2}) \leq \lambda(q_{Co1} + q_{Co2}) \\
q_{De1} + q_{De4} &\leq \lambda_{12}(q_{Co1} + q_{De3}) \leq \lambda(q_{Co1} + q_{De3})
\end{aligned} \tag{10}$$

[12] It is unnecessary to specify the entire proportion vector, as explained in Remark 6. This is also illustrated in the empirical application.

Extrapolation is characterized by q, so the analysis here does not depend on the value of Z. Regardless of whether the new 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 $Pr(Co2|C)$ to be close to 1 and $Pr(Co1|N)$ close to zero. In contrast, with $Z = 100$, or a very different propensity score, it is analogous to a large extrapolation with $Pr(Co1|N)$ 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 6. *(Specifying Proportions). It is often unnecessary to specify the entire vector of defier proportions, because it does not matter what proportions we specify for groups that we are not interested in. For instance, when doing an interpolation and we are interested in TE for Co1, it is sufficient to specify q_D and q_{Co1}. q_{Co1} controls how different the policy environment is from what we have data for, and λ continues to be the sensitivity parameter for the presence of defiers. For all valid q_{At} chosen, the result will be the same, since only the identified distribution of the old complier group matters.*

A setting where we would be interested in extrapolation by using a different instrument value is when contracting or extending a policy that is the instrument. Another interesting setting is when the instrument is a randomly assigned incentive in an experiment. Extrapolation to different instrument values places bounds on what the outcomes of the experiment would have been were it designed differently (e.g., by assigning a different incentive intensity). This exercise is especially useful for checking the external validity of one's causal conclusions in these counterfactual environments.

3.2 Changing Threshold Crossing Rule

Suppose we have a policy that incentivises treatment for everyone, so the selection equation is now given by: $T_* = 1[\nu(Z, u) \geq -\alpha]$. Fix q_D as before. In our policy counterfactual, always-takers will still be always-takers. Compliers can remain compliers, or they can be come always takers when the policy is strong enough to shift their $Z = 0$ treatment to $T = 1$. Same argument can be made for defiers. The never-takers are the most interesting 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.

Use notation (G, Gc) to denote the subset of the population who are in group G in the original dataset and group Gc in this counterfactual policy. As before, we need additional assumptions on the proportion of each response group. The four observed distributions will now each be a mixture of the counterfactual groups:

- T=0, Z=0: N,C: CC, CA, NN, NC, ND, NA
- T=1, Z=0: D,A,: DD, DA, A
- T=0, Z=1: D,N: DD, DA, NN, NC, ND, NA
- T=1, Z=1: C,A: A, CC, CA

Mean restrictions are analogous to the discussion in the previous section. In this context, the $LATE^*$ is the treatment effect on the new group of compliers, which is expressed as $LATE^* = E[Y_1 - Y_0|CC, NC]$. As before, we can calculate the TE for each of the counterfactual groups, and these in themselves may be objects of interest. For instance, DA are those who have a gender preference for two boys or two girls, but their response are sensitive to childcare subsidies. We can then study the effect of having a third child on labor force participation for this subset of women.

Although we still have a single pair of instrument values in this setting, they occur under different policy environments. They are hence characterized by different sensitivity parameters. I use (0) to denote the original policy environment and (1) to denote the new policy environment.

$$\begin{aligned} q_{DD} + q_{DA} &\leq \lambda_{(0)}(q_{CC} + q_{CA}) \\ q_{DD} + q_{ND} &\leq \lambda_{(1)}(q_{CC} + q_{NC}) \end{aligned} \tag{11}$$

The mechanics for the two classes of policies are quite similar, which makes the bounds on some subgroups identical under an analogous setup. In the threshold-crossing environment, CC comes from the old complier group and $q_{CC} = Pr(CC|C)q_C$; in the changing instrument value environment, $Co1$ comes from the old complier group and $q_{Co1} = Pr(Co1|C)q_C$. Since they tap on the same complier group, under the same data-generating process in the original environment, Proposition 2 claims that the bounds on the TE of these extrapolated groups will be identical.

Proposition 2. *Given q_D and $Pr(CC|C) = Pr(Co1|C)$, bounds on $\mu_{CC1} - \mu_{CC0}$ are identical to bounds on $\mu_{Co1,1} - \mu_{Co1,0}$.*

The proof of this proposition comes from the observation that, when extrapolating, no restrictions are placed on these subgroups beyond trimming bounds. I emphasize that the PRTE problem can be addressed by the framework presented in Section 2.

Corollary 1. *PRTE that are written as the average TE for various subpopulations can be written as a linear program with a convex identified set.*

The corollary follows trivially from Lemma 1 and Theorem 1 since we can write the PRTE's as linear programs that follow the framework of Section 2. This did not require monotonicity, which allows us to do sensitivity analysis with respect to Mogstad et al. (2018).

While the threshold-crossing and instrument-value environments are mechanically identical for some bounds, they are conceptually different. In the (right-extrapolation) instrument-value setup, the uncertainty is in expected outcomes when $Z = 2$, while the data identifies expected outcomes for $Z = 0$ and $Z = 1$. In the threshold-crossing setup, the uncertainty is in the expected outcomes for both $Z = 0$ and $Z = 1$ under the new policy, which is not in the data. Even if we impose some monotonicity such that people treated in $Z = 1$ cannot be untreated in $Z = 2$ (thereby ruling out De1 and De4), we only reduce the problem to 6 response groups, which is distinct from the threshold-crossing problem where propensity score is uniformly increased, yielding 9 response groups. Depending on the application, one counterfactual environment is often more sensible than another.

Remark 7. *When monotonicity holds, the setup here simplifies to the class of policy environments discussed in Carneiro et al. (2011). In the threshold crossing rule with a separable selection equation,*

α denotes the increase in the probability of treatment, regardless of the instrument value assigned. Ruling out defiers in our context, having an increase in treatment probability for $Z = 0$ and $Z = 1$ respectively means:

$$q_{NA} + q_{CA} = \alpha$$
$$q_{NA} + q_{NC} = \alpha$$

When changing the instrument value without defiers, we only have four groups $(At, Nt, Co1, Co2)$. In the right-extrapolation exercise, changing the instrument value by α^* requires mapping the effect to some change in probability of treatment, say α. Then, $q_{Co2} = \alpha$.

4 Implementation

To implement this method on some given data, we can simply plug in the sample analog. The vector of μ's will be parameters that we optimize over. Let n denote the number of observations. An implementable algorithm follows:

1. Estimate probability objects p_{tz} by $\hat{p}_{tz} = \frac{\sum_{i=1}^{n} 1[T=t, Z=z]}{\sum_{i=1}^{n} 1[Z=z]}$. 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 $E[Y|T = t, Z = z]$.
2. For q_D assumed and $\hat{p}_{tz}$ calculated, calculate vector $\hat{q}$ using Equation 2.
3. Use the empirical cdf to calculate bounds in Inequality 5 and plug in $\hat{E}[Y|T = t, Z = z]$ and $\hat{q}$ into Equation 6.
4. Set up the objective function and solve the linear program in Equation 7.
5. When doing sensitivity analysis, create an outer loop as in Equation 8 using the sample analog.

For sensitivity analysis observe that condition $q_D \leq \lambda q_C$ can be written as $(1 - \lambda)q_D \leq \lambda(q_C - q_D)$, which is $q_D \leq \frac{\lambda}{1-\lambda} FS$, where $FS = q_C - q_D = Pr(T = 1|Z = 1) - Pr(T = 1|Z = 0)$. Since FS is obtained directly from the data, specifying λ gives us an upper bound on q_D. The q_D that optimizes the problem would either be $q_D = 0$, $q_D = \frac{\lambda}{1-\lambda} FS$, or some interior point. The linear program is cheap to implement, but the algorithm can be sped up even further when we restrict the search space. Depending on the specification, it may be possible to solve for the interior point, so we simply have to check three points to find the optimum q_D.

Parametric and further smoothness assumptions can also be imposed in $\mathcal{M}$ to get tighter bounds. Similar to MST, this is equivalent to making assumptions on $m_t(g)$ in Equation 1. Assumptions like monotonicity that rule out response groups impose restrictions on distribution $\nu(g)$. Assumptions on $m_t(g)$ can play a similar role by imposing that several groups have the same conditional expectation, which helps to reduce the dimension of the response group space. MST's finite linear basis in their unobserved u is one way to implement this. This maintains the linear program and implementation proceeds as before. Nonetheless, the rest of this paper proceeds without making these parameterizations.

Denote the estimate obtained from the sample $(\hat{\underline{\beta}}, \hat{\overline{\beta}})$ for the lower and upper bounds respectively for the problem in Equation 7. These estimates are consistent.

Theorem 2. *(Consistency). Suppose Assumption 1 holds. Then,* $\hat{\overline{\beta}} \xrightarrow{p} \overline{\beta}$ *and* $\hat{\underline{\beta}} \xrightarrow{p} \underline{\beta}$.

Consistency comes from the fact that the objective value is a continuous function of the distribution. We require data to be iid in Assumption 1 to apply the Glivenko-Cantelli theorem that gives convergence in the empirical cdf. Consistency then follows from continuous mapping theorem after proving continuity in the program.

I make some practical recommendations when implementing:

1. In most linear program solvers, a nonnegativity constraint is imposed on the solution vector μ, and I do the same here. This is reasonable in most empirical contexts when we consider binary outcomes or wages.
2. For some values of q_D, the linear program solver may state that the problem is infeasible. This is symptomatic that the defier value is too low, and increasing the q_D imposed would generally make the problem feasible. An infeasible problem at q_D is symptomatic of monotonicity violations (a testable implication).
3. The maximum q_D permitted in the data is $\min\{p_{01}, p_{10}\}$. This occurs when the defiers take up the entire $Y|T, Z$ distribution, and there are either no N or no A. Suppose there are no N. If we still require the upper and lower bounds of μ_{N0}, a reasonable thing to do is to use the max and min of distribution $Y|T = 0, Z = 1$ where only defiers are present.
4. Without a binary outcome space, it is often difficult to tell what the bounds K on the outcome space should be. A rule of thumb (ROT) would be to use the max and min of all outcomes Y observed.

Inference is a difficult problem for linear programs. In general, the nonparametric bootstrap will not have good coverage properties. For the problem at hand, it is possible to rewrite the constraints as a moment inequality model, which allows one to do inference using generalized moment selection. Details are in Appendix B, and plausible implementation is given in Fang et al. (2020).

5 Empirical Application

Going back to 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 we instrument with the first two kids being of the same sex (Z). All variables are thus 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 PUMS data for mothers using their posted dataset. The implementation follows their Table 5 where no additional covariates were included.

We have $n = 380007$ observations and the proportions are given by $\hat{P}(Z) = 0.504$, $\hat{P}(T = 1|Z = 1) = 0.402$ and $\hat{P}(T = 1|Z = 0) = 0.339$. Hence, the first stage, equivalently $q_C - q_D$, is 0.063.

Table 4: Summary of Results. TEC: Treatment Effect on Compliers; TEM: Treatment Effect on Marginal Population; ATE: Average Treatment Effect. PRTE: Policy-Relevant Treatment Effect in this context is $E[Y_1 - Y_0|CC, NC]$. PRTE use the following counterfactual environments: Policy Environment 1 has $Pr(CA|C) = 0, Pr(NC|N) = 0.1$; Policy Environment 2 has $Pr(CA|C) = 0.1, Pr(NC|N) = 0$; Policy Environment 3 has $Pr(CA|C) = 0.1, Pr(NC|N) = 0.1$. UB: Upper Bound; LB: Lower Bound.

	λ=0,LB	λ=0,UB	λ=0.2,LB	λ=0.2,UB	λ=0.5,LB	λ=0.5,UB
TEC	-0.08292	-0.08292	-0.26633	0.13367	-0.54146	0.45854
TEM	-0.08292	-0.08292	-0.38861	0.27806	-0.69431	0.63903
ATE	-0.56541	0.37153	-0.56541	0.37153	-0.56541	0.37153
PRTE1	-0.52940	0.44430	-0.57811	0.50182	-0.67808	0.61986
PRTE2	-0.20324	0.01898	-0.40704	0.25963	-0.71273	0.62060
PRTE3	-0.61213	0.52243	-0.67440	0.59345	-0.80478	0.74217

A summary of results is presented in Table 4, which displays the sensitivity of various objects of interest to varying extents that monotonicity is violated. In the simplest case, when $q_D = 0$, $TSLS = TEC = TEM$, so the lower and upper bounds are the same. This section discusses PRTE. The TEC and policy effects are relegated to Appendix D.

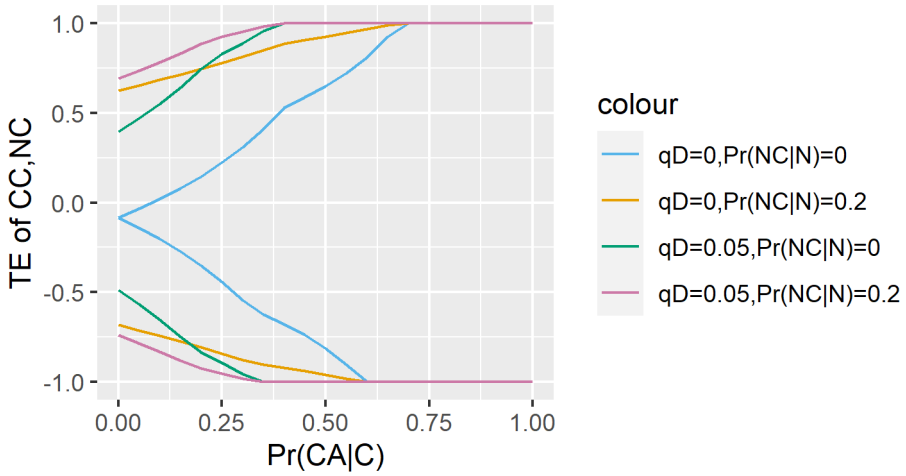
When considering counterfactual policy environments, the threshold shifting counterfactual is more sensible (e.g., childcare subsidy) than changing the instrument value. Then, PRTE is TE for new compliers, which are groups $\{CC, NC\}$. Table 4 considers three possible counterfactual policy environments. PRTE1 has $Pr(CA|C) = 0, Pr(NC|N) = 0.1$, which means that the childcare subsidy incentivized 10% of the never-takers to become compliers (i.e., have a third child if and only if the first two are of the same gender) while none of the existing compliers becomes always-takers. Similarly, PRTE2 with $Pr(CA|C) = 0.1, Pr(NC|N) = 0$ incentivized the C group to be A, but not the N group, and PRTE3 with $Pr(CA|C) = 0.1, Pr(NC|N) = 0.1$ incentivized both groups.

We can calculate the TE for each of the counterfactual groups, and these may be objects of interest. Namely, NC are those who have a weak preference for gender balance. In the original dataset, they would never have a third child, but once subsidies are available, they have a third child when the first two kids are of the same sex. $E[Y_1 - Y_0|NC]$ is effect of third child for this subgroup. CC are those who have a strong preference for gender balance, such that childcare subsidies did not incentivize them to have a third kid when their first two kids are of the same sex. Figure 1 plots bounds for $E[Y_1 - Y_0|CC, NC]$, which is a mixture of the old C and N groups.

Figure 1 plots the PRTE against $Pr(CA|C)$ to show how bounds change when the extent of extrapolation increases i.e., when the policy environment that we are interested in becomes more different from what we had data for. It reflects how assumptions on q_{CA}, q_{NC} and q_D play a role. Since $q_{CA} = Pr(CA|C)q_C$, making an assumption on $Pr(CA|C)$ is sufficient when we have q_D. Old compliers can only be C or A in the new policy environment, so $q_{CC} = Pr(CC|C)q_C = (1 - Pr(CA|C))q_C$.

We essentially deal with 6 response groups here: (A, CA, CC, D, NC, NC'), where NC' denotes the set of groups that switch from N to anything but a complier in the new policy environment.

Figure 1: Plot of $PRTE = E[Y_1 - Y_0|CC, NC]$ bounds against $Pr(CA|C)$ for various q_D imposed.



This occurs despite having 9 response groups in Section 3 because the mixture of DA relative to DD does not matter in the extrapolation. Similarly, the proportions of various groups in NC' is irrelevant to the object of interest. This means that we can choose q_{DD}, q_{DA} to be whatever we want as long as they are nonnegative and $q_{DD} + q_{DA} = q_D$. Consequently, it is sufficient to specify $(q_D, Pr(CA|C), Pr(NC|N))$ to solve the linear program.

For sensitivity analysis, objects $(Pr(CA|C), Pr(NC|N))$ are treated as parameters or assumptions on the policy environment. Instead of specifying q_D, we can specify the sensitivity parameter λ which satisfies $\lambda \geq \lambda_{(0)}$ and $\lambda \geq \lambda_{(1)}$ in Equation 11. Since we can innocuously set $q_{DD} = q_{ND} = 0$, $\lambda = \lambda_{(0)}$ and only the first inequality is binding. This will give us the optimum, because the objective value has to perform weakly better with one fewer constraint.

Consider the environment where monotonicity holds in Figure 1, so $q_D = 0$. When $Pr(NC|N) = 0$, $E[Y_1 - Y_0|CC, NC] = E[Y_1 - Y_0|CC]$, as all the weight on $E[Y_1 - Y_0|CC, NC]$ comes from the CC group. The blue line thus reflects the effect of $Pr(CC|C)$ on bounds of $\mu_{CC1} - \mu_{CC0}$ when monotonicity holds. The CC group comes from the old C group who would be split across CC and CA. When $Pr(CA|C) = 0$, the bounds are identical to TEC in the original formulation (i.e., TSLS under monotonicity). Since the TEC is a mixture between CC and CA, as the proportion of CA increases, the gap between the largest and smallest bounds widens. It is also evident that after $Pr(CA|C)$ increases above a certain point for a given q_D, the bounds become trivial.

Still working with the monotonicity assumption, consider the orange line where $Pr(NC|N) = 0.2$. $E[Y_1 - Y_0|CC, NC]$ is now the weighted TE of CC and NC. Since the NC group originally came from the never-takers, the trivial (-1,1) bounds are used for the NC group in this mixture. Consequently, the bounds are much wider than the blue case with $Pr(NC|N) = 0$.

Consider the case where monotonicity is violated for PRTE. The green line in Figure 1 sets $Pr(NC|N) = 0$, so the bounds are for $\mu_{CC1} - \mu_{CC0}$. Bounds at $Pr(CA|C) = 0$ are identical to TEC, so it matches the bounds on TEC when $q_D = 0.05$. For every $Pr(CA|C)$ value assumed (which corresponds to a different counterfactual environment), we can compare this green line with the original blue line with $q_D = 0, Pr(NC|N) = 0$ to observe how failure of monotonicity affects the bounds. Similarly, the magenta line with $q_D = 0.05, Pr(NC|N) = 0.2$ is the analogous plot

to the orange $q_D = 0, Pr(NC|N) = 0.2$, allowing us to analyze the sensitivity of the bounds to failures of monotonicity.

In light of trivial bounds under some proportion assumptions, one might consider when data becomes uninformative. Since bounds on TE of CC is similar to bounds on TE of $Co1$, I focus on interpolation and provide sufficient conditions for getting trivial bounds for $E[Y_1 - Y_0|Co1]$ without loss of generality in the policy regime. Recall that the old C group splits into Co1 and Co2, so a sufficient statistic for the proportions is $q_{Co1|C} = Pr(Co1|C)$.

Proposition 3 states the result for getting an uninformative upper bound for binary outcomes, and the result for the lower bound is analogous. Namely, it tells us the proportions required to obtain trivial bounds. Proportion $q_{Co1|C}$ being smaller than a particular value is analogous to $q_{Co2|C}$ being larger than some value. The latter is analogous to $Pr(CA|C)$ being larger than some value, yielding the result observed in Figure 1.

Let $\mu^U_{C1}(q_D)$ denote the largest achievable μ_{C1} for the given q_D and $\mu^L_{C0}(q_D)$ denote the smallest achievable μ_{C0}.

Proposition 3. *(Uninformative Bounds). Suppose we have a binary outcome and $q_{Co1|C} \neq 0$. When interpolating, consider conditions:*

(i) Either $q_{Co1|C} \leq \mu^U_{C1}(q_D)$ or $\mu^U_{C1}(q_D) = 1$.

(ii) Either $q_{Co1|C} \leq 1 - \mu^L_{C0}(q_D)$ or $\mu^L_{C0}(q_D) = 0$.

If (i) and (ii) hold, then $\max_{\mu \in \mathcal{S}} \mu_{Co1,1} - \mu_{Co1,0} = 1$.

The proof relies on two observations. First, since $\mu^U_{C1}(q_D) = 1$ is a convex combination $\mu_{Co1,1}$ and $\mu_{Co2,1}$, both of them must also be one. Second, the other group $Co2$ does not face further restrictions, so we can make $\mu_{Co2,1} = 0$ to get a large $\mu_{Co1,1}$.

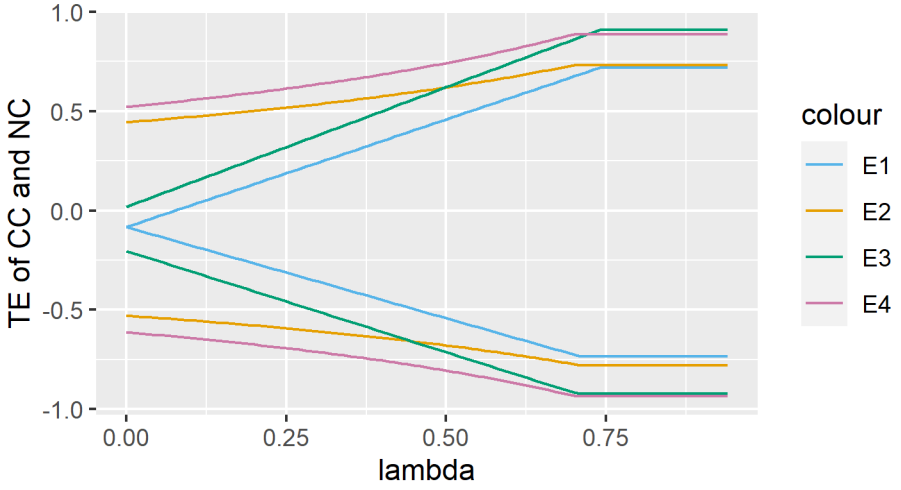
Figure 2 is more useful for sensitivity analysis. Environment E1 is the setting where we consider the original policy environment (i.e., no extrapolation), matching Figure 4 exactly. When imposing $Pr(NC|N) = 0$ in Environment E3, the bounds are linear, because we are placing bounds on the TE of a subpopulation of the original complier distribution. The properties of the original TEC (e.g., bounds are linear in λ) will still hold, and wider bounds are obtained from having a smaller subpopulation. Bounds are much wider when q_{NC} is nonzero, because we have never-takers in the mixture and worst-case bounds are imposed for μ_{NC1}. Linearity in λ is also lost in the presence of the NC group.

6 Extensions

In this section, introduce notation h where h_{g,y_0,y_1} denotes the proportion of the population that is group g, with potential outcomes y_t for the various treatment values $t \in \{0, 1\}$.

$$h_{g,yy'} := Pr(G = g, Y_0 = y, Y_1 = y')$$

Figure 2: Plot of $PRTE = E[Y_1 - Y_0|CC, NC]$ bounds against λ. Policy Environment E1 has $Pr(CA|C) = 0, Pr(NC|N) = 0$; Policy Environment E2 has $Pr(CA|C) = 0, Pr(NC|N) = 0.1$; Policy Environment E3 has $Pr(CA|C) = 0.1, Pr(NC|N) = 0$; Policy Environment E4 has $Pr(CA|C) = 0.1, Pr(NC|N) = 0.1$.



For discrete Y,

$$\sum_{g,y_0,y_1} h_{g,y_0,y_1} = 1$$

For continuous Y, let $H_{gt}(y)$ denote the proportion of population that is in group g with potential outcome up to y when given treatment t:

$$H_{g0}(y) = \int_{-\infty}^{y} \int_{-\infty}^{\infty} h_{g,y_0,y_1} dy_1 dy_0$$
$$H_{g0}(\infty) = h_g = q_g = H_{g1}(\infty)$$

In the binary representation, where $f_{T=t|Z=z}(y)$ denotes the observed $Pr(T = t, Y = y|Z = z)$,

$$f_{T=1|Z=1}(y) = \sum_j (h_{Ajy} + h_{Cjy})$$
$$1 = \sum_y (f_{T=1|Z=1}(y) + f_{T=0|Z=1}(y))$$

6.1 Sharp Bounds

It is shown in this section that sharp bounds can be obtained when the outcome space is finite and discrete. This is done by parameterizing the problem in Equation 7 in terms of point masses h instead of q and μ. For continuous outcomes, the problem becomes infinite-dimensional. Parameterizing the problem using the distribution allows sharp bounds, but this occurs at the cost of not having a linear objective function.

Using the h notation, (q, μ) can be expressed as:

$$\mu_{g0} = \frac{\sum_{y,y'} y h_{g,yy'}}{\sum_{y,y'} h_{g,yy'}} \qquad \mu_{g1} = \frac{\sum_{y,y'} y' h_{g,yy'}}{\sum_{y,y'} h_{g,yy'}}$$
$$q_g = \sum_{y,y'} h_{g,yy'}$$

For binary Y, $yy' \in \{00, 01, 10, 11\}$. We have 16 variables. Law of total probability implies $\sum_{g\in\mathcal{G}} \sum_{y,y'} h_{g,yy'} = 1$, and we require $h_{g,yy'} \in [0, 1]$.

In some cases, the object of interest (objective function) can be expressed as a linear combination of these point masses i.e., $c'h$, where h is a vector that stacks the point masses. For instance, the ATE $E[Y_1 - Y_0]$ can be expressed as

$$ATE = h_{A,01} + h_{C,01} + h_{D,01} + h_{N,01} - h_{A,10} - h_{C,10} - h_{D,10} - h_{N,10} \tag{12}$$

However, if we condition on any particular subgroup, then the objective is no longer linear in h, with $E[Y_1 - Y_0|G] = \frac{h_{G01} - h_{G10}}{q_G}$. This will be the case when we want the LATE or various PRTE. Hence, parameterizing the problem in terms of point masses gives sharp bounds at the cost of losing the linear program.

Constraints are given by the fact that the mass of the variables in each cell must add up to the observable primitives. For some discrete outcome space, with binary (T,Z), we have $2 \times 2 \times |\mathcal{Y}|$ constraints and $4 \times |\mathcal{Y}|^2$ parameters for a discrete support. Then, $\forall y \in \mathcal{Y}$, the mean probability restrictions are:

$$\begin{aligned} f_{T=1|Z=1}(y) &= \sum_j (h_{Ajy} + h_{Cjy}) \\ f_{T=1|Z=0}(y) &= \sum_j (h_{Ajy} + h_{Djy}) \\ f_{T=0|Z=1}(y) &= \sum_j (h_{Nyj} + h_{Dyj}) \\ f_{T=0|Z=0}(y) &= \sum_j (h_{Nyj} + h_{Cyj}) \end{aligned} \tag{13}$$

In the binary case,

$$\begin{aligned}
Pr(Y=0,T=0|Z=0) &= h_{N,00}+h_{N,01}+h_{C,00}+h_{C,01}\\
Pr(Y=0,T=1|Z=0) &= h_{D,00}+h_{D,10}+h_{A,00}+h_{A,10}\\
Pr(Y=1,T=0|Z=0) &= h_{N,10}+h_{N,11}+h_{C,10}+h_{C,11}\\
Pr(Y=1,T=1|Z=0) &= h_{D,01}+h_{D,11}+h_{A,01}+h_{A,11}\\
Pr(Y=0,T=0|Z=1) &= h_{N,00}+h_{N,01}+h_{D,00}+h_{D,01}\\
Pr(Y=0,T=1|Z=1) &= h_{C,00}+h_{C,10}+h_{A,00}+h_{A,10}\\
Pr(Y=1,T=0|Z=1) &= h_{N,10}+h_{N,11}+h_{D,10}+h_{D,11}\\
Pr(Y=1,T=1|Z=1) &= h_{C,01}+h_{C,11}+h_{A,01}+h_{A,11}
\end{aligned} \tag{14}$$

In addition, the h's must be in $[0,1]$. The problem with parameterizing the problem in this way is that the number of parameters grows exponentially as the outcome space grows. If we have a continuous outcome, we will have an infinite-dimensional parameter space.

Remark 8. *(Convexity of Sensitivity Region). With discrete outcomes, it is straightforward to see that when we have a convex set for h written in terms of linear equalities and inequalities, Equation 8 provides sharp bounds to the sensitivity region. This works even for a generic setup with multiple parameters $\lambda_p \leq \lambda$ for all p, when $\mathcal{Q}$ can be expressed as as a convex set of inequalities in h.*

An equivalent way to do this will be to use the cumulative distribution function that is amenable to continuous y. Using F to denote the cdf of the observed distribution and H to denote the cdf of the potential outcomes of various subgroups, the distributional information can be written as such:

$$\begin{aligned}
F_{T=0|Z=0}(y) &= H_{N0}(y)+H_{C0}(y)\\
F_{T=1|Z=0}(y) &= H_{A1}(y)+H_{D1}(y)\\
F_{T=0|Z=1}(y) &= H_{N0}(y)+H_{D0}(y)\\
F_{T=1|Z=1}(y) &= H_{A1}(y)+H_{C1}(y)
\end{aligned} \tag{15}$$

Proposition 4. *(Sufficiency). Parameters (q,H) in Equations 2 and 15 are sufficient to characterize all distributional information.*

Note that the cdf can also be written as a linear combination of point masses. In particular, $H_{gt}(y) = \int_{-\infty}^{y} h_{gt}(v)dv$. This preserves linearity in the problem, though the parameter space is still infinite-dimensional. Use kernel methods or basis functions because the problem is reduced entirely to proportions and conditional means.

Corollary 2. *Bounds obtained by parameterization of (q,H) are sharp.*

Since we have used all available information in the data in the constraint set, the linear program delivers sharp bounds. Consequently, even without binary outcomes, we can get Noack (2021) bounds for LATE, which are sharp. Running the ATE program gives bounds in Kitagawa (2021). Further, the parameterization in Equation 15 is analogous to Equation (3) in Kitagawa (2021), which is written in terms of the probability density function.

Table 5: Observe column indicates the observed response given instrument; group column indicates the groups present within the treatment-instrument pair; monotonicity column indicates groups present when there are no defiers.

Observe	Groups	Monotonicity
$T_1 = 0$	Nt, Co1, De1, Co2	Nt, Co1, Co2
$T_1 = 1$	De2, De3, De4, At	At
$T_2 = 0$	Nt, Co1, De2, De3	Nt, Co1
$T_2 = 1$	De1, Co2, De4, At	At, Co2
$T_3 = 0$	Nt, De1, De2, De4	Nt
$T_3 = 1$	Co1, Co2, De3, At	At1, Co1, Co2

Remark 9. *(Quantiles). Since the problem can be parameterized in terms of the cdf of various groups, it follows that we can solve for bounds on quantiles of various subgroups similarly using a linear program. The objective function is a linear function of the cdf that parameterizes various groups.*

6.2 Multivalued Instrument and Treatment

Response groups can be indexed by $G = (T_0, T_1, \cdots)$, where T_z denotes potential treatment when given instrument z. Hence, with $T \in \mathcal{T}$ and $Z \in \mathcal{Z}$, the number of response types is $|\mathcal{T}|^{|\mathcal{Z}|}$. Then, conditional means can be expressed as:

$$\sum_{g \in \mathcal{G}(T_z = t)} q_g \mu_{gt} = \left(\sum_{g \in \mathcal{G}(T_z = t)} q_g \right) E[Y|T = t, Z = z] \tag{16}$$

where $\mathcal{G}(T_z = t)$ denotes the set of groups groups with $T_z = t$. This is illustrated in two environments - one with nonbinary instrument and another with nonbinary treatment.

Consider the environment where T is binary, but $Z \in \{0, 1, 2\}$ takes 3 discrete values, and propensity score increases with Z. Denote the response using the triple (T_1, T_2, T_3), so there are 8 possible response groups. Under monotonicity, only four such responses are valid: (0,0,0) are never takers (Ne), (1,1,1) are always takers (At), and (0,0,1) and (0,1,1) are complier groups Co1 and Co2 respectively. We then have 4 defier groups because there is at least a pair of instruments where they would constitute defiers. Label (0,1,0), (1,0,0), (1,0,1), (1,1,0) as De1, De2, De3, De4. This is summarized in Table 5.

With binary instruments, knowing q_D gave us the proportion of all other groups. The analog here is that we assume q_{DeX} is known for X 1 to 4. To see this, the proportion constraints are given by:

$$\begin{aligned} p_{00} &= q_{Nt} + q_{Co1} + q_{Co2} + q_{De1} \\ p_{01} &= q_{De2} + q_{De3} + q_{Nt} + q_{Co1} \\ p_{02} &= q_{Nt} + q_{De1} + q_{De2} + q_{De4} \end{aligned}$$

Since the probabilities sum to one, proportion constraints on $1 - p_{1z}$ are redundant. We thus have 4 linearly independent constraints available and 4 objects free where proportions are concerned. Hence, in the three-value instrument setup, we need to specify the entire vector of defier proportions.

This reasoning generalizes to more instrument values. If there are J instrument values, then we will have $J + 1$ equations and 2^J unknowns. There is 1 At group, 1 Nt group, $(J - 1)$ complier groups, and $2^J - (J + 1)$ defier groups. Hence $2^J - (J - 1)$ restrictions on defiers are required to identify q. In many applications, further restrictions can reasonably be imposed, usually through weaker alternative monotonicity assumptions: for instance, unordered monotonicity from Heckman and Pinto (2018) reduces the number of defier groups to $(J - 1)$.

Proposition 5. *With J instrument values, $2^J - (J+1)$ restrictions on proportion of defier groups are required to identify q.*

Unlike extrapolation, we do observe outcomes from $Z = 2$ here. Trimming bounds are analogous. Namely, if we have R groups in the distribution, with $r \in \{g_1, g_2, \cdots, g_R\}$. Then, the max of μ_{g_1} is the mean of the top $\frac{Pr(g_1)}{\sum_r Pr(g_r)}$ mass of the distribution. In addition to the mean restrictions in Equation (1) of this note, we have two more restrictions:

$$\begin{aligned} \frac{q_{Nt}\mu_{Nt,0} + q_{De1}\mu_{De1,0} + q_{De2}\mu_{De2,0} + q_{De4}\mu_{De4,0}}{q_{Nt} + q_{De1} + q_{De2} + q_{De4}} &= E[Y|T = 0, Z = 2] \\ \frac{q_{At}\mu_{At,1} + q_{Co1}\mu_{Co1,1} + q_{Co2}\mu_{Co2,1} + q_{De3}\mu_{De3,1}}{q_{At} + q_{Co1} + q_{Co2} + q_{De3}} &= E[Y|T = 1, Z = 2] \end{aligned} \tag{17}$$

The mean restrictions when we have more instrument values will be analogous, so the system is easy to generalize.

Next, consider case where $t \in \{0, 1, 2\} = \mathcal{T}$ and $z \in \{0, 1\} = \mathcal{Z}$, so there are 9 groups. Conditional mean parameters are $\mu_{G,t}$ and we have $|\mathcal{T}|^{|\mathcal{Z}|}|\mathcal{Z}| = 27$ of them. Moment conditions denoting the mixture of distributions are hence:

$$\begin{aligned} q_{(0,0)}\mu_{(0,0),0} + q_{(0,1)}\mu_{(0,1),0} + q_{(0,2)}\mu_{(0,2),0} &= (q_{(0,0)} + q_{(0,1)} + q_{(0,2)})E[Y|T = 0, Z = 0] \\ q_{(1,0)}\mu_{(1,0),1} + q_{(1,1)}\mu_{(1,1),1} + q_{(1,2)}\mu_{(1,2),1} &= (q_{(1,0)} + q_{(1,1)} + q_{(1,2)})E[Y|T = 1, Z = 0] \\ q_{(2,0)}\mu_{(2,0),2} + q_{(2,1)}\mu_{(2,1),2} + q_{(2,2)}\mu_{(2,2),2} &= (q_{(2,0)} + q_{(2,1)} + q_{(2,2)})E[Y|T = 2, Z = 0] \\ q_{(0,0)}\mu_{(0,0),0} + q_{(1,0)}\mu_{(1,0),0} + q_{(2,0)}\mu_{(2,0),0} &= (q_{(0,0)} + q_{(1,0)} + q_{(2,0)})E[Y|T = 0, Z = 1] \\ q_{(0,1)}\mu_{(0,1),1} + q_{(1,1)}\mu_{(1,1),1} + q_{(2,1)}\mu_{(2,1),1} &= (q_{(0,1)} + q_{(1,1)} + q_{(2,1)})E[Y|T = 1, Z = 1] \\ q_{(0,2)}\mu_{(0,2),2} + q_{(1,2)}\mu_{(1,2),2} + q_{(2,2)}\mu_{(2,2),2} &= (q_{(0,2)} + q_{(1,2)} + q_{(2,2)})E[Y|T = 2, Z = 1] \end{aligned}$$

The six TE's that are point identified as thus $\Delta\mu_{(0,1)}$, $\Delta\mu_{(0,2)}$, $\Delta\mu_{(1,0)}$, $\Delta\mu_{(1,2)}$, $\Delta\mu_{(2,0)}$, and $\Delta\mu_{(2,1)}$. The mechanism in the framework generalizes straightforwardly.

To obtain sharp bounds we can parameterize the distribution instead. Index the treatment response as g as before, and index the potential outcomes as $o = (Y_0, Y_1, \cdots)$. Then, the joint distribution

can be written as:

$$f(y,t,z) = \int_o \int_g h_{g,o} d\mathcal{G}(T_z = t) d\mathcal{Y}(Y_t = y)$$

Here, $\mathcal{G}(T_z = t)$ denotes set of groups with $T_z = t$, and $\mathcal{Y}(Y_t = y)$ denotes set of groups with $Y_t = y$. We might settle for moment conditions that are non-sharp, so we don't need to solve an infinite-dimensional problem. As before, the advantage of sharp bounds by parameterizing the problem using the distribution comes as the cost of losing linearity for some objective functions.

7 Conclusion

This paper showed how one can use known monotonicity violations to partially identify objects of interest, including (local) average treatment effects (LATE) and policy-relevant treatment effects (PRTE) by writing them in the form of a general-purpose linear program. An interpretation for PRTE without the MTE framework is also provided. Identification uses assumptions on proportions of the population that have a particular response to the instrument instead of assumptions on the outcome function. Nontrivial bounds for objects of interest in counterfactual policies can be obtained without monotonicity. Sensitivity analysis of PRTE can be conducted by treating the proportion of defiers relative to compliers as a sensitivity parameter. The method can be extended to nonbinary instruments and to obtain sharp bounds for continuous outcomes.

The method presented is useful in many econometric tasks and empirical applications, including those suggested in Mogstad et al. (2018) and more. Econometric applications include: (1) sensitivity analysis of conventional treatment parameters (e.g., LATE, ATT) to monotonicity failure (2) evaluation of the robustness of a study's conclusions to different policy environments and its sensitivity to the presence of defiers (3) testing for independence and exclusion without using a joint test of these assumptions with monotonicity. The threshold crossing extrapolation is useful in environments when the instrument is based on natural experiments (such as gender composition) so individuals affected there are likely different from those affected by a relevant policy. Extrapolation by changing instrument value is useful when the instrument represents a known policy change, but the researcher is interested in extending or contracting the policy. Further, it helps place bounds on what the outcome of an experiment would have been should it have been designed differently.

A Proof of Results

Proof of Lemma 1. For some convex $\mathcal{M}$, $\mathcal{M}_\mathcal{S}$ is either empty or nonempty. If $\mathcal{M}_\mathcal{S}$ is empty, then by definition $\mathcal{B} = \emptyset$. Next consider a nonempty $\mathcal{M}_\mathcal{S}$. Since $\mathcal{M}$ is convex and the set of μ that satisfies $\mathcal{S}$ is convex, and the intersection of convex sets is also convex, $\mathcal{S}$ is convex. Since a linear mapping of a convex set also yields a convex set, $\mathcal{B}$ is a convex set. The identified set of β thus lies within the upper and lower bound described.

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. □

Proof of Theorem 1. Proof for an empty $\mathcal{B}$ is identical to the proof in Lemma 1. Since $c(q)$ is continuous in q, $c(q)'\mu$ is continuous in q. Using Theorem 2 from Wets (1985), the linear program is continuous, so $\underline{\beta}(q)$ and $\overline{\beta}(q)$ are continuous in q, where they are solutions to Equation 7.

It is sufficient to show that for any $b \in [\underline{\beta}_\lambda, \overline{\beta}_\lambda]$ is in the identified set of valid q. Let q_0 denote the q value where monotonicity holds, so $q_D = 0$. If $b \in [\underline{\beta}(q_0), \overline{\beta}(q_0)]$, then it is in an identified set for some q. Let $\overline{q}$ be the q whose maximum is $\overline{\beta}_\lambda$. If $b \in [\overline{\beta}(q_0), \overline{\beta}_\lambda]$, by intermediate value theorem and continuity of the program, there must exist some $q^* \in [q_0, \overline{q}]$ such that $\overline{\beta}(q^*) = b$. Hence, b is feasible for some valid q. An analogous argument can be made for $b \in [\underline{\beta}_\lambda, \underline{\beta}(q_0)]$. □

Proof of Proposition 1. I first set up the Balke and Pearl (1997) (henceforth BP) problem, then show how my formulation is equivalent to theirs. Since the BP bounds are sharp, it follows that the formulation in Equation 7 is also sharp.

BP have 8 equality constraints and several inequality constraints corresponding to a valid probability masses. Use notation in Section 6.1 and that $p_{yt.z} := Pr(Y = y, T = t|Z = z)$. Writing our mean and proportion parameters in terms of the BP notation:

$$\begin{aligned}
\mu_{C1} &= \frac{1}{q_C}(h_{C01} + h_{C11}) \\
\mu_{C0} &= \frac{1}{q_C}(h_{C10} + h_{C11}) \\
\mu_{C1} - \mu_{C0} &= \frac{1}{q_C}(h_{C01} - h_{C10}) \\
q_C &= h_{C00} + h_{C01} + h_{C10} + h_{C11}
\end{aligned}$$

Mean restrictions can thus be written as:

$$\begin{aligned}
q_A\mu_{A1} + q_C\mu_{C1} &= (q_A + q_C)E[Y|T = 1, Z = 0] \\
\Rightarrow h_{A01} + h_{A11} + h_{C01} + h_{C11} &= (q_A + q_C)\frac{p_{11.1}}{p_{11.1} + p_{01.1}}
\end{aligned}$$

Since $q_A + q_C = p_{11.1} + p_{01.1}$,

$$h_{A01} + h_{A11} + h_{C01} + h_{C11} = p_{11.1}$$

This is one of the eight BP constraints. We can use an analogous argument for three other constraints, so the mean restrictions are identical to 4 out of 8 BP equality constraints.

In the (A, C) combination, the BP constraint $p_{01.1} = h_{C,00} + h_{C,10} + h_{A,00} + h_{A,10}$ is not directly analogous to the mean constraint worked out above. Note that $p_{01.1} = q_A + q_C - p_{11.1}$ because it contains the (A,C) probability masses that do not feature in $p_{11.1}$. But when q_C and q_A known, substitution of various terms into the expression yields the $p_{01.1}$ equation from BP, so no new equation is contained here. This implies that the combination of proportion restrictions in Equation 2 and the mean restrictions are sufficient to obtain the four remaining BP equality constraints.

Similarly, the trimming bounds in Inequality 5, proportion constraints in Equality 2, and $\mathcal{M} = [0, 1]^8$ map to the BP inequalities.

Since the formulation in Equation 7 is sharp for all q, the bounds in Equation 8 must also be achievable for some distribution where $q \in \mathcal{Q}(\lambda)$. By convexity in Theorem 1, the bounds in Equation 8 are sharp. □

Proof of Proposition 2. Compliers can only go into CC or CA in the first setting and Co1 or Co2 in the second setting. Both initial problems (before extrapolation) are identical, so we get the same identified distribution of $\mu_{C1} - \mu_{C0}$. Both $\mu_{Co1,1} - \mu_{Co1,0}$ and $\mu_{CC1} - \mu_{CC0}$ have the same trimming bounds, as no other restrictions are placed on them. Their bounds must hence be the same. □

Proof of Corollary 1. Follows from Lemma 1, Theorem 1 and how PRTE can be written in the form of a linear program. □

Proof of Theorem 2. Let F_n denote the empirical joint cdf. Using Proposition 3.6.6 of Giné and Nickl (2021), the class of all half-spaces of $\mathbb{R}^d$ is a Vapnik-Cervonenkes (VC) class. The family of indicator functions of sets in VC class is a VC subgraph, so the condition for Glivenko-Cantelli theorem is satisfied. Applying Corollary 3.7.17 in Giné and Nickl (2021), $F_n \xrightarrow{a.s.} F$, where F is the asymptotic joint distribution of (Y, T, Z). This implies $F_n \xrightarrow{p} F$ since almost sure convergence implies convergence in probability. Consider the $\hat{\overline{\beta}} \xrightarrow{p} \overline{\beta}$ problem, since the lower bound is analogous.

$$\hat{\overline{\beta}}(F_n) = \max_{\mu \in \mathcal{S}(F_n)} c(q(F_n))'\mu$$

Using our parameterization, observe that the constraints are continuous. Thus, by continuous mapping theorem, $\mathcal{S}(F_n) \xrightarrow{p} \mathcal{S}(F)$ and $q(F_n) \xrightarrow{p} q(F)$. Further, c is a continuous function of q and a composition of continuous functions is also continuous, so $c(q(F))$ is a continuous function in F. Thus, both the constraint set and the objective function converges in probability.

Finally, apply Theorem 2 from Wets (1985) that the linear program is continuous in its hyperparameters. By continuous mapping theorem, $\overline{\beta}(F_n) \xrightarrow{p} \overline{\beta}(F)$. □

Proof of Proposition 3. $\mu_{Co1,1} - \mu_{Co1,0} = 1$ is obtained when $\mu_{Co1,1} = 1$ and $\mu_{Co1,0} = 0$. By construction, we require

$$q_{Co1|C}\mu_{Co1,1} + q_{Co2|C}\mu_{Co2,1} = \mu^U_{C1}(q_D)$$

Either $\mu^U_{C1}(q_D) = 1$ or $q_{Co1|C} \leq \mu^U_{C1}(q_D)$ will yield $\mu_{Co1,1} = 1$.

If $\mu^U_{C1}(q_D) = 1$, then $q_{Co1|C}\mu_{Co1,1} + (1 - q_{Co1|C})\mu_{Co2,1} = 1$. Further, $\mu_{Co1,1}, \mu_{Co2,1} \in [0, 1]$ for binary outcomes. If $\mu_{Co1,1} < 1$, then for a convex combination of $\mu_{Co1,1}$ and $\mu_{Co2,1}$ to be 1, we require $\mu_{Co2,1} > 1$, a contradiction.

If $q_{Co1|C} = \mu^U_{C1}(q_D)$, then $\mu_{Co1,1}\mu^U_{C1}(q_D) = \mu^U_{C1}(q_D)$ when setting $\mu_{Co2,1} = 0$. Next consider $q_{Co1|C} < \mu^U_{C1}(q_D)$. Rewriting the mean restriction yields $\mu_{Co1,1} = \frac{1}{q_{Co1|C}}(\mu^U_{C1}(q_D) - q_{Co2}\mu_{Co2,1}$, so $\frac{\mu^U_{C1}(q_D)}{q_{Co1|C}} > 1$, and we can choose some feasible $\mu_{Co2,1} > 0$ to make $\mu_{Co1,1} = 1$.

An analogous argument can be made to obtain $\mu_{Co1,0} = 0$. Argument for a convex combination is straightforward. If $q_{Co1|C} = 1 - \mu^L_{C0}(q_D)$, then $\mu_{Co1,0}(1 - \mu^L_{C0}(q_D)) + \mu^L_{C0}(q_D) = \mu^L_{C0}(q_D)$ when

setting $\mu_{Co2,0} = 1$. This makes $\mu_{Co1,1} = 0$ feasible. □

Proof of Proposition 4. The joint distribution of (Y, T, Z) can be written as its conditionals: $f(Y, T, Z) = f(Y|T, Z)f(T, Z)$. The vector q is sufficient to characterize distribution $f(T, Z)$ and H is sufficient to characterize $f(Y|T, Z)$. Since they individually parameterize each of the conditional distributions, they jointly exhaust all available data. □

Proof of Corollary 2. Follows directly from sufficiency theorem and applying Stoye (2010) Theorem 2. □

Proof of Proposition 5. It is sufficient to show that we have a just-identified system when there are no defiers. If this is true, then adding defiers results in an underidentified system, so we need an assumption for every group of defier that we allow for. There are $2^J - (J + 1)$ defier groups.

When monotonicity holds for J instrument values, the system of proportions with the law of total probability is:

$$\begin{aligned} p_{0J} &= q_{Nt} \\ p_{0(J-1)} &= q_{Nt} + q_{Co1} \\ &\vdots \\ p_{00} &= q_{Nt} + \sum_{X=1}^{J-1} q_{CoX} \\ 1 &= q_{Nt} + \sum_{X=1}^{J-1} q_{CoX} + q_{At} \end{aligned}$$

This is rewritten as $Aq = b$, with $q, b \in \mathbb{R}^{J+1}$ and A is a $(J + 1) \times (J + 1)$ invertible matrix. Then, $q = A^{-1}b$ is just identified. □

B System of Moment Conditions

In general, the nonparametric bootstrap is invalid because the procedure does not mimic the data-generating process when the sample does not correctly reflect the constraints that are binding. This property is similar to the better studied problem of doing inference on a parameter at the boundary, where bootstrap is known to be invalid. Hence, inference should be conducted by using moment conditions.

The constraint set in Equation 7 characterized by trimming bounds and mean restrictions can be written as three systems of constraints. The distribution of observables are (Y, T, Z) and the unknown parameters are $(q, \mu, r_q^{(T,Z)})$, where μ are the 8 means for each group and treatment and $r_q^{(T,Z)}$ are 8 quantile objects. Further, q is the vector of proportions: if q_D is known, then there are 3 unknowns here.

The first system of constraints comes from proportion restrictions in Equation 2.

$$\begin{aligned} E[1 - q_D - q_A - (1-T)(1-Z)] &= 0 \\ E[1 - q_D - q_N - TZ] &= 0 \\ E[TZ + (1-T)(1-Z) - 1 + q_D - q_C] &= 0 \end{aligned} \tag{18}$$

The second system comes from observed mean constraints in Equation 6.

$$\begin{aligned} E\left[\left(\frac{q_N \mu_{N0}}{q_N + q_C} + \frac{q_C \mu_{C0}}{q_N + q_C} - Y\right)(1-T)(1-Z)\right] &= 0 \\ E\left[\left(\frac{q_A \mu_{A1}}{q_A + q_D} + \frac{q_D \mu_{D1}}{q_A + q_D} - Y\right)T(1-Z)\right] &= 0 \\ E\left[\left(\frac{q_N \mu_{N0}}{q_N + q_D} + \frac{q_D \mu_{D0}}{q_N + q_D} - Y\right)(1-T)Z\right] &= 0 \\ E\left[\left(\frac{q_A \mu_{A1}}{q_A + q_C} + \frac{q_C \mu_{C1}}{q_A + q_C} - Y\right)TZ\right] &= 0 \end{aligned} \tag{19}$$

The final system comes from trimming bounds. Since there are 4 pairs of (T,Z) combinations with 2 groups each, and each with an upper and lower bound, we have 16 inequalities. The first block concerns the compliers and defiers while the second block concerns the always takers and never takers.

$$\begin{aligned} E[(\mu_{C1} - Y)TZ1\{Y \leq r^{(1,1)}_{q_C}\}] &\geq 0 \\ E[(Y - \mu_{C1})TZ1\{Y > r^{(1,1)}_{1-q_C}\}] &\geq 0 \\ E[(\mu_{C0} - Y)(1-T)(1-Z)1\{Y \leq r^{(0,0)}_{q_C}\}] &\geq 0 \\ E[(Y - \mu_{C0})(1-T)(1-Z)1\{Y > r^{(0,0)}_{1-q_C}\}] &\geq 0 \\ E[(\mu_{D1} - Y)T(1-Z)1\{Y \leq r^{(1,0)}_{q_D}\}] &\geq 0 \\ E[(Y - \mu_{D1})T(1-Z)1\{Y > r^{(1,0)}_{1-q_D}\}] &\geq 0 \\ E[(\mu_{D0} - Y)(1-T)Z1\{Y \leq r^{(0,1)}_{q_D}\}] &\geq 0 \\ E[(Y - \mu_{D0})(1-T)Z1\{Y > r^{(0,1)}_{1-q_D}\}] &\geq 0 \end{aligned} \tag{20}$$

$$
\begin{aligned}
E[(\mu_{A1} - Y)TZ1\{Y \leq r^{(1,1)}_{q_A}\}] &\geq 0 \\
E[(Y - \mu_{A1})TZ1\{Y > r^{(1,1)}_{1-q_A}\}] &\geq 0 \\
E[(\mu_{A1} - Y)T(1-Z)1\{Y \leq r^{(1,0)}_{q_A}\}] &\geq 0 \\
E[(Y - \mu_{A1})T(1-Z)1\{Y > r^{(1,0)}_{1-q_A}\}] &\geq 0 \\
E[(\mu_{N0} - Y)(1-T)(1-Z)1\{Y \leq r^{(0,0)}_{q_N}\}] &\geq 0 \\
E[(Y - \mu_{N0})(1-T)(1-Z)1\{Y > r^{(0,0)}_{1-q_N}\}] &\geq 0 \\
E[(\mu_{N0} - Y)(1-T)Z1\{Y \leq r^{(0,1)}_{q_N}\}] &\geq 0 \\
E[(Y - \mu_{N0})(1-T)Z1\{Y > r^{(0,1)}_{1-q_N}\}] &\geq 0
\end{aligned}
\tag{21}
$$

Although there are 16 quantile objects above, there are only 8 unknowns due to the following equalities:

$$
\begin{aligned}
r^{(1,1)}_{q_C} &= r^{(1,1)}_{1-q_A} & \qquad r^{(1,1)}_{q_A} &= r^{(1,1)}_{1-q_C} \\
r^{(1,0)}_{q_A} &= r^{(1,0)}_{1-q_D} & \qquad r^{(1,0)}_{q_D} &= r^{(1,0)}_{1-q_A} \\
r^{(0,1)}_{q_N} &= r^{(0,1)}_{1-q_D} & \qquad r^{(0,1)}_{q_D} &= r^{(0,1)}_{1-q_N} \\
r^{(0,0)}_{q_C} &= r^{(0,0)}_{1-q_N} & \qquad r^{(0,0)}_{q_N} &= r^{(0,0)}_{1-q_C}
\end{aligned}
\tag{22}
$$

Consequently, the problem can be written as a set of moment inequalities. We can then do inference on the unknowns. When the asymptotic distribution of these unknowns is found, we can do inference on β using simulation methods.

C More Empirical Examples

The first subsection provides further examples of monotonicity violations, which justifies this study. The second subsection provides further examples of PRTE in the applied literature.

C.1 Examples of Monotonicity Violations

To show the significance of this study, I show how defiers feature in many economic applications. I provide three further examples, with the first two pointed out in De Chaisemartin (2017).

First, in the judges design like Doyle (2007) and Aizer and Doyle Jr (2015), the random assignment of judges is used as an instrument for incarceration, or family removal in the context of foster care. Monotonicity in this context means that, if case worker (CW) A is more likely to send a kid to foster care than CW B, any kid who has been sent to foster care by CW B must be sent to foster care by CW A. But it is possible that CW B is less skilled, and made a mistake by sending the kid to foster care, and the more skilled CW A would rightly choose not to send the kid to foster care. Such a kid would then be considered a defier.

Second, defiers could be present in randomized controlled trials (RCT). Duflo and Saez (2003) study the impact of attending a meeting on the take-up of a retirement plan. To instrument for attending the meeting, they randomly assigned letters to subjects, and subjects who received a letter would receive a financial incentive upon attendance. Defiers are subjects who would attending the meeting if they had not received the letter, but receiving the letter would cause them not to attend the meeting. This is possible, and there is evidence in Gneezy and Rustichini (2000) that giving fines to parents who pick up their children late can crowd out their intrinsic motivation.

Third, Rao (2019) is interested in the effect of rich children being in the same study group as poor students on choosing to do charity work . The instrument used is whether the school's study group sorting is done alphabetically and that the rich child had a name that was alphabetically adjacent to poor students. Having two students with alphabetically adjacent names in schools that use alphabetical assignment would increase their probability of being put in the same group. A defier in this case is a student who would be with a poor child when he is in a school that does not use alphabetical assignment, but not if he were in a school with alphabetical assignment and has a name adjacent to a poor child. This can happen when schools without alphabetical assignment are intentional about mixing income groups (i.e., working against homophily), and schools with alphabetical assignment draw boundaries in the group based on their income level.

The framework presented in this paper is sufficiently general to deal with all cases above. However, the implementation using discrete instrument values is better suited when the instrument only takes on a few values. Knowledge of the extent of monotonicity violation in these contexts would be knowing that proportion of the population that are defiers. When there are many instruments, as in the judges case, an analogous approach can be taken, where, instead of assuming that the proportion of defiers is known, the researcher makes an assumption on the treatment selection equation.

C.2 Target PRTE in Empirical Applications

C.2.1 Marginal Returns to College

Carneiro et al. (2011) discuss PRTE for estimating marginal returns to education. One possible counterfactual is to increase the probability of going to college by a fixed proportion or fixed amount, though they were not specific as to what such a policy might be. Their other counterfactual policy is a change in tuition, which is a change in their instrument value.

Consider the cost of tuition as a specific example. Suppose the data has tuition taking values $Z \in \{0, 1\}$ but we have a counterfactual policy the increases tuition to $Z^* = 2$. In the base case, we have some colleges that offer free tuition, and others require one to pay. At the 0-1 margin, LATE can be calculated. Their PRTE is then the LATE at the 0-2 margin, so all colleges who require students to pay tuition increase their tuition. LATE only concerns compliers at the new margin. Under monotonicity, which they assumed, we would expect a different set of compliers, say C^*, who would go to college when $Z = 0$ but won't go to college when $Z = 2$. The new $LATE^*$ measures the average returns to college for this subgroup.

Under monotonicity, there are no defiers. This means never-takers remain never-takers under the new policy. The old C group are those who will go to college when tuition is free, and not when

they have to pay. Under the new policy, all of them will still remain compliers C^* because they would still go to college when it is free, and with a higher tuition cost, they would not go. Finally, we would expect some always-takers to now be compliers in C^*. Clearly, they would go to college when it's free. When they had to pay a small cost at $Z = 1$, they would still go to college. But with a higher tuition cost at $Z^* = 2$ some of them may choose to no longer go to college. This subset of A will now be in C^*, and the new A^* group is smaller. Since always-takers are those who are likely to have higher returns from college, under this counterfactual policy, we expect $LATE^* > LATE$. This analysis tells us that a possible PRTE is the causal effect of college on wages for the subpopulation that is somewhat sensitive to tuition cost. I use "somewhat sensitive", because they are less responsive than the group at the original 0-1 margin.

C.2.2 Meeting and Pension Plan

Duflo and Saez (2003) financially incentivized people to attend a meeting. People who received a letter (Z) would get money if they attended a meeting (T), and the research question is the effect of attending the meeting (T) on the take-up rate of a pension plan (Y).

Suppose we have a counterfactual policy that gives people more money if they attended the meeting. If we do not have defiers, then the analysis is entirely analogous to the Carneiro et al. (2011) example on returns to education. The analogous object of interest seems to be the TE of the meeting for people who are somewhat sensitive to financial incentives. Presumably, they are poorer individuals, and their take-up rate of a pension plan would hence be of interest.

Now suppose there are defiers. We can still place bounds on the TE of $Co1$ and $Co2$ as before, then we can interpret that object as the takeup rate of compliers who are somewhat sensitive to financial incentives. However, Co1 comes from pre-extrapolation group N, so without further assumption about its conditional means, its TE is in general not identified, as the worst case μ_{N1} is $\pm\infty$. But with bounded outcomes, the framework described is now able to do sensitivity analysis on the objects of interest in MST.

Various subpopulations might also be of interest. For instance, De3 is the group that are defiers at the 0-5 margin, never-takers at the 0-10 margin, and compliers at the 5-10 margin, for instrument taking values 0, 5 and 10 for the amount of money offered to attend the meeting. Behavioral studies on fund raisers in Gneezy and Rustichini (2000) show such behavior, where giving a bit of financial incentive disincentivises intrinsic effort, but offering a large financial incentive increases their effort. For people with such behavioral responses, what is their take-up rate of a pension plan? That could be an interesting research question.

Remark 10. *(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 to begin with. Exercise in extrapolation puts bounds on what the results of the experiment would have been if it had been designed differently.*

C.2.3 Price Subsidies for Bed Nets

Consider the counterfactual change in instrument value. Dupas (2014) conducted an RCT with randomly assigned prices for an antimalarial bed net (Olyset net), so Z is the randomly assigned price, T is whether the household purchases the net, and Y is whether the household uses the net. The experiment had a range of prices from 0 to 250 Kenyan Shillings (Ksh), but I will stick to the binary case of 0 and 250 Ksh for illustration. The counterfactual policy that MST considered in this example is offering the net at 150Ksh, which is the actual market price when the nets were rolled out a year later.

Their PRTE is the LATE at the 0-150 margin. They found that the PRTE's are decreasing in cost. This finding is consistent with higher prices excluding poor households who would use an Olyset net if they were able to purchase one. PRTE here corresponds to the usage rate of the net for the subpopulation of compliers at the price margin. Without defiers, the always-takers remain always-takers because if they were willing to buy at 250Ksh, they will buy at 150Ksh. Never-takers remain never-takers: if they would not take the net at 0Ksh, they would not take the net at 150Ksh, so they remain never-takers. Among the old complier group C, they would not buy the net at 250Ksh. But at 150Ksh, some of them would now be willing to buy the net, so they become always-takers in the new setting. Mapping this to the Table 1 response, this behavior is the Co2 group. Then, since the new LATE only measures the causal effect of C^*, it corresponds to $LATE^* = E[Y_1 - Y_0|Co1]$. This is a smaller subpopulation than the old C group, because the old C group consists of $Co1$ and $Co2$. The argument for being interested in this group is to find the usage rate among people who are highly price sensitive (i.e., more so than those who are sensitive at the 0-250 Ksh margin).

D More Objects of Interest

D.1 (Local) Average Treatment Effect

The objective is to (partially) identify the treatment effect $Y_1 - Y_0$. Under monotonicity, compliers are the largest subpopulation whose TE is point identified. But when there are defiers, both C and D are affected by the instrument, so their combined TE may also be interesting. Consequently, in this section, I explain three possible quantities we could be interested in: (i) TE on Compliers or Defiers (TEC/ TED), (ii) TE on the Marginal population (TEM), and (iii) Average Treatment Effect (ATE).

First observe that when outcomes are unbounded, treatment effects (TE) on A and N are not identified. Considering N, we only observe Y_0 and never Y_1, so in the worst case, Y_1 can be $\pm\infty$, so the treatment effect is unbounded. Thus, the objects that we can meaningfully identify are TE for compliers and defiers: $TEC := \mu_{C1} - \mu_{C0}$ and $TED := \mu_{D1} - \mu_{D0}$. We can simply solve the problem in Equation 7.

Taking the perspective that this method is a form of sensitivity analysis, TEC would be the relevant statistic. When monotonicity holds, TEC is point-identified. When we restrict $q_D \leq \lambda$, we can identify the maximum bias of the two-stage least squares estimator for TEC for a given λ.

In the presence of defiers and unbounded Y, we might be interested in the impact of treatment

on the marginal population, since this is the largest subpopulation that we can obtain nontrivial bounds of a treatment effect. Since A and N are unaffected by the instrument, the total mass of affected (marginal) population is $q_C + q_D$. This object is a weighted average of TEC and TED.

$$TEM := E[Y_1 - Y_0|\{C, D\}] = \frac{q_C(\mu_{C1} - \mu_{C0}) + q_D(\mu_{D1} - \mu_{D0})}{q_C + q_D}$$

If instead we have a bounded outcome space, an object that we might arguably be interested in is the ATE for the entire population $E[Y_1 - Y_0]$, which is also the object of interest in Balke and Pearl (1997). From a policy perspective, if we oblige treatment for everyone, the ATE is the relevant statistic. Since we know ex ante that $K_l \leq Y \leq K_h$, we can bound μ_{N1} and μ_{A0} by K to get nontrivial bounds, thereby allowing partial identification of ATE. Namely, $\mathcal{M} = \{\mu : \mu_{Gt} \in [K_l, K_h], \forall G, t\}$. Then, the upper bound is $\max_{\mu \in \mathcal{M}_\mathcal{S}} \sum_G q_G(\mu_{G1} - \mu_{G0})$, with an analogous expression for the lower bound.

As is already understood in literature, the worst-case bounds for the ATE occurs when there are no defiers. This occurs because ATE imputes the worst-case bounds for the A and N groups. When the proportion of D increases, the proportion of A and N decreases, so we have a smaller proportion that get worst-case bounds, thereby tightening the bounds. This means sensitivity analysis to monotonicity violations is irrelevant to ATE.

Applying this to Angrist and Evans (1998), they have a binary same sex instrument, when studying the impact of the third child on various outcomes (see their Tables 7 and 8). Working for pay is a binary outcome, so $\mu_{Gt} \in [0, 1]$, and one can calculate TEM and ATE to find the impact of the third child for either a subgroup or the entire population. There are other bounded outcomes like weeks worked and hours per week, which are amenable to the same methods. Finally, outcomes like labor income and family income are unbounded, so using ex ante bounds on μ_{Gt} may not be reasonable. Then, TEM gives meaningful information on the treatment effect for the largest possible subpopulation.

Remark 11. *TEC and TEM also reflect two natural options for dealing with the existence of defiers. TEC essentially ignores them, by truncating the switchers to only include compliers. TEM switches the sign of $E[Y|Z = 1] - E[Y|Z = 0]$ for a subset of the observations to reflect the treatment effect of defiers.*

D.2 Policy Effect and Intent to Treat

Instead of considering LATE, we can consider policy effects (PE). Suppose we are interested in expected outcomes Y for the new policy environment $Pol(1)$. The effect of the new policy can be written as $E[Y|Pol(1)] - E[Y|Pol(0)]$.

In counterfactual policies, the policy effect is often more interesting than the PRTE for policymakers, though the latter is more important for robustness checks. In Angrist and Evans (1998), "what is the impact on labour force participation of women with a $5000 childcare subsidy?" is a question on PE, while "what is the impact of the third child women's labor force participation for women who weakly prefer a balanced gender portfolio?" is a question on PRTE. Weak preference here refers to having a third child when the first two children are of the same gender when the $5000 childcare subsidy is available. In Dupas (2014), bed nets are sold at randomly selected prices

(Z), and they observe if the household bought a bed net (T) and if the household used the bed net. Here, "what proportion of the population uses a bednet when they are sold at 150Ksh instead of 250Ksh?" is a question on PE, while "what proportion of households who weakly prefer to buy a bed net use a bed net if they purchased one at price 150Ksh?" is a question on PRTE. In both situations, policy makers are more likely interested in the former rather than the latter question. The latter questions apply more to researchers who are interested in the external validity of their conclusions on the treatment effect.

In simple settings, where we impute the worst-case bounds (e.g., ± 1), PE will be perfectly linear, with slope equal to the coefficients given by p_z and q. In more complicated mixtures, however, such as when trimming bounds are binding, the linear program will be useful.

These policy effects are based on the intent to treat (ITT), the difference in outcomes when the instrument is assigned compared to when the instrument is not assigned. The notable difference here compared to the previous subsection is that we are no longer taking a weighted average of treatment effects. Instead, we put negative weights on the TE of defiers. If we do not extrapolate, policy effects are point-identified:

- Intent to Treat (ITT): $ITT = E[Y|Z=1] - E[Y|Z=0] = q_C(\mu_{C1} - \mu_{C0}) - q_D(\mu_{D1} - \mu_{D0})$ (Reduced Form coefficient)
- Policy Effect for Marginal Population (PEM): $PEM = \frac{ITT}{q_C + q_D}$
- Policy Effect for Net Compliers (PENC): $PENC = \frac{ITT}{q_C - q_D} = TSLS$. From a budget standpoint, $q_C - q_D$ is the relevant subpopulation of switchers, and we want to see the effect on those the policy paid for.

The reduced form is sometimes called the ITT in the policy evaluation literature, as it tells us the expected policy impact on the entire population. $ITT := E[Y|Z=1] - E[Y|Z=0]$. The PEM simply scales the ITT to capture only the switching population. Since $\frac{q_C - q_D}{q_C + q_D} < 1$ and $q_C + q_D < 1$, $|ITT| \leq |PEM| \leq |TSLS|$. Both TSLS and ITT are point identified, so PEM is also point-identified. The policy effect for net compliers makes TSLS interpretable, though the interpretation is now different from simply being the TE for compliers.

PEM is the weighted average of the TEC and the negative of the TED, which tells us the impact that the instrument has on the marginal population that has been switched between treatment and non-treatment. This would be most relevant if we were to roll out the instrument to the entire population as it is, and we want to know how the outcomes are affected among the population that switched between treatment and non-treatment.

$$PEM := \frac{q_C(\mu_{C1} - \mu_{C0}) - q_D(\mu_{D1} - \mu_{D0})}{q_C + q_D} = TSLS \cdot \frac{q_C - q_D}{q_C + q_D} = \frac{ITT}{q_C + q_D} \tag{23}$$

However, if we are interested in counterfactual policy environments, the ITT is not longer point identified. Recall our three objects are $ITT, \frac{ITT}{q_C + q_D}, \frac{ITT}{q_C - q_D}$. All of them are built from ITT and knowing proportions q^*. The q^* object is obtained directly from assumption. For instance, when we extrapolate right, $C^* : Co1, Co2$ and $D^* : De2, De4$. It hence suffices to show how ITT can be written as a linear combination of μ's.

One possible extrapolation is where we observe data for $Z \in \{0,1\}$, but we are interested in $E[Y|Z=2]$, as the instrument value is the policy implemented. Then, the policy effect can be written as ITT. To save on notation, let $\Delta\mu_{Ge} := \mu_{Ge,1} - \mu_{Ge,0}$. ITT only concerns the compliers and defiers at the instrument pair.

$$\begin{aligned}
ITT^*_{02} &= E[Y|Z=2] - E[Y|Z=0] \\
&= q_{Co1}(\Delta\mu_{Co1}) + q_{Co2}(\Delta\mu_{Co2}) - q_{De2}(\Delta\mu_{De2}) - q_{De4}(\Delta\mu_{De4}) \\
ITT^*_{01} &= q_{De1}(\Delta\mu_{De1}) + q_{Co2}(\Delta\mu_{Co2}) - q_{De2}(\Delta\mu_{De2}) - q_{De3}(\Delta\mu_{De3}) \\
ITT^*_{12} &= q_{Co1}(\Delta\mu_{Co1}) + q_{De3}(\Delta\mu_{De3}) - q_{De1}(\Delta\mu_{De1}) - q_{De4}(\Delta\mu_{De4})
\end{aligned}$$

Without parameterizing Y as a function of Z, the framework provides a viable method for placing bounds on these policy effects. These ITT objects are linear in μ, so the linear program is maintained.

While in some settings, the instrument is the policy itself (so that ITT is interpretable as the effect of the policy), this is no longer the case for threshold-crossing policies. In Angrist and Evans (1998), the same sex instrument is clearly not a policy, but the childcare subsidy is. Thus, when considering the policy effect (PE), we are interested in the effect of childcare policy on outcome Y. This is represented by $PE := E[Y|Pol(1)] - E[Y|Pol(0)]$. To save on notation, let $p_z := Pr(Z=1)$.

The expected outcome in the original setting is a linear combination of potential outcomes. This can be written in terms of the original $\{A, C, D, N\}$ groups (in the first equality below), or the threshold-crossing groups (second equality below).

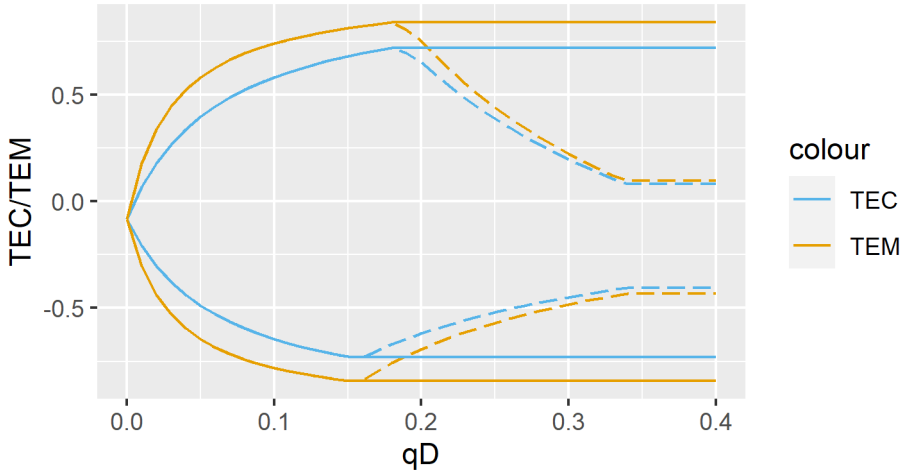
$$\begin{aligned}
E[Y|Pol(0)] &= q_A\mu_{A1} + q_C p_z\mu_{C1} + q_c(1-p_z)\mu_{C0} + q_D p_z\mu_{D0} + q_D(1-p_z)\mu_{D1} + q_N\mu_{N0} \\
&= q_A\mu_{A1} + q_{CC}p_z\mu_{CC1} + q_{CA}p_z\mu_{CA1} + q_{CC}(1-p_z)\mu_{CC0} + q_{CA}(1-p_z)\mu_{CA0} \\
&\quad + q_{DA}(1-p_z)\mu_{DA1} + q_{DD}(1-p_z)\mu_{DD1} + q_{DA}p_z\mu_{DA0} + q_{DD}p_z\mu_{DD0} + \sum_G q_{NG}\mu_{NG0} \\
E[Y|Pol(1)] &= q_A\mu_{A1} + q_{CA}\mu_{CA1} + q_{DA}\mu_{DA1} + q_{NA}\mu_{NA1} + q_{NN}\mu_{NN0} \\
&\quad + q_{CC}p_z\mu_{CC1} + q_{CC}(1-p_z)\mu_{CC0} + q_{NC}p_z\mu_{NC1} + q_{NC}(1-p_z)\mu_{NC0} \\
&\quad + q_{DD}(1-p_z)\mu_{DD1} + q_{DD}p_z\mu_{DD0} + q_{ND}(1-p_z)\mu_{ND1} + q_{ND}p_z\mu_{ND0}
\end{aligned}$$

Consequently, the policy effect is:

$$\begin{aligned}
PE &= E[Y|Pol(1)] - E[Y|Pol(0)] \\
&= q_{CA}(1-p_z)(\mu_{CA1} - \mu_{CA0}) + q_{DA}p_z(\mu_{DA1} - \mu_{DA0}) + q_{NA}(\mu_{NA1} - \mu_{NA0}) \\
&\quad + q_{NC}p_z(\mu_{NC1} - \mu_{NC0}) + q_{ND}(1-p_z)(\mu_{ND1} - \mu_{ND0})
\end{aligned}$$

Observe that groups $\{A, CC, DD, NN\}$ do not feature in PE, since their behavior in response to the instrument is the same in both environments, so their outcomes are unaffected by the policy. This PE object is thus still a linear function of μ, and can be solved using the procedure described, though it is a much simpler object to deal with.

Figure 3: Plot of TEC and TEM bounds against proportion of defiers q_D. q_D values are in increment of 0.01. $q_C - q_D = 0.0504$ in the data. Solid lines are bounds when q_D can take any value up to those indicated on the x-axis, and dashed lines are what we would obtain when imposing q_D to be a particular value.



D.3 Gender Preference Illustration

Consider the empirical application to Angrist and Evans (1998).

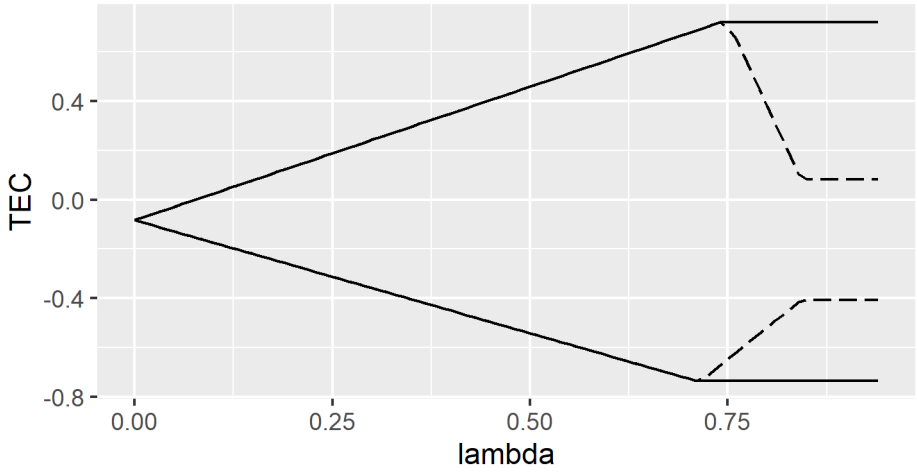
D.3.1 Treatment Effect for Compliers

Bounds on TEC and TEM from solving Equation 7 are presented in Figure 3. I present both the plot when imposing a particular q_D in dashed lines and the plot allowing q_D to be up to a particular value. When doing sensitivity analysis, bounds will gradually increase until it hits the maximum or minimum, then level off. Plotting it against q_D is more informative of how the optimization works, but not useful for sensitivity analysis.

When $q_D = 0$, we can point identify TEC, since the result is TSLS. As q_D increases, the bounds get wider, then gets narrower. It levels off when we have hit the maximum permissible proportion of defiers. To understand this behavior, observe that as q_D increases, there are two effects that work in opposite directions. One effect is that μ_{A1} and μ_{N0} have wider permissible bounds when q_D increases, since they no longer have to take up the entire $Y|T, Z$ distribution as in the $q_D = 0$ case. With observed mean restrictions, these give wider permissible bounds on μ_{C1} and μ_{C0}. The other effect is that q_C increases from Equation 2. With a larger mass of q_C, the trimming bounds on μ_{C1} and μ_{C0} get more restrictive. Thus, at small values of q_D, the first effect dominates, which results in wider bounds, but eventually the second effect dominates. The bounds flatten out after a certain point as we have reached the maximum q_D permitted by the data: this occurs when either q_A or q_N is zero.

A noteworthy point is that the solution vector for TEC, TED, and TEM are identical. This might initially be surprising since they are optimizing different objective functions, but it is in fact reasonable when we note that these objective functions optimize in the same direction. In the max

Figure 4: Plot of TEC against λ. Solid lines are bounds when imposing $q_D \leq \lambda q_C$ and dashed lines are what we would obtain when setting $q_D = \lambda q_C$.



problem, we aim to maximize μ_{C1} and μ_{D1} while minimizing μ_{C0} and μ_{D0}. In trimming bounds, we want μ_{C1} to be as large as possible, and this is equivalent to making μ_{A1} as small as possible in the $Y|T = 1, Z = 1$ distribution. Similarly, we want to make μ_{A1} as small as possible in the $Y|T = 1, Z = 0$ distribution to make μ_{D1} large. The same reasoning applies to μ_{N0}. Hence, the value of μ_{C1} that maximizes TEC also maximizes TED. Consequently, the same solution vector optimizes TEC, TED and TEM. In the ATE opimization problem, since $q_C + q_D$ is larger than q_N and q_A, the solver chooses the vertex that maximizes TEM, resulting in the same solution vector.

Some insight can be gained when plotting TEC against the sensitivity parameter λ directly, as in Figure 4. The bounds are now linear, which is not surprising when considering the closed-form bounds given in Equation (6) of Noack (2021), because the bounds are linear in $\frac{q_D}{q_C}$. This observation further supports how λ is an appropriate sensitivity parameter, especially when we are interested in objects like the TEC and its variations.

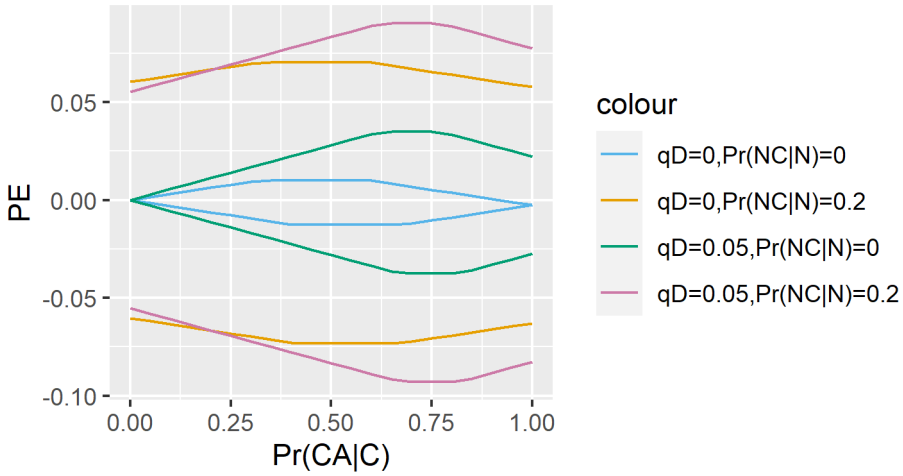
D.3.2 Policy Effect

In the Angrist and Evans (1998) setting, the relevant policy effect (PE) is $E[Y|Pol(0)]-E[Y|Pol(1)]$. In the simplest setting, set $Pr(DA|D) = Pr(NA|N) = Pr(ND|N) = 0$ to give some insight to the mechanics. Since $\{A, CC, DD, NN\}$ do not feature in PE, so the groups left that can influence PE are $\{CA, NC\}$. The result for this exercise is presented in Figure 5.

First consider the blue lines where monotonicity holds and $Pr(NC|N) = 0$, so the only group that affects PE is CA. When $Pr(CA|C) = 0$, PE will trivially be point identified and be 0, since $q_{CA} = 0$, so the only group that affects PE receives zero weight. This is a situation where the entire population only consists of $\{A, CC, DD, NN\}$ that do not feature in PE. On the other extreme, when $Pr(CA|C) = 1$, PE is point identified. Here, $q_{CA} = q_C$ and $\mu_{CA,t} = \mu_{C,t}$. Then, $PE = q_C(\mu_{C1} - \mu_{C0})(1 - p_z)$, which is a point-identified object, since $\mu_{C1} - \mu_{C0} = TEC$ is point-identified when monotonicity holds.

The intuition behind the bounds obtained in other counterfactual environments is similar to the

Figure 5: Plot of PE bounds against $Pr(CA|C)$ for various $q_D, Pr(NC|N)$ imposed. Set $Pr(DA|D) = 0, Pr(NA|N) = Pr(ND|N) = 0$.



discussion in PRTE. A notable feature is that, with the assumptions imposed, the worst-case bounds (± 1) are imputed for small values of $Pr(CA|C)$. Since $PE = q_{CA}(1 - p_z)(\mu_{CA1} - \mu_{CA0})$ and $\mu_{CA1} - \mu_{CA0} = \pm 1$, the upper and lower bounds will be symmetric around zero, with slope given by $q_{CA}(1-p_z)$. In general, this need not be true because a continuous conditional distribution of Y can yield trimming bounds that are strictly increasing or decreasing. The green bounds shows how bounds expand in the presence of non-monotonicity: even though defiers do not feature in PE (all of them are DD), they widen the bounds on $\mu_{CA1} - \mu_{CA0}$, which leads to wider PE bounds. A similar argument can be made when worst-case bounds are imputed when we allow for the NC group in the mixture.

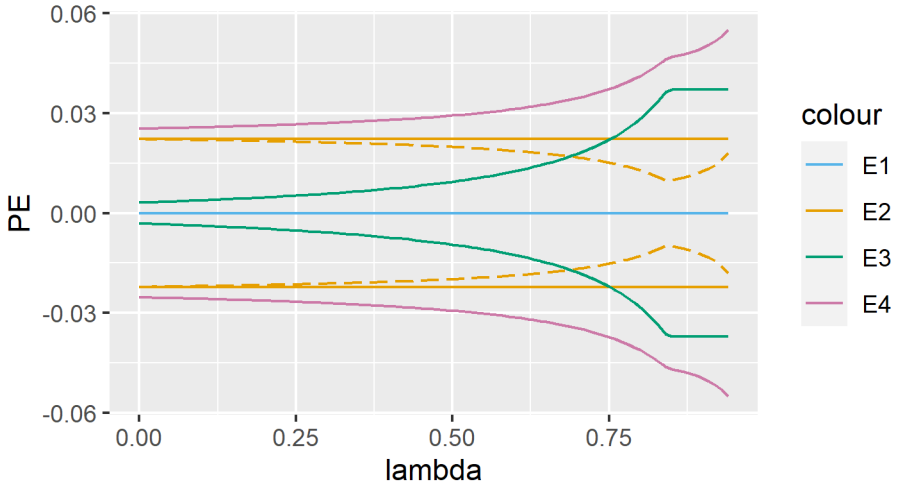
For sensitivity analysis to monotonicity violations, we may plot PE bounds against λ as in Figure 6. Unlike PRTE, we can no longer innocuously set some proportions to zero. Instead, one should consider what policy environment is reasonable. The first candidate restriction is where C and D switch proportionately into A i.e., $Pr(CA|C) = Pr(DA|D) = \tau_1$, which is reasonable when the subsidy incentivizes the third child in the dispreferred state for C and D in the same way. The second candidate restriction is that the N group switch proportionately into A, C and D groups in the new policy i.e., $\exists \tau_2$ such that $q_{NA} = \tau_2 q_A, q_{NC} = \tau_2 q_C, q_{ND} = \tau_2 q_D$, so proportions are given by response types in the original policy. τ_2 has to be sufficiently small such that the sum of these proportions is weakly smaller than q_N. This is reasonable when the gender preference among those who were never-takers are similar to those observed in the data, such that a similar proportion will switch when they are incentivized to have a child.

Using the policy environment described, observe that $q_{CA} = \tau_1 q_C, q_{CC} = (1 - \tau_1) q_C$ and the same can be written for q_{DA}, q_{DD}. Then, for some generic pair (τ_1, τ_2), Equation 11 can be written as:

$$q_D \leq \lambda_{(0)} q_C$$
$$(1 - \tau_1 + \tau_2) q_D \leq \lambda_{(1)} (1 - \tau_1 + \tau_2) q_C$$

This means that the two inequalities are identical for $\lambda = \lambda_{(0)} = \lambda_{(1)}$. For the four variations of policy environments in Figure 6, either one, both, or neither of the τ objects are set to zero, and

Figure 6: Plot of PE bounds against λ. Set $Pr(CA|C) = Pr(DA|D) = \tau_1$, and $q_{NA} = \tau_2 q_A, q_{NC} = \tau_2 q_C, q_{ND} = \tau_2 q_D$. Policy Environment E1 has $\tau_1 = 0, \tau_2 = 0$; Policy Environment E2 has $\tau_1 = 0, \tau_2 = 0.1$; Policy Environment E3 has $\tau_1 = 0.1, \tau_2 = 0$; Policy Environment E4 has $\tau_1 = 0.1, \tau_2 = 0.1$.



the same sensitivity argument applies. This makes λ sufficient as a sensitivity parameter.

When we set $\tau_1 = \tau_2 = 0$ (i.e. E1), all response types are in $\{A, CC, DD, NN\}$, which do not feature in PE, so we would obtain 0 identically. As with ATE, assuming that some defiers exist can sometimes tighten bounds, as in E2. This occurs because we only have N groups in the mixture, and having more defiers imputes worst-case bounds for a smaller proportion of the subpopulation of interest. Bounds widen in E3 and E4 for $\tau_1 \neq 0$ as more defiers implies a larger population enter CA and DA, so the policy affects a larger subpopulation and worst-case bounds broaden.

References

Aizer, A. and J. J. Doyle Jr (2015): "Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges," *The Quarterly Journal of Economics*, 130, 759–803.

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

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

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

Armstrong, T. B. and M. Kolesár (2021): "Sensitivity analysis using approximate moment condition models," *Quantitative Economics*, 12, 77–108.

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

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

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

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

DiNardo, J. and D. S. Lee (2011): "Program evaluation and research designs," in *Handbook of labor economics*, Elsevier, vol. 4, 463–536.

Dobbie, W., J. Goldin, and C. S. Yang (2018): "The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges," *American Economic Review*, 108, 201–40.

Doyle, J. J. J. (2007): "Child protection and child outcomes: Measuring the effects of foster care," *American Economic Review*, 97, 1583–1610.

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

Dupas, P. (2014): "Short-run subsidies and long-run adoption of new health products: Evidence from a field experiment," *Econometrica*, 82, 197–228.

Fang, Z., A. Santos, A. Shaikh, and A. Torgovitsky (2020): "Inference for Large-Scale Linear Systems with Known Coefficients," *University of Chicago, Becker Friedman Institute for Economics Working Paper*.

Frandsen, B. R., L. J. Lefgren, and E. C. Leslie (2019): "Judging judge fixed effects," Tech. rep., National Bureau of Economic Research.

Giné, E. and R. Nickl (2021): *Mathematical foundations of infinite-dimensional statistical models*, Cambridge University Press.

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

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

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

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

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

Ito, K., T. Ida, and M. Tanaka (2021): "Selection on Welfare Gains: Experimental Evidence from Electricity Plan Choice," Tech. rep., National Bureau of Economic Research.

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

——— (2021): "The identification region of the potential outcome distributions under instrument independence," *Journal of Econometrics*.

Kline, P. and C. R. Walters (2019): "On Heckits, LATE, and numerical equivalence," *Econometrica*, 87, 677–696.

Kowalski, A. E. (2018): "Reconciling seemingly contradictory results from the Oregon health insurance experiment and the Massachusetts health reform," Tech. rep., National Bureau of Economic Research.

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

Manski, C. F. (1989): "Anatomy of the selection problem," *Journal of Human resources*, 343–360.

Masten, M. A. and A. Poirier (2018): "Identification of treatment effects under conditional partial independence," *Econometrica*, 86, 317–351.

——— (2020): "Inference on breakdown frontiers," *Quantitative Economics*, 11, 41–111.

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

Mourifié, I. and Y. Wan (2017): "Testing local average treatment effect assumptions," *Review of Economics and Statistics*, 99, 305–313.

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

Noack, C. (2021): "Sensitivity of LATE Estimates to Violations of the Monotonicity Assumption," *arXiv preprint arXiv:2106.06421*.

Rao, G. (2019): "Familiarity does not breed contempt: Generosity, discrimination, and diversity in Delhi schools," *American Economic Review*, 109, 774–809.

Semenova, V. (2020): "Better Lee Bounds," *arXiv preprint arXiv:2008.12720*.

Stoye, J. (2010): "Partial identification of spread parameters," *Quantitative Economics*, 1, 323–357.

Vytlacil, E. (2002): "Independence, monotonicity, and latent index models: An equivalence result," *Econometrica*, 70, 331–341.

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*, Springer, 14–29.