Sharp Bounds on Treatment Effects for Policy Evaluation[*]

Sukjin Han
Department of Economics
University of Bristol
sukjin.han@gmail.com

Shenshen Yang
Department of Economics
University of Texas at Austin
shenshenyang@utexas.edu

This Draft: September 28, 2020

Abstract

For counterfactual policy evaluation, it is important to ensure that treatment parameters are relevant to the policies in question. This is especially challenging under unobserved heterogeneity, as is well featured in the definition of the local average treatment effect (LATE). Being intrinsically local, the LATE is known to lack external validity in counterfactual environments. This paper investigates the possibility of extrapolating local treatment effects to different counterfactual settings when instrumental variables are only binary. We propose a novel framework to systematically calculate sharp nonparametric bounds on various policy-relevant treatment parameters that are defined as weighted averages of the marginal treatment effect (MTE). Our framework is flexible enough to incorporate a large menu of identifying assumptions beyond the shape restrictions on the MTE that have been considered in prior studies. We apply our method to understand the effects of medical insurance policies on the use of medical services.

JEL Numbers: C14, C32, C33, C36

Keywords: Heterogeneous treatment effects, local average treatment effects, marginal treatment effects, extrapolation, partial identification.

[*] The authors are grateful to Jason Abrevaya, Brendan Kline, Xun Tang, Alex Torgovitsky, Ed Vytlacil, Haiqing Xu, and participants in the 2020 Texas Econometrics Camp and the workshop at UT Austin for helpful comments and discussions.

1 Introduction

For counterfactual policy evaluation, it is important to ensure that treatment parameters are relevant to the policies in question. This is especially challenging in the presence of unobserved heterogeneity. This challenge is well featured in the definition of the local average treatment effect (LATE). The LATE has been one of the most popular treatment parameters used by empirical researchers since it was introduced by Imbens and Angrist (1994). It induces a straightforward linear estimation method that requires only a binary instrumental variable (IV), and yet, allows for unrestricted treatment heterogeneity. The unfortunate feature of the LATE is that, as the name suggests, the parameter is intrinsically local, recovering the average treatment effect (ATE) for a specific subgroup of population called compliers. This feature leads to two major challenges in making the LATE a valuable parameter for counterfactual policy evaluation. First, the subpopulation for which the effect is measured may not be the population of policy interest. Second, the definition of the subpopulation depends on the IV chosen, rendering the parameter even more difficult to extrapolate to new environments.

Dealing with the lack of external validity of the LATE has been an important theme in the literature. One approach in theoretical work (Angrist and Fernandez-Val (2010); Bertanha and Imbens (2019)) and empirical research (Dehejia et al. (2019); Muralidharan et al. (2019)) has been to show the similarity between complier and non-complier groups based on observables. This approach, however, cannot attend to possible unobservable discrepancies between these groups. Heckman and Vytlacil (2005) unify well-known treatment parameters by expressing them as weighted averages of what they define as the marginal treatment effect (MTE). This MTE framework has a great potential for extrapolation because a class of treatment parameters that are policy-relevant can also be generated as weighted averages of the MTE. The only obstacle is that the MTE is identified via a method called local IV (Heckman and Vytlacil (1999)), which requires the continuous variation of the IV that is sometime large depending on the targeted support. This in turn reflects the intrinsic difficulty of extrapolation when available exogenous variation is only discrete. Acknowledging this nature of the challenge, previous studies in the literature have proposed imposing shape restrictions on the MTE, which is a function of the treatment-selection unobservable, while allowing for binary instruments in the framework of Heckman and Vytlacil (2005). Brinch et al. (2017) introduce shape restrictions (e.g., linearity) on the MTE functions in an attempt to identify the LATE extrapolated to different subpopulations or to test for its externality validity. More recently, Mogstad et al. (2018) propose a general partial identification framework where bounds on various policy-relevant treatment parameters can be obtained from a set of "IV-like estimands" that are directly identified from the data and routinely obtained in empirical work. Kowalski (2020) applies an approach similar to these studies to extrapolate the results from one health insurance experiment to an external setting.

This paper continues this pursuit and investigates the possibility of extrapolating local treatment parameters to different policy settings in the MTE framework when IVs are only binary. In a partial identification framework similar in spirit to Mogstad et al. (2018), we show how to systematically calculate sharp nonparametric bounds on various extrapolated treatment parameters for binary (or more generally, discrete) outcomes using instruments that are allowed to be binary. These parameters are defined as weighted averages of the MTE. Examples include the ATE, the treatment on the treated, the LATE for subgroups induced by new policies, and the policy-relevant treatment effect (PRTE). We also show how to place in this procedure restrictions from a large menu of identifying assumptions beyond the shape restrictions considered in earlier work.

In this paper, we make four main contributions. First, we propose a novel framework for calculating bounds on policy-relevant treatment parameters. We introduce the probability of the latent state of the outcome-generating process conditional on the treatment-selection unobservable. This latent conditional probability is the key ingredient for our analysis, as both the target parameter and the distribution of the observables can be written as linear functionals of it. Therefore, having it as a decision variable, we can formulate infinite-dimensional linear programming that produces bounds on a targeted treatment parameter. This approach is reminiscent of Balke and Pearl (1997) and can be viewed as its generalization to the MTE framework. Balke and Pearl (1997) introduce a linear programming approach to characterize bounds on the ATE with a binary outcome, treatment and instrument. The main distinction of our approach is that the latent probability is conditioned on the selection unobservable, which is important for our extrapolation purpose. To make it feasible to solve the resulting infinite-dimensional program, we use a sieve-like approximation of the program and produce a finite-dimensional linear program (LP). This approximation approach builds on Mogstad et al. (2018), although they use approximation directly on the MTE function. We also propose a conservative approach to choosing the sieve dimension in practice.

Second, by formulating the LP based on the latent conditional probability rather than the MTE, it creates a flexible environment where we can introduce identifying assumptions that have not been used in the context of the MTE framework or the LATE extrapolation. We propose assumptions that there exist exogenous variables other than IVs. We propose two types of exogenous variables that have been used in the context of identifying the ATE in the literature: Mourifié (2015), Han and Vytlacil (2017), Vuong and Xu (2017), and Han and Lee (2019) use the first type, and Vytlacil and Yildiz (2007), Shaikh and Vytlacil (2011), and

Balat and Han (2018) use the second type. We utilize these variables in this novel context of the MTE framework. Also, while the earlier papers exploit these variables in combination with rank similarity or rank invariance, we show that they independently have identifying power for treatment parameters, including the ATE. We also propose identifying assumptions such as uniformity and the direction of endogeneity in this MTE framework. The direction of endogeneity is sometimes assumed in empirical work to characterize selection bias and has been shown to have identifying power (Manski and Pepper (2000)). The uniformity assumption is related to rank similarity or rank invariance (Chernozhukov and Hansen (2005)). The shape restrictions on the MTE considered in the literature can also be nested within our framework, since the MTE is just a sum of the latent conditional probabilities. The assumptions on the existence of exogenous variables complement the identifying assumptions that rely on the researcher's prior, in that its identifying power comes from actual data. When a confidence set is constructed under one of the latter assumptions, we can conduct a specification test for that assumption.

Third, we show that our approach yields straightforward proof of the sharpness of the resulting bounds. This feature stems from the use of the latent conditional probability in the linear programming and the convexity of the feasible set in the program. When the MTE itself is the target parameter, we distinguish between the notions of point-wise and uniform sharpness and argue why uniform sharpness is often difficult to achieve.

Fourth, as an application, we study the effects of insurance on medical service utilization by considering various counterfactual policies related to insurance coverage. The LATE for compliers and the bounds on the LATE for always-takers and never-takers reveal that possessing private insurance has the largest effect on medical visits for never takers, i.e., those who face higher insurance cost. This provides a policy implication that lowering the cost of private insurance is important, because the high cost might hinder people with most need from receiving adequate medical services.

The linear programming approach to partial identification of treatment effects was pioneered by Balke and Pearl (1997) and recently gained attention in the literature; see, e.g., Mogstad et al. (2018), Torgovitsky (2019a), Machado et al. (2019), Kamat (2019), Gunsilius (2019), and Han (2020b).[1] As these papers suggest, there are many settings, including ours, where analytical derivation of bounds is cumbersome or nearly impossible due to the complexity of the problems.

This paper will proceed as follows. The next section introduces the main observables, maintained assumptions, and target parameters. Section 3 defines the latent conditional

[1] For the computational approach in contexts other than program evaluation, see Manski (2007), Kitamura and Stoye (2019), Deb et al. (2017), and Tebaldi et al. (2019).

probability and formulates the infinite-dimensional LP, and Section 4 introduces sieve approximation to the program. Section 5 then generalize the analysis to incorporate additional exogenous variables. Section 6 proposes a menu of identifying assumptions and shows how they can easily be incorporated in the LP. Section 7 provides numerical illustrations, and Section 8 contains an empirical application. In the Appendix, Section A lists the examples of target parameters. Section B discusses (i) the point-wise and uniform sharpness for the MTE bounds, (ii) inference, especially how to conduct specification tests for identifying assumptions, (iii) an extension with continuous covariates, and (iv) the relationship between this paper's LP and those in Mogstad et al. (2018). All proofs are contained in Section C.

2 Observables and Target Parameters

Assume that we observe the binary outcome $Y \in \{0, 1\}$, binary treatment $D \in \{0, 1\}$, and binary instrument $Z \in \{0, 1\}$. We may additionally observe (possibly endogenous) discrete covariates $X \in \mathcal{X}$.[2] Binary Y is common in empirical work. Binary Z is also common, especially in randomized experiments, and allowing for this minimal exogenous variation is the key challenge for extrapolation that we want to address in this paper. Still, the analysis of this paper can be extended to allow for general discrete Y and Z. Let $Y(d)$ be the counterfactual outcome given $D = d$, which is consistent with the observed outcome: $Y = DY(1) + (1 - D)Y(0)$. We maintain the following assumptions:

Assumption SEL. $D = 1\{U \leq P(Z, X)\}$ *where* $P(Z, X) \equiv \Pr[D = 1|Z, X]$ *and* $U|_{X=x} \sim Unif[0, 1]$ *for* $x \in \mathcal{X}$*.*

Assumption EX. $(Y(d), D(z)) \perp Z|X$*.*

Assumption SEL imposes a selection model for D, which is important in motivating and interpreting marginal treatment effects later. This assumption is also equivalent to Imbens and Angrist (1994)'s monotonicity assumption (Vytlacil (2002)). We introduce the standard normalization that $U \sim Unif[0, 1]$ conditional on $X = x$.[3] Assumption EX imposes the exclusion restriction and conditional independence for Z.

[2] We focus on discrete X as it simplifies the exposition. Section B.3 in the Appendix extends the framework to incorporate continuously distributed X.

[3] Note that for any index function $g(z, x)$ and an unobservable ε with any distribution, the selection model satisfies $D = 1\{\varepsilon \leq g(Z, X)\} = 1\{F_{\varepsilon|X}(\varepsilon|X) \leq F_{\varepsilon|X}(g(Z, X)|X)\} = 1\{U \leq P(Z, X)\}$, since $P(z, x) = \Pr[\varepsilon \leq g(z, x)|X = x] = \Pr[U \leq F_{\varepsilon|X}(g(z, x)|x)|X = x] = F_{\varepsilon|X}(g(z, x)|x)$ and $F_{\varepsilon|X}(\varepsilon|X) = U$ is uniformly distributed conditional on X.

Heckman and Vytlacil (2005) establish that various treatment parameters can be expressed as integral equations of the MTE, defined as

$$E[Y(1) - Y(0)|U = u, X = x].$$

Following Mogstad et al. (2018), it is convenient to introduce the marginal treatment response (MTR) function

$$\begin{aligned} m_d(u, x) &\equiv E[Y(d)|U = u, X = x] \\ &= \Pr[Y(d) = 1|U = u, X = x]. \end{aligned}$$

Then, the MTE can be expressed as $m_1(u, x) - m_0(u, x)$. Now, we define the target parameter τ to be a weighted average of the MTE:

$$\tau = E[\tau_1(Z, X) - \tau_0(Z, X)], \tag{2.1}$$

where

$$\tau_d(z, x) = \int m_d(u, x) w_d(u, z, x) du \tag{2.2}$$

by using $F_{U|X}(u|x) = u$, and $w_d(u, z, x)$ is a known weight specific to the parameter of interest.[4] This definition agrees with the insight of Heckman and Vytlacil (2005). The target parameter includes a wide range of policy-relevant treatment parameters. With a Dirac delta function for a given value u as the weight, the MTE itself can be an example. We list a few examples of the target parameter here; other examples can be found in Table 4 in the Appendix.

Example 1. *The ATE can be a target parameter with $w_d(u, z, x) = 1, \forall u, z, x$.*

$$\tau_{ATE} = E\left[\int_0^1 m_1(u, X) du - \int_0^1 m_0(u, X) du\right]$$

Example 2. *The generalized LATE for always-takers and never-takers are also target parameters. Here, we give the expression of the LATE for always-takers as an example. Assume $P(z, x)$ increases in z for any given $x \in \mathcal{X}$. For the always-taker (AT) LATE, we give weight $\frac{1}{P(0,x)}$ to individuals with $u \in [0, P(0, x)]$ and thus, we have $w_d(u, z, x) = \frac{1(u \in [0, p(0,x)])}{p(0,x)}$.*

[4] Mogstad et al. (2018) define the weight in a slightly different way.

$$\tau_{LATE\text{-}AT} = E\left[\int_0^1 m_1(u,X)\frac{1(u \in [0,p(0,X)])}{p(0,X)}du - \int_0^1 m_0(u,X)\frac{1(u \in [0,p(0,X)])}{p(0,X)}du\right]$$

Example 3. *The policy relevant treatment effect (PRTE) is a target parameter that is particularly useful for policy evaluation. It is defined as the welfare difference between two different policies. Let Z and Z' be two instrument variables under two policies and $P(Z,X)$ and $P'(Z',X)$ be propensity scores under the two policies.*

$$\begin{aligned}\tau_{PRTE} = E\Bigg[&\int_0^1 m_1(u,X)\frac{\Pr\left[u \le P'(Z',X)\right] - \Pr\left[u \le P(Z,X)\right]}{E\left[P'(Z',X)\right] - E\left[P(Z,X)\right]}du \\ &- \int_0^1 m_0(u,X)\frac{\Pr\left[u \le P'(Z',X)\right] - \Pr\left[u \le P(Z,X)\right]}{E\left[P'(Z',X)\right] - E\left[P(Z,X)\right]}du\Bigg]\end{aligned}$$

In these examples, the weights w_0 and w_1 can be set asymmetrically to define a broader class of parameters. All the parameters we consider in this paper can be defined conditional on X, although we omit them for succinctness.

3 Distribution of Latent State and Infinite-Dimensional Linear Program

As a crucial first step of our analysis, we define a state variable that determines a specific mapping of

$$d \mapsto y.$$

Since $d \in \{0,1\}$ and $y \in \{0,1\}$, there are four possible maps from d onto y. Define a discrete latent variable ϵ whose value e corresponds to each possible map:

$$\epsilon \in \mathcal{E},$$

where $|\mathcal{E}| = 4$ with $\mathcal{E} \equiv \{1,2,3,4\}$. That is, ϵ is a decimal transformation of a binary sequence $(Y(1),Y(0))$, which captures the treatment effect heterogeneity. For the later purpose, it is helpful to explicitly define the map as

$$y = g_e(d)$$

and write

$$Y(d) = g_\epsilon(d), \tag{3.1}$$

which implies $Y = g_\epsilon(D)$. It is important to note that no structure is imposed in introducing $g_e(\cdot)$ because the mapping is saturated by binary Y and D. By (3.1) and Assumption SEL, Assumption EX can be equivalently stated as $(\epsilon, U) \perp Z|X$. Still, ϵ and X can be correlated as X is allowed to be endogenous.

Now, as a key component of our LP, we define the probability mass function of ϵ conditional on (U, X): for $e \in \mathcal{E}$,

$$q(e|u,x) \equiv \Pr[\epsilon = e|U = u, X = x] \tag{3.2}$$

with $\sum_{e\in\mathcal{E}} q(e|u,x) = 1$ for any u, x. The quantity $q(e|u,x)$ captures endogenous treatment selection. It is shown below that this latent conditional probability is a building block for various treatment parameters and thus serves as the decision variable in the LP. The introduction of $q(e|u,x)$ distinguishes our approach from those in Balke and Pearl (1997) and Mogstad et al. (2018). Since the probability is conditional on continuously distributed U, the simple finite-dimensional linear programming approach of Balke and Pearl (1997) is no longer applicable. Instead, we use an approximation method similar to Mogstad et al. (2018). However, Mogstad et al. (2018) uses the MTR function as a building block for treatment parameters and introduces the "IV-like" estimands as a means of funneling the information from the data. Unlike in Mogstad et al. (2018), $q(e|u,x)$ can be directly related to the distribution of data. This allows us to later incorporate identifying assumptions that are difficult to incorporate within the framework of Mogstad et al. (2018).

By (3.1) and (3.2), note that

$$\begin{aligned}\Pr[Y(d) = 1|U = u, X = x] &= \Pr[\epsilon \in \{e \in \mathcal{E} : g_e(d) = 1\}|U = u, X = x] \\ &= \sum_{e\in\mathcal{E}:g_e(d)=1} q(e|u,x).\end{aligned}$$

Therefore, the MTR can be expressed as

$$m_d(u,x) = \sum_{e:g_e(d)=1} q(e|u,x). \tag{3.3}$$

Combining (3.3) and (2.2), we have $\tau_d(z,x) = \sum_{e:g_e(d)=1} \int q(e|u,x)w_d(u,z,x)du$, and thus the target parameter $\tau = E[\tau_1(Z,X)] - E[\tau_0(Z,X)]$ in (2.1) can be written as

$$\tau = \sum_{e:g_e(1)=1} \int E[q(e|u,X)w_1(u,Z,X)]du - \sum_{e:g_e(0)=1} \int E[q(e|u,X)w_0(u,Z,X)]du \tag{3.4}$$

for some q that satisfies the properties of probability.

The goal of this paper is to (at least partially) infer the target parameter τ based on the data, i.e., the distribution of (Y,D,Z,X). The key insight is that there are observationally equivalent $q(e|u,x)$'s that are consistent with the data, which in turn produces observationally equivalent τ's that define the identified set.

Let $p(y,d|z,x) \equiv \Pr[Y=y, D=d|Z=z, X=x]$ be the observed conditional probability. This data distribution imposes restrictions on $q(e|u,x)$. For instance, for $D=1$,

$$\begin{aligned} p(y,1|z,x) &= \Pr[Y(1)=y, U \leq P(z,x)|Z=z, X=x] \\ &= \Pr[Y(1)=y, U \leq P(z,x)|X=x] \end{aligned}$$

by Assumption EX, but

$$\begin{aligned} \Pr[Y(1)=y, U \leq P(z,x)|X=x] &= \int_0^{P(z,x)} \Pr[Y(1)=y|U=u, X=x]du \\ &= \sum_{e:g_e(1)=y} \int_0^{P(z,x)} q(e|u,x)du, \end{aligned} \tag{3.5}$$

where the second equality is by $\Pr[Y(d)=y|U=u,X=x] = \sum_{e:g_e(d)=y} q(e|u,x)$.

To define the identified set for τ, we introduce some simplifying notation. Let $q(u,x) \equiv \{q(e|u,x)\}_{e\in\mathcal{E}}$ and

$$\mathcal{Q} \equiv \{q(\cdot) : \sum_{e\in\mathcal{E}} q(e|u,x) = 1 \, \forall (u,x) \text{ and } q(e|u,x) \geq 0 \, \forall (e,u,x)\}$$

be the class of $q(u,x)$, and let $p \equiv \{p(1,d|z,x)\}_{(d,z,x)\in\{0,1\}^2\times\mathcal{X}}$. Also, let $R_\tau : \mathcal{Q} \to \mathbb{R}$ and $R_0 : \mathcal{Q} \to \mathbb{R}^{d_p}$ (with d_p being the dimension of p) denote the linear operators of $q(\cdot)$ that satisfy

$$R_\tau q \equiv \sum_{e:g_e(1)=1} \int E[q(e|u,X)w_1^\tau(u,Z,X)]du - \sum_{e:g_e(0)=1} \int E[q(e|u,X)w_0^\tau(u,Z,X)]du,$$
$$R_0 q \equiv \sum_{e:g_e(d)=1} \int_{\mathcal{U}_{z,x}^d} q(e|u,x)du,$$

where $\mathcal{U}_{z,x}^d$ denotes the intervals $\mathcal{U}_{z,x}^1 \equiv [0, P(z,x)]$ and $\mathcal{U}_{z,x}^0 \equiv (P(z,x), 1]$.

Definition 3.1. *The identified set of τ is defined as*

$$\mathcal{T}^* \equiv \{\tau \in \mathbb{R} : \tau = R_\tau q \text{ for some } q \in \mathcal{Q} \text{ such that } R_0 q = p\}.$$

In what follows, we formulate the infinite-dimensional LP (∞-LP) that characterizes $\mathcal{T}^*$. This program conceptualizes sharp bounds on τ from the data and the maintained assumptions (Assumptions SEL and EX). The upper and lower bounds on τ are defined as

$$\overline{\tau} = \sup_{q\in\mathcal{Q}} R_\tau q, \tag{∞-LP1}$$
$$\underline{\tau} = \inf_{q\in\mathcal{Q}} R_\tau q, \tag{∞-LP2}$$

subject to

$$R_0 q = p. \tag{∞-LP3}$$

Observe that the set of constraints (∞-LP3) does not include

$$\sum_{e:g_e(d)=0} \int_{\mathcal{U}_{z,x}^d} q(e|u,x)du = p(0,d|z,x) \qquad \forall(d,z,x) \in \{0,1\}^2 \times \mathcal{X}. \tag{3.6}$$

This is because we know a priori that they are redundant in the sense that they do not further restrict the *feasible set*, i.e., the set of $q(e|u,x)$'s that satisfy all the constraints ($q \in \mathcal{Q}$ and (∞-LP3)).

Lemma 3.1. *In the linear program* (∞-LP1)–(∞-LP3)*, the feasible set defined by $q \in \mathcal{Q}$ and* (∞-LP3) *is identical to the feasible set defined by $q \in \mathcal{Q}$,* (∞-LP3)*, and* (3.6)*.*

Theorem 3.1. *Under Assumptions SEL and EX, suppose $\mathcal{T}^*$ is non-empty. Then, the bounds $[\underline{\tau}, \overline{\tau}]$ in* (∞-LP1)–(∞-LP3) *are sharp for the target parameter τ, i.e., $cl(\mathcal{T}^*) = [\underline{\tau}, \overline{\tau}]$, where $cl(\cdot)$ is the closure of a set.*

The result of this theorem is immediate due to the convexity of the feasible set $\{q : q \in \mathcal{Q}\} \cap \{q : R_0 q = p\}$ in the LP and the linearity of $R_\tau q$ in q, which implies that $[\underline{\tau}, \overline{\tau}]$ is convex.

4 Sieve Approximation and Finite-Dimensional Linear Programming

Although conceptually useful, the LP (∞-LP1)–(∞-LP3) is not feasible in practice because $\mathcal{Q}$ is an infinite-dimensional space. In this section, we approximate (∞-LP1)–(∞-LP3) with a finite-dimensional LP via a sieve approximation of the conditional probability $q(e|u,x)$. We use Bernstein polynomials as the sieve basis. Bernstein polynomials are useful in imposing restrictions on the original function (Joy (2000); Chen et al. (2011); Chen et al. (2017)) and therefore have been introduced in the context of linear programming (Mogstad et al. (2018); Masten and Poirier (2018); Mogstad et al. (2019)).

Consider the following sieve approximation of $q(e|u,x)$ using Bernstein polynomials of order K

$$q(e|u,x) \approx \sum_{k=1}^{K} \theta_k^{e,x} b_k(u),$$

where $b_k(u) \equiv \binom{K}{k} x^k (1-x)^{K-k}$ is a univariate Bernstein basis, $\theta_k^{e,x} \equiv \theta_{k,K}^{e,x} \equiv q(e|k/K,x)$ is its coefficient, and K is finite. It is important to note that x can index θ, because $q(e|u,x)$ is a saturated function of x. By the definition of the Bernstein coefficient, for any (e,x), it satisfies $q(e|u,x) \geq 0$ for all u if and only if $\theta_k^{e,x} \geq 0$ for all k. Also, $\sum_{e\in\mathcal{E}} q(e|u,x) = 1$ for all (u,x) is approximately equivalent to $\sum_{e\in\mathcal{E}} \theta_k^{e,x} = 1$ for all (k,x). To see this, first, $\sum_{e\in\mathcal{E}} q(e|u,x) = 1$ for all (u,x) implies $\sum_{e\in\mathcal{E}} \theta_k^{e,x} = \sum_{e\in\mathcal{E}} q(e|k/K,x) = 1$ for all (k,x). Conversely, when $\sum_{e\in\mathcal{E}} \theta_k^{e,x} = 1$ for all (k,x),

$$\sum_{e\in\mathcal{E}} q(e|u,x) \approx \sum_{e\in\mathcal{E}} \sum_{k=1}^{K} \theta_k^{e,x} b_k(u) = \sum_{k=1}^{K} b_k(u) = 1$$

by the binomial theorem (Coolidge (1949)). Motivated by this approximation, we formally define the following sieve space for $\mathcal{Q}$:

$$\mathcal{Q}_K \equiv \left\{ \left\{ \sum_{k=1}^{K} \theta_k^{e,x} b_k(u) \right\}_{e\in\mathcal{E}} : \sum_{e\in\mathcal{E}} \theta_k^{e,x} = 1 \, \forall (k,x) \text{ and } \theta_k^{e,x} \geq 0 \, \forall (e,k,x) \right\} \subseteq \mathcal{Q}. \tag{4.1}$$

Let $\mathcal{K} \equiv \{1, ..., K\}$ and $p(z,x) \equiv \Pr[Z = z, X = x]$. For $q \in \mathcal{Q}_K$, by (3.4) and (4.1), the target parameter $\tau = E[\tau_1(Z,X)] - E[\tau_0(Z,X)]$ can be expressed with

$$\begin{aligned} E[\tau_d(Z,X)] &= \sum_{e:g_e(d)=1} \sum_{(k,x)\in\mathcal{K}\times\mathcal{X}} \theta_k^{e,x} \int b_k(u) \sum_{z\in\{0,1\}} w_d(u,z,x)p(z,x)du \\ &\equiv \sum_{e:g_e(d)=1} \sum_{(k,x)\in\mathcal{K}\times\mathcal{X}} \theta_k^{e,x}\gamma_k^d(x), \end{aligned} \tag{4.2}$$

where $\gamma_k^d(x) \equiv \int b_k(u) \sum_{z\in\{0,1\}} w_d(u,z,x)p(z,x)du$. Also, for $q \in \mathcal{Q}_K$ and $D = 1$, by (3.5), we have

$$\begin{aligned} p(y,1|z,x) &= \sum_{e:g_e(d)=y} \sum_{k\in\mathcal{K}} \theta_k^{e,x} \int_0^{P(z,x)} b_k(u)du \\ &\equiv \sum_{e:g_e(d)=y} \sum_{k\in\mathcal{K}} \theta_k^{e,x}\delta_k^1(z,x), \end{aligned} \tag{4.3}$$

where $\delta_k^d(z,x) \equiv \int_{\mathcal{U}_{z,x}^d} b_k(u)du$.

From (4.2) and (4.3), we can expect that a finite-dimensional LP can be obtained with respect to $\theta_k^{e,x}$. Let $\theta \equiv \{\theta_k^{e,x}\}_{(e,k,x)\in\mathcal{E}\times\mathcal{K}\times\mathcal{X}}$ and let

$$\Theta_K \equiv \left\{ \theta : \sum_{e\in\mathcal{E}} \theta_k^{e,x} = 1\,\forall(k,x) \text{ and } \theta_k^{e,x} \geq 0\,\forall(e,k,x) \right\}.$$

Then, we can formulate the following finite-dimensional LP that corresponds to the ∞-LP in (∞-LP1)–(∞-LP3):

$$\overline{\tau}_K = \max_{\theta\in\Theta_K} \sum_{(k,x)\in\mathcal{K}\times\mathcal{X}} \left\{ \sum_{e:g_e(1)=1} \theta_k^{e,x}\gamma_k^1(x) - \sum_{e:g_e(0)=1} \theta_k^{e,x}\gamma_k^0(x) \right\} \tag{LP1}$$

$$\underline{\tau}_K = \min_{\theta\in\Theta_K} \sum_{(k,x)\in\mathcal{K}\times\mathcal{X}} \left\{ \sum_{e:g_e(1)=1} \theta_k^{e,x}\gamma_k^1(x) - \sum_{e:g_e(0)=1} \theta_k^{e,x}\gamma_k^0(x) \right\} \tag{LP2}$$

subject to

$$\sum_{e:g_e(d)=1} \sum_{k\in\mathcal{K}} \theta_k^{e,x}\delta_k^d(z,x) = p(1,d|z,x) \qquad \forall(d,z,x) \in \{0,1\}^2 \times \mathcal{X}. \tag{LP3}$$

This LP is computationally very easy to solve using standard algorithms, such as the simplex algorithm; conditional on x, when $K = 50$ and $\dim(\theta) = 204$, it takes only around 10 seconds to calculate $\overline{\tau}_K$ and $\underline{\tau}_K$ with moderate computing power. The important remaining question is how to choose K in practice. We discuss this issue in Section 7. Finally, it is worth noting that, extending Proposition 4 in Mogstad et al. (2018), we may exactly calculate $\overline{\tau}$ and $\underline{\tau}$ (i.e., $\overline{\tau} = \overline{\tau}_K$ and $\underline{\tau} = \underline{\tau}_K$) under the assumptions that (i) the weight function $w_d(u, z, x)$ is piece-wise constant in u and (ii) the constant spline that provides the best mean squared error approximation of $q(e|u, x)$ satisfies all the maintained assumptions (possibly including the identifying assumptions introduced later) that $q(e|u, x)$ itself satisfies; see Mogstad et al. (2018) for details.

5 General Analysis

Now we generalize the analysis in Sections 2–4 to incorporate additional exogenous variables other than the instrument Z that researchers may be equipped with. We show that these variables are fruitful for narrowing bounds on the target parameter. This is the first paper that introduces this type of variable in the MTE framework. This is also the first paper that shows the usefulness of these variables without necessarily combining them with assumptions related to rank similarity or rank invariance.

Let $W \in \mathcal{W}$ be such an exogenous variable. We assume that W is discrete. We show that even binary variation in W can be useful in improving the bounds. We modify our maintained assumptions to consider two different scenarios related to W: (a) W directly affects Y but not D and (b) W directly affects both Y and D. Let $Y(d, w)$ be the extended counterfactual outcome of Y given (d, w).

Assumption SEL$_W$. *(a) Assumption SEL; (b)* $D = 1\{U \leq P(Z, X, W)\}$ *where* $P(Z, X, W) \equiv \Pr[D = 1|Z, X, W]$.

Assumption EX$_W$. *(a)* $(Y(d, w), D(z)) \perp (Z, W)|X$; *(b)* $(Y(d, w), D(z, w)) \perp (Z, W)|X$.

Case (a) is where W is a reversely excluded exogenous variable, which we call *reverse IV*. This type of exogenous variables was considered by Vytlacil and Yildiz (2007), Shaikh and Vytlacil (2011), and Balat and Han (2018). However, unlike those studies, we exploit W without rank similarity or rank invariance. In Case (b), we show that a reverse IV is not necessary, and W can be present in the selection equation. This type of exogenous variables was considered by Mourifié (2015), Han and Vytlacil (2017), Vuong and Xu (2017), and Han and Lee (2019), but again, unlike these papers, we do not necessarily assume rank similarity or rank invariance. Below, we combine the existence of W (for both scenarios) with assumptions that are related to rank similarity. Another distinct feature of our approach in comparison to the prior studies is that we consider a broad class of the generalized LATEs as our target parameter, including the ATE considered in those studies.

In what follows, we modify the linear programming framework from Sections 2 and 3 to reflect Assumptions SEL_W and EX_W. For notational simplicity, we focus on Case (a) here; it is straightforward to draw analogous results for Case (b). With the existence of W, the MTR is defined as

$$\begin{aligned} m_d(u,w,x) &\equiv E[Y(d,w)|U=u,X=x] \\ &= \Pr[Y(d,w)=1|U=u,X=x], \end{aligned}$$

where W does not appear as a conditioning variable due to Assumption EX_W(a). Then, the target parameter can be expressed as

$$\tau = E[\tau_1(Z,W,X) - \tau_0(Z,W,X)],$$

where

$$\tau_d(z,w,x) = \int m_d(u,w,x)w_d(u,z,x)du.$$

Note that the weight $w_d(u,z,x)$ is not a function of w due to Assumption SEL_W(a).[5] Now, consider a mapping

$$(d,w) \mapsto y,$$

which is coded in the value e of $\epsilon \in \mathcal{E}$ where $|\mathcal{E}| = 16$ (redefining the variable ϵ introduced in Section 3). Conveniently, let $\mathcal{E} \equiv \{1,2,...,16\}$; Table 1 lists all 16 maps. Equivalently, define

$$y = g_e(d,w),$$

which implies

$$Y(d,w) = g_\epsilon(d,w) \tag{5.1}$$

and $Y = g_\epsilon(D,W)$. By (5.1) and Assumption SEL_W(a), Assumption EX_W(a) can be equivalently stated as $(\epsilon,U) \perp (Z,W)|X$. Define the probability mass function of ϵ conditional on (U,X) as

$$q(e|u,x) \equiv \Pr[\epsilon = e|U=u,X=x] = \Pr[\epsilon = e|U=u,X=x,W=w].$$

[5] Apparently, with Assumption SEL_W(b), the weight will be written as $w_d^*(u,z,w,x)$ since the propensity score is a function of W.

ϵ	d	w	$Y(d,w)$	ϵ	d	w	$Y(d,w)$	ϵ	d	w	$Y(d,w)$	ϵ	d	w	$Y(d,w)$
1	0	0	0	5	0	0	0	9	0	0	0	13	0	0	0
	0	1	0		0	1	0		0	1	0		0	1	0
	1	0	0		1	0	1		1	0	0		1	0	1
	1	1	0		1	1	0		1	1	1		1	1	1
2	0	0	1	6	0	0	1	10	0	0	1	14	0	0	1
	0	1	0		0	1	0		0	1	0		0	1	0
	1	0	0		1	0	1		1	0	0		1	0	1
	1	1	0		1	1	0		1	1	1		1	1	1
3	0	0	0	7	0	0	0	11	0	0	0	15	0	0	0
	0	1	1		0	1	1		0	1	1		0	1	1
	1	0	0		1	0	1		1	0	0		1	0	1
	1	1	0		1	1	0		1	1	1		1	1	1
4	0	0	1	8	0	0	1	12	0	0	1	16	0	0	1
	0	1	1		0	1	1		0	1	1		0	1	1
	1	0	0		1	0	1		1	0	0		1	0	1
	1	1	0		1	1	0		1	1	1		1	1	1

Table 1: All Possible Maps from (d, w) to y

Then, the MTR can be expressed as

$$m_d(u, w, x) = \sum_{e \in \mathcal{E}: g_e(d,w)=1} q(e|u, x) \tag{5.2}$$

and the target parameter as

$$\tau = \int E[\sum_{e:g_e(1,W)=1} q(e|u, X)w_1(u, Z, X)]du - \int E[\sum_{e:g_e(0,W)=1} q(e|u, X)w_0(u, Z, X)]du. \tag{5.3}$$

Recall, the sieve approximation of the conditional probability is given by $q(e|u, x) \approx \sum_{k \in \mathcal{K}} \theta_k^{e,x} b_k(u)$. Then, for $q \in \mathcal{Q}_K$, by (5.3), we have $\tau = E[\tau_1(Z, W, X)] - E[\tau_0(Z, W, X)]$ with

$$E[\tau_d(Z, W, X)] = \sum_{(w,x) \in \mathcal{W} \times \mathcal{X}} \sum_{e:g_e(d,w)=1} \sum_{k \in \mathcal{K}} \theta_k^{e,x} \gamma_k^d(w, x),$$

where $\gamma_k^d(w, x) \equiv \sum_{z \in \{0,1\}} p(z, w, x) \int b_k(u)w_d(u, z, x)du$ and $p(z, w, x) \equiv \Pr[Z = z, W = w, X = x]$. In terms of the data distribution, we can derive, e.g.,

$$\begin{aligned} p(y,1|z,w,x) &= \Pr[Y(1,w)=y, U \le P(z,x)|X=x] \\ &= \int_0^{P(z,x)} \Pr[Y(1,w)=y|U=u, X=x]du \\ &= \sum_{e:g_e(1,w)=y} \int_0^{P(z,x)} q(e|u,x)du \\ &= \sum_{e:g_e(1,w)=y} \sum_{k\in\mathcal{K}} \theta_k^{e,x}\delta_k^1(z,x), \end{aligned}$$

where $\delta_k^d(z,x) \equiv \int_{\mathcal{U}_{z,x}^d} b_k(u)du$, as in the baseline case.

This modification yields a modified LP:

$$\overline{\tau} = \max_{\theta\in\Theta_K} \sum_{(k,w,x)\in\mathcal{K}\times\mathcal{W}\times\mathcal{X}} \left\{ \sum_{e:g_e(1,w)=1} \theta_k^{e,x}\gamma_k^1(w,x) - \sum_{e:g_e(0,w)=1} \theta_k^{e,x}\gamma_k^0(w,x) \right\} \tag{LP$_W$1}$$

$$\underline{\tau} = \min_{\theta\in\Theta_K} \sum_{(k,w,x)\in\mathcal{K}\times\mathcal{W}\times\mathcal{X}} \left\{ \sum_{e:g_e(1,w)=1} \theta_k^{e,x}\gamma_k^1(w,x) - \sum_{e:g_e(0,w)=1} \theta_k^{e,x}\gamma_k^0(w,x) \right\} \tag{LP$_W$2}$$

subject to

$$\sum_{e:g_e(d,w)=1} \sum_{k\in\mathcal{K}} \theta_k^{e,x}\delta_k^d(z,x) = p(1,d|z,w,x) \qquad \forall(d,z,w,x)\in\{0,1\}^2\times\mathcal{W}\times\mathcal{X}. \tag{LP$_W$3}$$

Although omitted in the paper, similar modification can be made for Case (b), i.e., under Assumptions SEL$_W$(b) and EX$_W$(b). Since the number of maps increases to 16 instead of four of the baseline case, the dimension of the decision variable θ in the LP (LP$_W$1)–(LP$_W$3) is four times larger than that in the baseline. For example, assuming binary W and setting $K=50$, we have $dim(\theta)=816=4\times 204$. Still, it takes only about 13 seconds to solve the program.

We argue that in either Cases (a) or (b), the variation from W is in fact helpful in narrowing the bounds $[\underline{\tau},\overline{\tau}]$ as long as W is a relevant variable. For the remainder of Section 6, we assume $W\in\{0,1\}$ and suppress "conditional on X" for simplicity, unless otherwise noted.

Assumption R. *(i)* $\Pr[Y(d,w)\neq Y(d,w')]>0$ *for some* d *and* $w\neq w'$*; (ii) either (a)* $P(z)>0$ *for all* z *or (b)* $P(z,w)>0$ *for all* z,w*.*

Theorem 5.1. *Under Assumptions SEL$_W$, EX$_W$, and R, the variation of W poses non-redundant constraints on $\theta\in\Theta_K$ in the LP* (LP$_W$1)*–*(LP$_W$3) *(suppressing x).*

Assumption R(i) is a relevance condition for W in determining Y. Heuristically, the improvement occurs because, with R(i), the constraint matrix (i.e., the matrix multiplied to the vector θ in (LP_W3)) has greater rank with the variation of W than without. See the proof of the theorem for a formal argument. Note that non-redundant constraints on θ do not always guarantee an improvement of the bounds in (LP_W1)–(LP_W3), because these constraints may still be non-binding. Nevertheless, non-redundancy is a necessary condition for the improvement.

6 Possible Identifying Assumptions

Bounds on the target parameter are generally uninformative in the absence of additional assumptions besides Assumptions SEL and EX. This is because a binary instrument has no extrapolative power for general non-compliers, e.g., always-takers and never-takers, but only identifies the effect for compliers. Prior studies have tried to overcome this challenge by imposing shape restrictions on the MTE (Cornelissen et al. (2016); Brinch et al. (2017); Kowalski (2020)), although these restrictions are not always empirically justified. Evidently, it would be useful to provide researchers with a larger variety of assumptions so that it is easier to find justifiable assumptions that suit their specific examples.

In Section 5, the existence of W (Assumptions SEL_W and EX_W) is shown to be one useful source for extrapolation. In this section, we propose identifying assumptions that can be incorporated within our framework and that help shrink the bounds on the target parameters. The shape restrictions employed in the literature can be used within our framework. We also propose other assumptions that have not been previously used in the LATE extrapolation. These assumptions can be incorporated as additional equality and inequality restrictions in the linear programming: Given the LP (∞-LP1)–(∞-LP3), identifying assumptions can be imposed by appending

$$R_1 q = a_1, \tag{∞-LP4}$$

$$R_2 q \leq a_2, \tag{∞-LP5}$$

where R_1 and R_2 are linear operators on $\mathcal{Q}$ that correspond to equality and inequality constraints, respectively, and a_1 and a_2 are some vectors.

When an assumption violates the true data-generating process, then the identified set will be empty. This corresponds to the situation where the LP does not have a feasible solution. When we reflect sampling errors, this corresponds to the case where the confidence set is empty.[6]

6.1 Uniformity

Researchers may be willing to restrict the degree of treatment heterogeneity to yield informative bounds. This restriction has not been used before in the context of the MTE framework. This restriction may be combined with the assumptions related to the existence of W (Assumptions SEL$_W$, EX$_W$, and R). We suppress the conditioning on X throughout this subsection.

Assumption U. *For every* $w \in \mathcal{W}$, $\Pr[Y(1,w) \geq Y(0,w)] = 1$ *or* $\Pr[Y(1,w) \leq Y(0,w)] = 1$*.*

When W is not available at all, this assumption can be understood with $\mathcal{W}$ being degenerate. The following assumption is stronger than Assumption U.

Assumption U*. *For every* $w, w' \in \mathcal{W}$, $\Pr[Y(1,w) \geq Y(0,w')] = 1$ *or* $\Pr[Y(1,w) \leq Y(0,w')] = 1$*.*

Note that w and w' may be the same or different, i.e., the uniformity is for all combinations of $(w,w') \in \{(0,0),(1,1),(1,0),(0,1)\}$. Therefore, Assumption U* implies Assumption U. Assumptions U and U* are weaker than the monotone treatment response assumption in Manski (1997) and Manski and Pepper (2000) in that they do not impose the direction of monotonicity. Assumptions U and U* are also closely related to the rank similarity and rank invariance assumptions in the literature (e.g., Chernozhukov and Hansen (2005)). Namely, given a structural model $Y = 1[s(D,W) \geq V_D]$, when Assumption U* is violated, then rank similarity ($F_{V_1|U} = F_{V_0|U}$) cannot hold, and thus rank invariance ($V_1 = V_0$) cannot hold.

Assumptions U and U* can be imposed by "deactivating" relevant maps. For example, suppose $Y(1,w) \geq Y(0,w)$ almost surely for all $w \in \{0,1\}$ under Assumption U. This assumption can be imposed as equality constraints (∞-LP4), i.e., in the form of $R_1 q = a_1$, using the labeling of Table 1:

$$q(3|u) = q(4|u) = q(7|u) = q(8|u) = 0,$$
$$q(2|u) = q(4|u) = q(10|u) = q(12|u) = 0,$$

[6] In order to verify whether the identified set is empty, we need to check whether the feasible set of θ is empty. An efficient way to do this is to identify vertices of the feasible polytope, if any. This process is no simpler than the simplex algorithm that we use to solve the LP. Therefore, we recommend that one first solves the LP and check if infeasibility is reported.

respectively, corresponding for $w = 1$ and $w = 0$. Therefore, the corresponding $\theta_k^e = 0$. Then, the effective dimension of θ will be reduced in (LP$_W$1)–(LP$_W$3) and thus yields narrower bounds. As another example, suppose the following holds almost surely under Assumption U*: $Y(1,1) \geq Y(0,0)$, $Y(1,0) \leq Y(0,1)$, $Y(1,1) \geq Y(0,1)$, and $Y(1,0) \geq Y(0,0)$. These inequalities respectively imply

$$\begin{aligned}
&q(2|u) = q(4|u) = q(6|u) = q(8|u) = 0,\\
&q(5|u) = q(6|u) = q(13|u) = q(14|u) = 0,\\
&q(3|u) = q(4|u) = q(7|u) = q(8|u) = 0,\\
&q(2|u) = q(4|u) = q(10|u) = q(12|u) = 0.
\end{aligned}$$

It is worth mentioning that, in Assumption U (Assumption U*), the direction of monotonicity is allowed to be different for different w ((w, w') pairs). This direction will be identified from the data. Specifically, the direction can be automatically determined from the LP by inspecting whether the LP has a feasible solution; when wrong maps are removed, there is no feasible solution. Note that this result holds regardless of the existence of W. It is easy to see that the direction of the monotonicity coincides with the sign of the ATE. Previous work has discussed the role of the rank similarity assumption on determining the sign of the ATE (Bhattacharya et al. (2008); Shaikh and Vytlacil (2011); Han (2020b)), and the result above shows that Assumptions U and U* play a similar role in the linear programming approach. In the next two subsections, we suppress W for simplicity.

6.2 Direction of Endogeneity

In some applications, researchers are relatively confident about the direction of treatment endogeneity. The idea of imposing the direction of the selection bias as an identifying assumption appears in Manski and Pepper (2000), who introduce monotone treatment selection (MTS), in addition to the monotone treatment response assumption mentioned above.

Assumption MTS. *$E[Y(d)|D = 1, X = x] \geq E[Y(d)|D = 0, X = x]$ for $d \in \{0, 1\}$ and $x \in \mathcal{X}$.*

Under our framework, this assumption can be imposed in the form of $R_2 q \leq a_2$. To see this, Assumption MTS is equivalent to

$$\sum_{e:g_e(d)=1} E\left[\int_{P(Z,X)}^{1} q(e|u)du - \int_{0}^{P(Z,X)} q(e|u)du \,\middle|\, X = x\right] \leq 0$$

for all $d, x \in \{0,1\} \times \mathcal{X}$. As is clear from this expression, Assumption MTS imposes restrictions on the joint distribution of (ϵ, U).

6.3 Shape Restrictions

It is straightforward to incorporate the shape restrictions on the MTR or MTE function introduced in the literature. They can be imposed via constraints on θ.

Assumption M. *For $x \in \mathcal{X}$, $m_d(u,x)$ is weakly increasing in $u \in [0,1]$.*

Assumption C. *For $x \in \mathcal{X}$, $m_d(u,x)$ is weakly concave in $u \in [0,1]$.*

Assumption M appears in Brinch et al. (2017) and Mogstad et al. (2018) and Assumption C appears in Mogstad et al. (2018). These assumptions can be imposed as inequality constraints (∞-LP4), i.e., in the form of $R_2 q \leq a_2$. For implications on the finite-dimensional LP (LP1)–(LP3), recall that for $q \in \mathcal{Q}_K$, the MTR satisfies

$$m_d(u,x) = \sum_{e:g_e(d)=1} q(e|u,x) = \sum_{k \in \mathcal{K}} \sum_{e:g_e(d)=1} \theta_k^{e,x} b_k(u).$$

According to the property of the Bernstein polynomial, Assumption M implies that $\sum_{e:g_e(d)=1} \theta_k^{e,x}$ is weakly increasing in k, i.e.,

$$\sum_{e:g_e(d)=1} \theta_1^{e,x} \leq \sum_{e:g_e(d)=1} \theta_2^{e,x} \leq \cdots \leq \sum_{e:g_e(d)=1} \theta_K^{e,x}.$$

Assumption C implies that

$$\sum_{e:g_e(d)=1} \theta_k^{e,x} - \sum_{e:g_e(d)=1} 2\theta_{k+1}^{e,x} + \sum_{e:g_e(d)=1} \theta_{k+2}^{e,x} \leq 0 \qquad \text{for } k = 0, ..., K-2.$$

One can obtain analogous assumptions and their implications in the presence of W.

Another shape restriction introduced in the literature is separability. Although it is not particularly appealing with binary Y, if one is willing to assume a separable model for $m_d(u,x) = \Pr[Y(d) = 1|U = u, X = x] = m_{1d}(x) + m_{2d}(u)$, then such a structure can be imposed on θ.

7 Simulation

This section provides numerical results to illustrate our theoretical framework and to show the role of different identifying assumptions in improving bounds on the target parameters.

For target parameters, we consider the ATE and the LATEs for always-takers (LATE-AT), never-takers (LATE-NT), and compliers (LATE-C). We calculate the bounds on them based only on the information from the data and then show how additional assumptions on the existence of additional exogenous variables, uniformity, and shape restrictions tighten the bounds.

7.1 Data-Generating Process

We generate the observables (Y, D, Z, X, W) from the following data-generating process (DGP). We assume that W is a reverse IV, i.e., we maintain Assumptions SEL_W(a) and EX_W(a). We allow covariate X to be endogenous. All the variables are set to be binary with $\Pr[Z = 1] = 0.5$, $\Pr[X = 1] = 0.6$ and $\Pr[W = 1] = 0.4$. The treatment D is determined by Z and X through the threshold crossing model specified in Assumption SEL_W(a), where the propensity scores $P(z, x)$ are specified as follows: $P(0, 0) = 0.1$, $P(1, 0) = 0.4$, $P(0, 1) = 0.4$, and $P(1, 1) = 0.7$. The outcome Y is generated from (D, X, W) through $Y = DY_1 + (1-D)Y_0$ where

$$Y_d = 1\left[m_d(U, X, W) \geq \epsilon\right] \tag{7.1}$$

and the MTR functions are defined as

$$\begin{aligned}
m_0(u,0,0) &= 0.02b_0^4(u) + 0.08b_1^4(u) + 0.14b_2^4(u) + 0.20b_3^4(u) + 0.21b_4^4(u)\\
m_1(u,0,0) &= 0.12b_0^4(u) + 0.28b_1^4(u) + 0.44b_2^4(u) + 0.52b_3^4(u) + 0.54b_4^4(u)\\
m_0(u,1,0) &= 0.22b_0^4(u) + 0.48b_1^4(u) + 0.64b_2^4(u) + 0.72b_3^4(u) + 0.74b_4^4(u)\\
m_1(u,1,0) &= 0.44b_0^4(u) + 0.71b_1^4(u) + 0.88b_2^4(u) + 0.97b_3^4(u) + 0.99b_4^4(u)\\
m_0(u,0,1) &= 0.10b_0^4(u) + 0.25b_1^4(u) + 0.30b_2^4(u) + 0.32b_3^4(u) + 0.33b_4^4(u)\\
m_1(u,0,1) &= 0.20b_0^4(u) + 0.45b_1^4(u) + 0.60b_2^4(u) + 0.65b_3^4(u) + 0.66b_4^4(u)\\
m_0(u,1,1) &= 0.30b_0^4(u) + 0.60b_1^4(u) + 0.80b_2^4(u) + 0.85b_3^4(u) + 0.86b_4^4(u)\\
m_1(u,1,1) &= 0.35b_0^4(u) + 0.70b_1^4(u) + 0.92b_2^4(u) + 0.99b_3^4(u) + 1.00b_4^4(u)
\end{aligned}$$

where b_k^K stands for the k-th basis function in the Bernstein approximation of degree K. These MTR functions are chosen to be consistent with Assumptions M and C, i.e., to be positively monotone and weakly concave in u for all $(d, x, w) \in \{0, 1\}^3$. Also, the DGP in (7.1) satisfies Assumption U* because ϵ does not depend on $d = 0, 1$ and the MTR functions satisfy $m_1(u, x, w) > m_0(u, x, w)$ for all $(d, x, w) \in \{0, 1\}^3$. Following the second example in Section 6.1, the DGP satisfy the following uniform order for the counterfactual outcomes $Y(d, w)$: $Y(1, 1) \geq Y(0, 1) \geq Y(1, 0) \geq Y(0, 0)$ a.s. We generate a sample containing

1,000,000 observations and choose $K = 50$. We choose the large sample size to mimic the population. Our choice of K is discussed below. The number of unknown parameters θ in the linear programming is equal to $\dim(\theta) = |\mathcal{E}| \times |\mathcal{X}| \times (K+1)$.

7.2 Bounds on Target Parameters under Different Assumptions

7.2.1 ATE

Figure 1 contains the bounds on the ATE under different assumptions. The true ATE value is 0.21, depicted as the solid red line in the figure. First, the worst-case bounds on the ATE with no additional assumptions (and without using variation from W) are $[-0.25, 0.45]$. Since the mappings do not involve W, we have $|\mathcal{E}| = 4$, and the linear programming is solved with $\dim(\theta) = |\mathcal{E}| \times |\mathcal{X}| \times (K+1) = 4 \times 2 \times 51 = 408$.

For comparison, we calculate the bounds that incorporate the existence of W. We build up the target parameters with mappings involving W and use data distribution conditional on $W = 0$ and $W = 1$ as the constraints. Using constraints conditional on different values of W allows us to fully exploit the variations from W; see (LP_W3). As shown in the setup with W (Section 5), we have $|\mathcal{E}| = 16$, which gives $\dim(\theta) = |\mathcal{E}| \times |\mathcal{X}| \times (K+1) = 16 \times 2 \times 51 = 1,632$. The resulting bounds are depicted in the dotted greenish-blue line. When the variation from W is exploited, the bounds on the ATE are $[-0.24, 0.44]$, which is slightly narrower than without using W. This result is consistent with our theoretical finding presented in Theorem 5.1 that W can help tighten the bounds as long as it is a relevant variable. Nonetheless, these worst-case bounds are not that informative, e.g., they do not determine the sign of the ATE.

Next, we impose the uniformity assumption without W (Assumption U) and with W (Assumption U*). First, under Assumption U, the bounds on the ATE are tightened as some mappings occur with probability zero reducing the dimension of θ. As mentioned in Section 6.1, the direction of monotonicity in Assumption U (i.e., which mapping does not occur) is determined by the LPs. We solve the LPs with different directions imposed, then choose the one with a feasible solution. This means that the corresponding direction of monotonicity is consistent with the DGP. Under Assumption U, we obtain a narrower bound $[0.06, 0.45]$. Second, under Assumption U*, the bounds become $[0.06, 0.33]$. In Figure 1, these bounds under Assumptions U and U* are depicted as violet and green dashed lines, respectively. Both sets of bounds identify the sign of the ATE, consistent with the theoretical discussion. While their lower bounds coincide, Assumption U* yields a lower upper bound compared to Assumption U.

Next, we impose the shape restrictions (Assumptions M and C). As discussed in Section

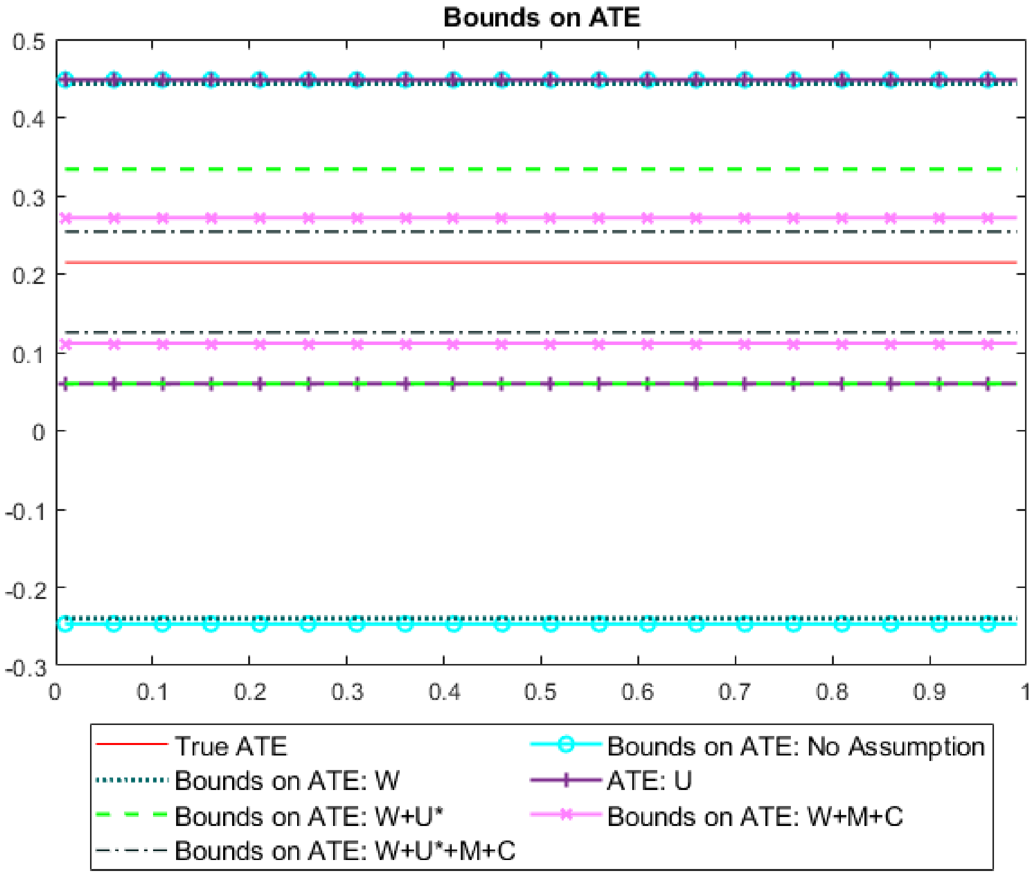

Figure 1: Bounds on the ATE under Different Assumptions

6.3, these assumptions can be easily incorporated in the linear programming by directly imposing inequality constraints on θ. Under these assumptions (and the existence of W), the bounds on the ATE shrink to $[0.11, 0.27]$, which is displayed with the pink line in Figure 1. We find that shape restrictions are powerful assumptions and yield narrower bounds compared to those with Assumption U*. They function differently in the linear programming: unlike the uniformity assumption, which maintains the ranking of individuals across counterfactual groups, shape restrictions directly control the MTR functions. Finally, the dash-dotted black line in Figure 1 shows the bounds on the ATE under the uniformity assumption and the shape restrictions. These assumptions all together yield the narrowest bounds, $[0.13, 0.25]$, for the true ATE, 0.21.

7.2.2 Generalized LATEs

Next, we construct bounds on the generalized LATEs. The original definition of the LATE is the ATE for compliers (C). Researchers may also have interests in other local treatment effects. We consider two other parameters—LATEs for always-takers (AT) and never-takers (NT). Figure 2 displays the bounds on the LATE-AT, LATE-C, and LATE-NT under different assumptions. This analysis is analogous to that with the ATE. Since the covariate X affects the decision of compliance, to avoid confusion in the definition of the compliance groups, we instead establish bounds on the LATEs conditional on X. We draw the conditional MTE functions with solid red lines in both panels as a reference.

The feature that there exists no defiers in the DGP is known. When there is no defier, the LATE-C is point identified, which has an analytical expression of the two-stage least squares estimand. As a confirmation exercise, we numerically calculate the LATE-C using the linear programming, which yields point estimates as shown in Figure 2. The true LATE-Cs conditional on $X = 0$ and $X = 1$ are equal to 0.21 and 0.22, respectively. Regardless of assumptions imposed, the estimates remain close to the true values throughout.

The true values of the conditional LATE-AT and the LATE-NT are 0.15 and 0.28 when $X = 0$ and 0.14 and 0.25 when $X = 1$. First, as before, we consider the worst-case bounds where the existence of W is ignored versus where W is taken into account. Without W, we get the bounds $[-0.71, 0.24]$ and $[-0.28, 0.72]$ on the LATE-AT and the LATE-NT conditional on $X = 0$, and $[-0.48, 0.52]$ and $[-0.56, 0.43]$ conditional on $X = 1$; with W, we get the bounds $[-0.62, 0.2]$ and $[-0.28, 0.7]$ on the LATE-AT and the LATE-NT conditional on $X = 0$, and $[-0.48, 0.5]$ and $[-0.55, 0.41]$ conditional on $X = 1$. The upper bounds with W are lower than the ones without W, although the gain is not substantial. For the lower bounds, the one on the LATE-AT conditional on $X = 0$ is significantly higher with W than without W, and all the other ones have negligible differences with and without W.

We then apply Assumptions U and U*. Under Assumption U, the bounds on the LATE-AT and the LATE-NT turn to $[0, 0.24]$ and $[0, 0.72]$ conditional on $X = 0$, and $[0, 0.52]$ and $[0, 0.43]$ conditional on $X = 1$; when W is used and Assumption U* is applied, the bounds shrink to $[0, 0.18]$ and $[0, 0.47]$ conditional on $X = 0$, and $[0, 0.36]$ and $[0, 0.35]$ conditional on $X = 1$. As before, Assumptions U and U* determine the sign of the effects.

When the shape restrictions are imposed instead, the bounds on the LATE-AT and the LATE-NT were improved to $[0.11, 0.17]$ and $[0.03, 0.3]$ conditional on $X = 0$, and $[0.05, 0.31]$ and $[0.15, 0.31]$ conditional on $X = 1$. Under Assumption U* combined with the shape restrictions, we get the narrowest bounds of $[0.11, 0.15]$ and $[0.04, 0.3]$ conditional on $X = 0$, and $[0.08, 0.26]$ and $[0.15, 0.31]$ conditional on $X = 1$.

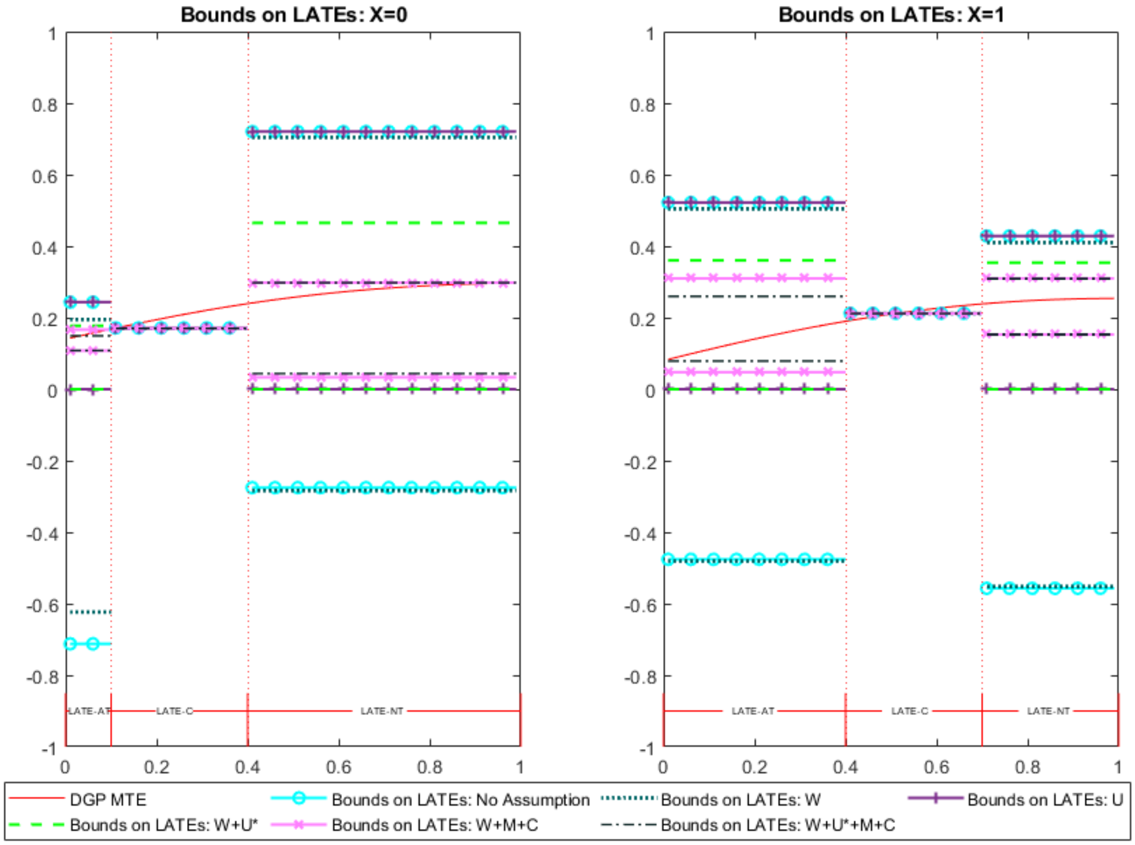

Figure 2: Bounds on the LATEs under Different Assumptions

7.3 The Choice of K

As a tuning parameter in the LP, we need to choose the order of Bernstein polynomials, K. In general, K should be chosen based on the sample size and the smoothness of the function to be approximated, in our case, $q(\cdot)$. The choice of the sieve dimension or more generally, regularization parameters, is a difficult question (Chen (2007)) and developing data-driven procedure is a subject of on-going research in various nonparametric contexts of point identification; see, e.g., Chen and Christensen (2015) and Han (2020a). In this partial identification setup, we propose the following heuristic and conservative approach, which is in spirit consistent with the very motivation of partial identification.

First, we do not want to claim any prior knowledge about the smoothness of $q(\cdot)$ because it is the distribution of a latent variable. Because K determines the dimension of unknown parameter θ in the linear programming, the width of the bounds tends to increase with K. At the same time, the computational burden increases with K. One interesting numerical finding is that, when K is sufficiently large, the increase of the width slows down and the bounds become stable. This suggests that we may be able to conservatively choose K that acknowledges our lack of knowledge of the smoothness but, at the same time, produces a reasonable computational task for the linear programming.

To illustrate this point, we consider the conditional MTE and ATE as the target parameters and show how their bounds change as we increase K. We consider the MTE because it is a fundamental parameter that generates other target parameters, and hence, it is important to understand the sensitivity of its bounds to K. Figures 3 and 4 show the evolution of the bounds on the MTE and the ATE as K grows. When $K = 5$, the bounds are narrow. Although it may be tempting to choose this value of K, this attempt should be avoided as it may be subject to the misspecification of smoothness. When K increases beyond 30, the bounds start to converge and become stable. We choose $K = 50$, and this is the choice we made in our previous numerical exercises.[7]

As discussed in Section B.1 in the Appendix, it is worth mentioning that the bounds on the MTE are point-wise sharp but *not* uniformly sharp. The graph for the MTE bounds are drawn by calculating the point-wise sharp bounds on MTE at each point of u (after properly discretizing it) and then connecting them. Therefore, these bounds should *not* be viewed as uniformly sharp bounds. Nonetheless, this graph is still useful for the purpose of our illustration. Given the current DGP, we find that there are no uniformly sharp bounds for the MTE.

[7] Note that with larger K, some LP solvers would ignore coefficients with negligible (e.g., 10^{-13}) values that cause a large range of magnitude in the coefficient matrix. It may be recommended to simultaneously rescale a column and a row to achieve a smaller range in the coefficients.

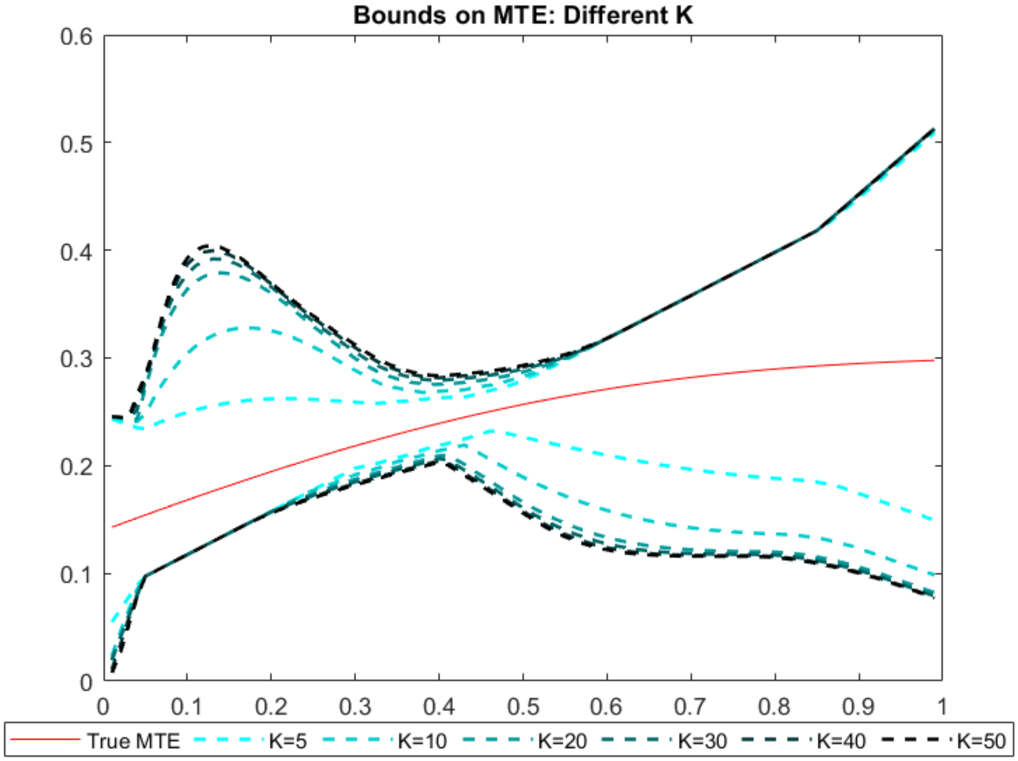

Figure 3: Bounds on MTE with Different K

8 Empirical Application

It is widely recognized in the empirical literature that health insurance coverage can be an essential factor for the utilization of medical services (Hurd and McGarry (1997); Dunlop et al. (2002); Finkelstein et al. (2012); Taubman et al. (2014)). Prior studies on this topic typically make use of parametric econometric models for the analysis. In their application, Han and Lee (2019) relax this common approach by introducing a semiparametric bivariate probit model to measure the average effect of insurance coverage on patients' medical visits. By applying our theoretical framework of partial identification, we further relax the parametric and semiparametric structures used in these studies. More importantly, we try to understand how much we can learn about the effect of insurance that is utilized through various counterfactual policies by learning the effect of different compliance groups.

We use the 2010 wave of the Medical Expenditure Panel Survey (MEPS) and focus on all the medical visits in January 2010. The sample is restricted to contain individuals aged between 25 and 64 and exclude those who had any kind of federal or state insurance in 2010. The outcome Y is a binary variable indicating whether or not an individual has visited a doctor's office; the treatment D is whether an individual has private insurance. We choose

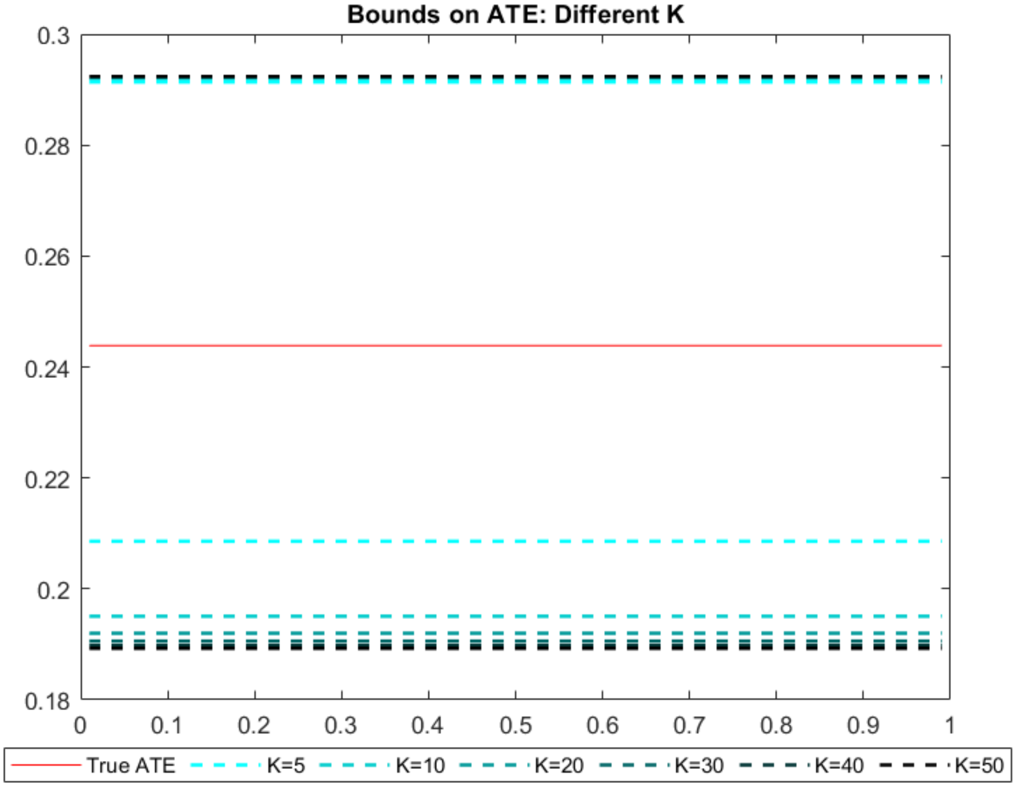

Figure 4: Bounds on ATE with Different K

whether a firm has multiple locations as the binary instrument Z. This IV reflects the size of the firm, and larger firms are more likely to provide fringe benefits, including health insurance. On the other hand, the number of branches of a firm does not directly affect employee decisions about medical visits. To justify the IV, self-employed individuals are excluded. For potentially endogenous covariates X, we include the age being 45 and older, gender, income above median, and health condition. Lastly, for an exogenous covariate W, we use the percentage of workers who are provided with paid sick leave benefits within each industry. Following Han and Lee (2019), we assume W satisfies Assumptions SEL_W(b) and EX_W(b), as X is controlled. We construct a categorical variable such that $W = 0$ for less than 50%, $W = 1$ for between 50–80%, and $W = 2$ for above 80%. Table 2 summarizes the observables.

Table 2: Summary Statistics

	Variable	Mean	S.D	Min	Max
Y	Whether or not visit doctors	0.18	0.39	0	1
D	Whether or not have insurance	0.66	0.47	0	1
Z	Firm has multiple locations	0.68	0.47	0	1
X	Age above 45	0.41	0.49	0	1
	Gender	0.50	0.50	0	1
	Income above median	0.50	0.50	0	1
	Good health	0.36	0.48	0	1
W	Pay sick leave provision	1.25	0.73	0	2
	Number of observations = 7,555				

First, as a benchmark, we report that the LATE-C estimate calculated via our linear programming approach is equal to a singleton of 0.17, which is in fact identical to the 2SLS estimate we separately calculate. In what follows, we extrapolate this LATE beyond the complier group to the ATE. The presence of covariates reduces the effective sample size and thus leads to larger sampling errors in estimating the p of the ∞-LP (∞-LP1)–(∞-LP3). This may create inconsistencies in the set of equality constraints (∞-LP3), resulting in no feasible solution. This is in fact what happens in this application. To resolve this estimation problem, we introduce a slackness parameter η and modify (∞-LP3) so that, with some slackness, it satisfies

$$\|R_0 q - p\| \leq \eta. \tag{8.1}$$

A similarly modified constraint can then be followed in the finite-dimensional LP after approximation, as well as by combining (∞-LP4)–(∞-LP5). The appropriate value of η should depend on the sample size, the dimension of covariates, and the dimension of the unknown parameter θ. To explain the latter, as K increases, the dimension of θ (i.e., unknowns) increases, while the number of constraints (i.e., simultaneous equations for the unknowns) is fixed. Therefore, as K increases, the chance that the LP does not have a feasible solution would decrease. Based on the method discussed in the previous section, we set $K = 50$ in this application.

We calculate worst-case bounds on the ATE, as well as bounds after imposing Assumptions U and M and after using covariate W. Under Assumption U, the data rules out the possibility that $Y(0) > Y(1)$, indicating that individuals with private insurance are more likely to visit a doctor. Assumption M imposes that the MTR function is weakly increasing in $U = u$. Usually, U is interpreted as the latent cost of obtaining treatment. Kowalski (2020) interpreted U as eligibility in a similar setup for Medicaid insurance. The eligibility for Medicaid is related to income level and age. In our setup, because the treatment is having the private insurance, we interpret the eligibility as the health status, which is reflected in the premium. Interpreting U as a latent cost (e.g., premium) of getting private insurance, Assumption M states that the chance of making a medical visit (with or without insurance) increases for those with higher cost. This is a reasonable assumption given that sicker individuals typically face higher insurance costs and also visit doctors more often. We choose the slackness parameter η to be 0.05 under no assumption and Assumption U and 0.07 when Assumption M is added. When W is used, we choose η to be 0.08 under no assumption and 0.1 with Assumption M.

The bounds on the ATE are shown in Figure 5. The worst-case bound on the ATE equals $[-0.45, 0.37]$. The bounds become $[0.01, 0.37]$ under Assumption U and $[0.06, 0.37]$ under Assumption M. It is interesting to note that the identifying power of the uniformity and the shape restriction is similar in this example. When both Assumption U and Assumption M are imposed, the bounds are further tightened to $[0.07, 0.37]$, although not substantially, indicating that the two assumptions are complementary. Lastly, we see improvements when the variation in W is exploited than when it is not, although the gains are not large.

Next, we consider the always-taker, complier, and never-taker LATEs. We consider these generalized LATEs conditional on $X = x$. Specifically, we focus on the treatment effects for males above age 45, with income below the median and bad health conditions. The results are shown in Table 3 and depicted in Figure 6. The LATE-C is analytically calculated via TSLS.[8] For the LATE-AT and LATE-NT, Assumption U identifies the sign of the effects,

[8] When the alternative constraint (8.1) is used with the slackness parameter, the LATE-C is no longer a singleton.

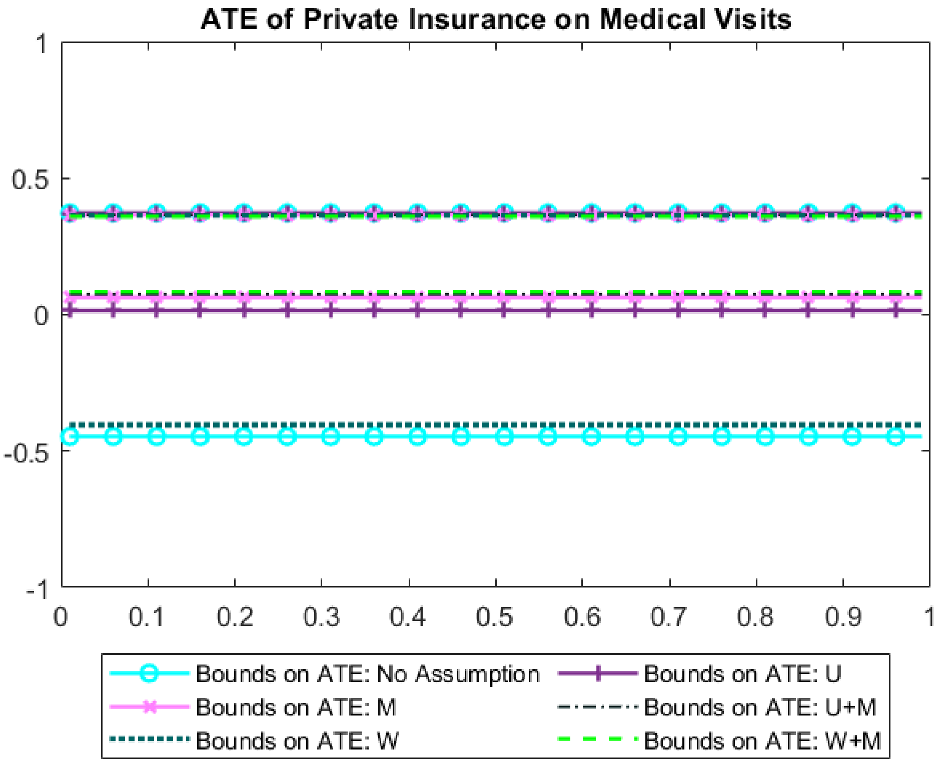

Figure 5: Bounds on the ATE of Private Insurance on Medical Visits

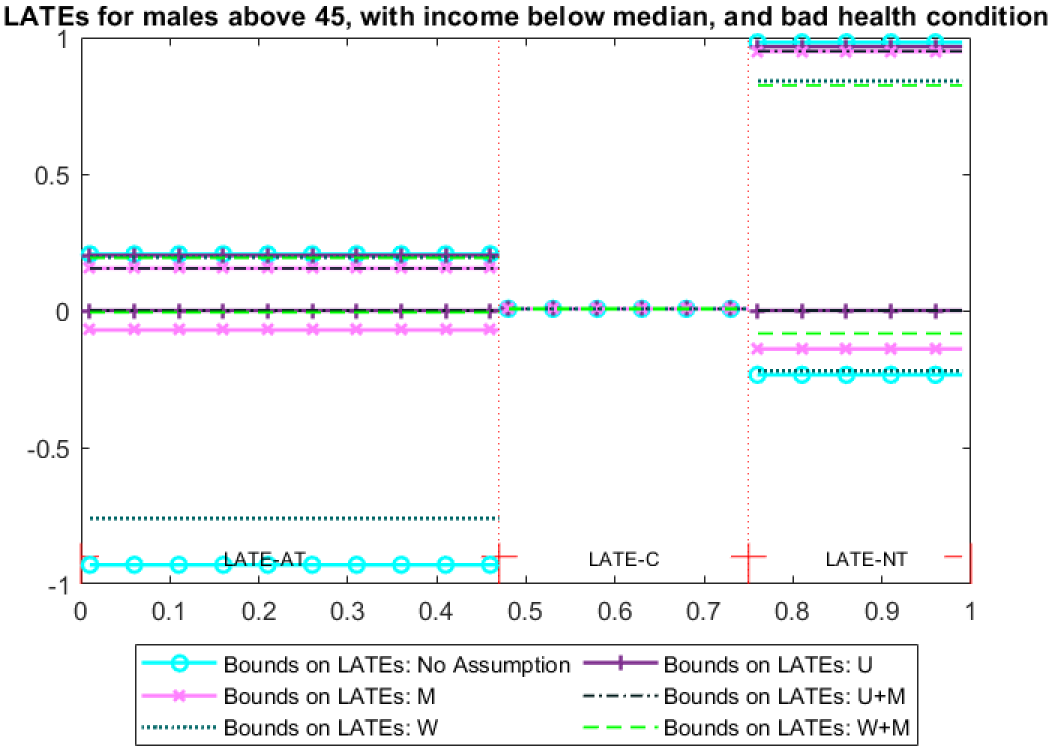

Figure 6: Bounds on the generalized LATEs of Private Insurance on Medical Visits for Male Above 45, with Income Below Median, of Bad Healthiness

Table 3: Estimated Bounds on generalized LATEs for Males Above 45, with Income Below Median, Bad Health Condition

	No Assumption	U	M	U+M	W	M+W
LATE-AT	[-0.93,0.21]	[0,0.20]	[-0.07,0.15]	[0,0.15]	[-0.76,0.20]	[-0.01,0.19]
LATE-C	0.01	0.01	0.01	0.01	0.01	0.01
LATE-NT	[-0.24,0.98]	[0,0.97]	[-0.14,0.95]	[0,0.95]	[-0.22,0.84]	[-0.08,0.82]
Slackness parameter η	0.05	0.05	0.07	0.07	0.08	0.10
Number of observations = 7,555						

and Assumption M nearly identifies it. Using the variation in W mostly improves the bounds compared to the ones without it. From the results, we can conclude that possessing private insurance has the greatest effect on medical visits for never takers, i.e., people who face higher insurance cost. This provides a policy implication that lowering the cost of private insurance is important, because high costs might hinder those with the most need from receiving enough medical services.

A Examples of the Target Parameters

Table 4 contains the list of target parameters. The table is taken from Mogstad et al. (2018).

B More Discussions

B.1 Point-wise and Uniform Sharp Bounds on MTE

In Section 2, we provided some examples of target parameters. The building block for these parameters is the MTE, $m_1(u) - m_0(u)$ (suppressing x). Heckman and Vytlacil (2005) show why this fundamental parameter can be of independent interest. Unlike other target parameters proposed here, we may want to allow the MTE to be a function of u (beyond evaluating it at a fixed u). In this section, we discuss the subtle issue of point-wise and uniform sharp bounds on $\tau_{MTE}(u) \equiv m_1(u) - m_0(u)$ as a function of u.

Suppress X for simplicity. Recall $q(u) \equiv \{q(e|u)\}_{e\in\mathcal{E}}$ and $\mathcal{Q} \equiv \{q(\cdot) : \sum_e q(e|u) = 1\,\forall u \text{ and } q(e|u) \geq 0\,\forall(e,u)\}$. Let $\mathcal{M}$ be the set of MTE functions, i.e.,

$$\mathcal{M} \equiv \left\{m_1(\cdot) - m_0(\cdot) : m_d(\cdot) = E[Y_d|U = \cdot] = \sum_{e\in\mathcal{E}:g_e(d)=1} q(e|\cdot)\,\forall d \in \{0,1\} \text{ for } q(\cdot) \in \mathcal{Q}\right\}.$$

Target Parameters	Expressions	Ranges of u	Weights
			$w_d(u,z,x)$
Average Treatment Effect (ATE)	$E[Y(1)-Y(0)]$	$[0,1]$	$2d-1$
LATE for Compliers (LATE-C) given $x \in \mathcal{X}$	$E\{Y(1)-Y(0) \mid u \in [P(z_0,x),P(z_1,x)]\}$	$[P(z_0,x),P(z_1,x)]$	$(2d-1) \times \frac{1(u \in [P(z_0,x),P(z_1,x)])}{P(z_1,x)-P(z_0,x)}$
LATE for Always-Takers (LATE-AT) given $x \in \mathcal{X}$	$E\{Y(1)-Y(0) \mid u \in [0,P(z_0,x)]\}$	$[0,P(z_0,x)]$	$(2d-1) \times \frac{1(u \in [0,P(z_0,x)])}{P(z_0,x)}$
LATE for Never Takers (LATE-NT) given $x \in \mathcal{X}$	$E\{Y(1)-Y(0) \mid u \in [P(z_1,x),1]\}$	$[P(z_1,x),1]$	$(2d-1) \times \frac{1(u \in [P(z_1,x),1])}{1-P(z_1,x)}$
LATE for $[\underline{u},\overline{u}]$	$E[Y(1)-Y(0) \mid u \in [\underline{u},\overline{u}]]$	$[P(z_0,x),P(z_1,x)]$	$(2d-1) \times \frac{1(u \in [\underline{u},\overline{u}])}{\overline{u}-\underline{u}}$
Marginal Treatment Effect (MTE)*	$E[Y(1)-Y(0) \mid u']$	u'	$(2d-1) \times 1(u=u')$
Policy Relevant Treatment Effect (PRTE) for a new policy (P',Z')	$\frac{E(Y')-E(Y)}{E(D')-E(D)}$	$[0,1]$	$(2d-1) \times \frac{\Pr[u \leq P'(z')]-\Pr[u \leq P'(z)]}{E[P(Z')]-E[P(Z)]}$

* The MTE uses the Dirac measure at u', while the other target parameters use the Lebesgue measure on $[0,1]$.

Table 4: Examples of the Target Parameters

The bounds on $\tau_{MTE} \in \mathcal{M}$ in the ∞-LP are given by using a Dirac delta function as a weight. Therefore, given evaluation point $u \in [0,1]$, (∞-LP1)–(∞-LP3) can be simplified as follows, defining the upper and lower bounds $\overline{\tau}(u)$ and $\underline{\tau}(u)$ (being explicit about the evaluation point) on $\tau_{MTE}(u)$:

$$\overline{\tau}(u) = \sup_{q \in \mathcal{Q}} \sum_{e \in \mathcal{E}: g_e(1)=1} q(e|u) - \sum_{e \in \mathcal{E}: g_e(0)=1} q(e|u) \tag{B.1}$$

$$\underline{\tau}(u) = \inf_{q \in \mathcal{Q}} \sum_{e \in \mathcal{E}: g_e(1)=1} q(e|u) - \sum_{e \in \mathcal{E}: g_e(0)=1} q(e|u) \tag{B.2}$$

subject to

$$\sum_{e: g_e(d)=1} \int_{\mathcal{U}_z^d} q(e|\tilde{u}) d\tilde{u} = p(1,d|z) \qquad \forall (d,z) \in \{0,1\}^2. \tag{B.3}$$

Then, for any fixed $u \in [0,1]$,

$$\underline{\tau}(u) \leq \tau_{MTE}(u) \leq \overline{\tau}(u).$$

We argue that these bounds are point-wise sharp but not necessarily uniformly sharp for $\tau_{MTE}(\cdot)$.[9]

[9] See Firpo and Ridder (2019) for related definitions of point-wise and uniform sharpness.

Definition B.1 (Point-wise Sharpness). *$\overline{\tau}(\cdot)$ and $\underline{\tau}(\cdot)$ are point-wise sharp if, for any $\bar{u} \in [0,1]$, there exist $\overline{\tau}_{MTE,\bar{u}}, \underline{\tau}_{MTE,\bar{u}} \in \mathcal{M}$ such that $\overline{\tau}(\bar{u}) = \overline{\tau}_{MTE,\bar{u}}(\bar{u})$ and $\underline{\tau}(\bar{u}) = \underline{\tau}_{MTE,\bar{u}}(\bar{u})$.*

Theorem B.1. *$\overline{\tau}(\cdot)$ and $\underline{\tau}(\cdot)$ are point-wise sharp bounds on $\tau_{MTE}(\cdot)$.*

The proofs of this and other theorems appear later. Note that point-wise bounds will maintain some properties of an MTE function, but not all. For uniform sharpness, $\overline{\tau}(\cdot)$ and $\underline{\tau}(\cdot)$ themselves have to be MTE functions on $[0,1]$, i.e., $\overline{\tau}(\cdot)$ and $\underline{\tau}(\cdot)$ should be elements in $\mathcal{M}$.

Definition B.2 (Uniform Sharpness). *$\overline{\tau}(\cdot)$ and $\underline{\tau}(\cdot)$ are uniformly sharp if $\overline{\tau}(\cdot), \underline{\tau}(\cdot) \in \mathcal{M}$.*

The following theorem is almost immediate.

Theorem B.2. *$\overline{\tau}(\cdot)$ is uniformly sharp if and only if there exists $q^*(\cdot) \in \mathcal{Q}$ such that $q^*(\cdot)$ is in the feasible set and $\overline{\tau}(u) = \sum_{e \in \mathcal{E}: g_e(1)=1} q^*(e|u) - \sum_{e \in \mathcal{E}: g_e(0)=1} q^*(e|u)$ for all $u \in [0,1]$. Similarly, $\underline{\tau}(\cdot)$ is uniformly sharp if and only if there exists $q^\dagger(\cdot) \in \mathcal{Q}$ such that $q^\dagger(\cdot)$ is in the feasible set and $\underline{\tau}(u) = \sum_{e \in \mathcal{E}: g_e(1)=1} q^\dagger(e|u) - \sum_{e \in \mathcal{E}: g_e(0)=1} q^\dagger(e|u)$ for all $u \in [0,1]$.*

The following is a more useful result that relates point-wise bounds with uniform bounds. For each $\bar{u}$, let $q^*_{\bar{u}}(\cdot)$ and $q^\dagger_{\bar{u}}(\cdot)$ be the point-wise maximizer and minimizer of (B.1)–(B.3), respectively.

Corollary B.1. *$\overline{\tau}(\cdot)$ is uniformly sharp if and only if there exists $q^*(\cdot) \in \mathcal{Q}$ such that $q^*(\cdot)$ is in the feasible set and $q^*_{\bar{u}}(\bar{u}) = q^*(\bar{u})$ for all $\bar{u} \in [0,1]$. Also, $\underline{\tau}(u)$ is uniformly sharp if and only if there exists $q^\dagger(\cdot) \in \mathcal{Q}$ such that $q^\dagger(\cdot)$ is in the feasible set and $q^\dagger_{\bar{u}}(\bar{u}) = q^\dagger(\bar{u})$ for all $\bar{u} \in [0,1]$.*

Based on the Bernstein approximation we introduce, this corollary implies that for a uniform upper bound to exist, there should exist a common maximizer θ^* such that θ^* is in the feasible set of the LP and $\overline{\tau}(u) = \sum_{k \in \mathcal{K}} \left\{ \sum_{e \in \mathcal{E}: g_e(1)=1} \theta^{e*}_k b_k(u) - \sum_{e \in \mathcal{E}: g_e(0)=1} \theta^{e*}_k b_k(u) \right\}$ *for all u*. In other words, if $\theta^*_{\bar{u}}$ is the maximizer of the LP for given $\bar{u}$, then there should exist θ^* in the feasible set such that $\theta^*_{\bar{u}} = \theta^*$ for all $\bar{u} \in [0,1]$. Since this condition will not generally hold, uniformly sharp bounds on the MTE may not exist. The condition can be verified in practice by implementing the LP in a finite grid of u in $[0,1]$ and checking whether θ^*_u is constant for all values in the grid.

B.2 Inference

It is important to construct a confidence set for our target parameter or its bounds in order to account for the sampling variation in measuring treatment effectiveness. It will also be interesting to develop a procedure to conduct a specification test for the identifying assumptions discussed in Section 6. The problem of statistical inference when the identified set is constructed via linear programming has been studied in, e.g.,Deb et al. (2017), Mogstad et al. (2018), Hsieh et al. (2018), and Torgovitsky (2019b) . Among these papers, Mogstad et al. (2017)'s setting is closest to ours, and their inference procedure can be directly adapted to our problem. Instead of repeating their result here, we only briefly discuss the procedure.

Recall $q(u,x) \equiv \{q(e|u,x)\}_{e\in\mathcal{E}}$ is the latent distribution and $p \equiv \{p(1,d|z,x)\}_{d,z,x}$ is the distribution of the data, and R_τ, R_0, R_1, and R_2 denote the linear operators of $q(\cdot)$ that correspond to the target and constraints. Consider the following hypotheses:

$$H_0 : p \in \mathcal{P}_0, \qquad H_1 : p \in \mathcal{P}\backslash\mathcal{P}_0,$$

where

$$\mathcal{P}_0 \equiv \{p \in \mathcal{P} : Rq = a \text{ for some } q \in \mathcal{Q}\}$$

and

$$\begin{aligned} R &\equiv (R'_\tau, R'_0, R'_1, R'_2)' \\ a &\equiv (\tau, p', a'_1, a'_2)' \end{aligned}$$

Suppose $\hat{R}$ and $\hat{a}$ are sample counterparts of R and a. Then, a minimum distance test statistic can be constructed as

$$T_n(\tau) \equiv \inf_{q\in\mathcal{Q}_K} \sqrt{n} \left\| \hat{R}q - \hat{a} \right\|.$$

Similar to Mogstad et al. (2017), $T_n(\tau)$ is the solution to a convex optimization problem that can be reformulated as an LP using duality. A $(1-\alpha)$-confidence set for the target parameter τ can be constructed by inverting the test:

$$CS_{1-\alpha} \equiv \{\tau : T_n(\tau) \leq \hat{c}_{1-\alpha}\}$$

where $\hat{c}_{1-\alpha}$ is the critical value for the test. The resulting object is of independent interest, and it can further be used to conduct specification tests. The large sample theory for $T_n(\tau)$, as well as a bootstrap procedure to calculate $\hat{c}_{1-\alpha}$, will directly follow according to Mogstad et al. (2017), which is omitted for succinctness.

B.3 Linear Programming with Continuous X

Suppose X is continuously distributed and assume $\mathcal{X} = [0,1]^{d_X}$. Let $q(u,x) \equiv \{q(e|u,x)\}_{e \in \mathcal{E}}$ and $p(x) \equiv \{p(1,d|z,x)\}_{d,z}$. Recall that $R_\tau : \mathcal{Q} \to \mathbb{R}$ and $R : \mathcal{Q} \to \mathbb{R}^{d_p}$ are the linear operators of $q(\cdot)$ where d_p is the dimension of p. Consider the following LP:

$$\overline{\tau} = \sup_{q \in \mathcal{Q}} R_\tau q, \tag{B.4}$$

$$\underline{\tau} = \inf_{q \in \mathcal{Q}} R_\tau q, \tag{B.5}$$

$$s.t. \quad (Rq)(x) = p(x) \qquad \text{for all } x \in \mathcal{X}, \tag{B.6}$$

where $(Rq)(x) = p(x)$ emphasizes the dependence on x, and thus contains infinitely many constraints. Therefore, this LP is infinite dimensional because of not only the decision variable but also the constraints. The problem with q is addressed with the sieve approximation. To address the problem with continuous X, we proceed as follows. Note that, for any measurable function $h : \mathcal{X} \to \mathbb{R}$, $E\,|h(X)| = 0$ if and only if $h(x) = 0$ almost everywhere in $\mathcal{X}$. Therefore, each j-th equation in the equality restrictions (B.6) can be replaced by

$$E\,|(Rq)_j(X) - p_j(X)| = 0.$$

Now, for the sieve space of $\mathcal{Q}$, we consider

$$\tilde{\mathcal{Q}}_K \equiv \left\{ \Big\{ \sum_{k=1}^{\tilde{K}} \theta_k^e b_k(u,x) \Big\}_{e \in \mathcal{E}} : \sum_{e \in \mathcal{E}} \theta_k^e = 1 \, \forall k \in \tilde{\mathcal{K}} \text{ and } \theta_k^e \geq 0 \, \forall (e,k) \right\} \subseteq \mathcal{Q}, \tag{B.7}$$

where $b_k(u,x)$ is a bivariate Bernstein polynomial and $\tilde{\mathcal{K}} \equiv \{1, ..., \tilde{K}\}$. Then,

$$\begin{aligned} E[\tau_d(Z,X)] &= \sum_{e: g_e(d)=1} \sum_{k \in \tilde{\mathcal{K}}} \theta_k^e \int E[b_k(u,X) w_d(u,Z,X)] du \\ &\equiv \sum_{e: g_e(d)=1} \sum_{k \in \tilde{\mathcal{K}}} \theta_k^e \tilde{\gamma}_k^d, \end{aligned} \tag{B.8}$$

where $\tilde{\gamma}_k^d \equiv \int E[b_k(u,X) w_d(u,Z,X)] du$. Also,

$$\begin{aligned} p(y,d|z,x) &= \sum_{e: g_e(d)=y} \sum_{k \in \tilde{\mathcal{K}}} \theta_k^e \int_{\mathcal{U}_{z,x}^d} b_k(u,x) du \\ &\equiv \sum_{e: g_e(d)=y} \sum_{k \in \tilde{\mathcal{K}}} \theta_k^e \tilde{\delta}_k^d(z,x), \end{aligned} \tag{B.9}$$

where $\tilde{\delta}_k^d(z,x) \equiv \int_{\mathcal{U}_{z,x}^d} b_k(u,x)du$. Let $\tilde{\theta} \equiv \{\theta_k^e\}_{(e,k)\in\mathcal{E}\times\tilde{\mathcal{K}}}$ and let

$$\tilde{\Theta}_{\tilde{K}} \equiv \left\{\tilde{\theta} : \sum_{e\in\mathcal{E}} \theta_k^e = 1\,\forall k \in \tilde{\mathcal{K}} \text{ and } \theta_k^e \geq 0\,\forall(e,k)\right\}.$$

Then, we can formulate the following finite-dimensional LP:

$$\overline{\tau}_{\tilde{K}} = \max_{\theta\in\Theta_{\tilde{K}}} \sum_{k\in\tilde{\mathcal{K}}} \left\{ \sum_{e:g_e(1)=1} \theta_k^e \tilde{\gamma}_k^1 - \sum_{e:g_e(0)=1} \theta_k^e \tilde{\gamma}_k^0 \right\} \tag{B.10}$$

$$\underline{\tau}_{\tilde{K}} = \min_{\theta\in\Theta_{\tilde{K}}} \sum_{k\in\tilde{\mathcal{K}}} \left\{ \sum_{e:g_e(1)=1} \theta_k^e \tilde{\gamma}_k^1 - \sum_{e:g_e(0)=1} \theta_k^e \tilde{\gamma}_k^0 \right\} \tag{B.11}$$

subject to

$$E\left| \sum_{e:g_e(d)=1} \sum_{k\in\tilde{\mathcal{K}}} \theta_k^e \tilde{\delta}_k^d(Z,X) - p(1,d|Z,X) \right| = 0. \tag{B.12}$$

In estimation, we use the sample counterparts $\hat{\tilde{\gamma}}_k^d$ and $\hat{\tilde{\delta}}_k^d$ for $\tilde{\gamma}_k^d$ and $\tilde{\delta}_k^d$, and (B.12) can be estimated with slackness by

$$\frac{1}{n}\sum_{i=1}^n \left| \sum_{e:g_e(d)=1} \sum_{k\in\tilde{\mathcal{K}}} \theta_k^e \hat{\tilde{\delta}}_k^d(Z_i,X_i) - \hat{p}(1,d|Z_i,X_i) \right| \leq \eta,$$

where $\hat{p}(1,d|z,x)$ is some preliminary estimate of $p(1,d|z,x)$ and η is the slackness parameter.

Later, we want to introduce additional constraints from some identifying assumptions:

$$R_1 q = a_1 \tag{B.13}$$

$$R_2 q \leq a_2 \tag{B.14}$$

For the equality restrictions, we can use the same approach that transforms (B.6). For the inequality restrictions (B.14), we can allow any identifying assumptions for which R_2 is a matrix rather than an operator:

Assumption MAT. *R_2 is a* $\dim(a_2) \times \dim(q)$ *matrix.*

Assumptions M and C and the unconditional version of Assumption MTS satisfy this condition.

B.4 Equivalence with the IV-Like Estimands

We draw a connection between our approach and the approach used in Mogstad et al. (2018). In particular, we show that the identified set of the MTR functions $\mathcal{M}_{id}$ used in Mogstad et al. (2018) is equivalent to the set of MTR functions derived from the feasible set used in this paper. Therefore, the feasible set in this paper contains no less information about the data than those contained in $\mathcal{M}_{id}$ via IV-like estimands in their paper.

The IV-like estimand is defined in Proposition 3 in Mogstad et al. (2018), and is stated as below.

Proposition B.1 (IV-like Estimand from Mogstad et al. (2018))**.** *Suppose that $s : \{0, 1\} \times \mathbf{R}^{d_z \times d_x} \to \mathbf{R}$ is an identified (or known) function that is measurable and has a finite second moment. We refer to such a function s as an IV-like specification and to $\beta_s \equiv E\Big[s(D, Z, X)Y\Big]$ as an IV-like estimand. If (Y, D) are generated according to Assumption SEL and Assumption EX, then*

$$\beta_s = E\Big[\int_0^1 m_0(u, X)\omega_{0s}(u, Z, X)du\Big] + E\Big[\int_0^1 m_1(u, X)\omega_{1s}(u, Z, X)du\Big], \tag{B.15}$$

where $\omega_{0s}(u, z, x) = s(0, z, x)1[u > p(z, x)]$, and $\omega_{1s}(u, z, x) = s(1, z, x)1[u \leq p(z, x)]$.

For the MTR functions to be consistent with the data, the following conditions need to be satisfied:

$$E[Y|D = 0, Z, X] = E[Y_0|U > p(Z, X), Z, X] = \frac{1}{1 - P(Z, X)} \int_{p(Z,X)}^1 m_0(u, X)du, \tag{B.16}$$

$$E[Y|D = 1, Z, X] = E[Y_1|U \leq p(Z, X), Z, X] = \frac{1}{P(Z, X)} \int_0^{p(Z,X)} m_1(u, X)du. \tag{B.17}$$

Define the identified set as:

$$\mathcal{M}_{id} = \Big\{m = (m_0, m_1), m_0, m_1 \in L^2 : m_0, m_1 \text{ satisfies equation (B.16) and (B.17) a.s}\Big\}.$$

This identified set is defined in Mogstad et al. (2018, Section 2.5). The definition follows the fact that the MTR functions in $\mathcal{M}_{id}$ are compatible with the observed conditional means of Y. In this sense, it exhausts the information of the data contained in the conditional means. When Y is binary, the conditional means of Y contain the information of the complete distribution.

Define the feasible set $\mathcal{Q}_f$ as

$$\mathcal{Q}_f = \Big\{ q \in L^2 : q \in \mathcal{Q} \text{ and satisfies equation } (\infty\text{-LP3}) \Big\}.$$

To establish the connection with $\mathcal{M}_{id}$, we construct the set of MTR functions based on the feasible set:

$$\mathcal{M}_f = \Big\{ m = (m_0, m_1) : m_d = \sum_{e: g_e(d)=1} q(e|u, x), d = \{0, 1\}, q \in \mathcal{Q}_f \Big\}.$$

Then the following holds, proof of which appears later:

Theorem B.3. *Suppose Y is discretely distributed. Under the Assumption SEL and EX, $\mathcal{M}_f = \mathcal{M}_{id}$.*

Proposition 3 in Mogstad et al. (2018) shows an equivalence relationship between the identified set $\mathcal{M}_{id}$ and the set of MTR functions satisfying constraints based on selected IV-like estimands. Theorem B.3 shows that the information contained in our feasible set used in the LP is the same as the selected IV-like estimands that exhaust the available information. Theorem B.3 can be extended to the case where Y is discrete and X is continuous. When Y is a non-binary discrete outcome variable, $\mathcal{M}_{id}$ and $\mathcal{M}_f$ only exhaust the information on the conditional means, but not other distributional information. Nonetheless, that missing information is captured by $\mathcal{Q}_f$ that we use as our constraint set, because $q(e|u)$ is defined as the conditional probability of Y taking each value.

C Proofs

C.1 Proof of Lemma 3.1

Fix (d, z, x). By $\sum_{e \in \mathcal{E}} q(e|u, x) = 1$ for $q \in \mathcal{Q}$, we have

$$1 = \sum_{e \in \mathcal{E}} q(e|u, x) = \sum_{e: g_e(d)=1} q(e|u, x) + \sum_{e: g_e(d)=0} q(e|u, x).$$

Then, in (∞-LP3), the constraint with $p(0, d|z, x)$ can be written as

$$\begin{aligned} p(0, d|z, x) &= \int_{\mathcal{U}^d_{z,x}} \sum_{e: g_e(d)=0} q(e|u, x) du = \int_{\mathcal{U}^d_{z,x}} \Big\{ 1 - \sum_{e: g_e(d)=1} q(e|u, x) \Big\} du \\ &= \Pr[D = d|Z = z, X = x] - \int_{\mathcal{U}^d_{z,x}} \sum_{e: g_e(d)=1} q(e|u, x) du. \end{aligned}$$

Then by rearranging terms, this constraint becomes

$$p(1,d|z,x)=\int_{\mathcal{U}^d_{z,x}}\sum_{e:g_e(d)=1}q(e|u,x)du,$$

since $\Pr[D=d|Z=z,X=x]-p(0,d|z,x)=p(1,d|z,x)$. Therefore, the constraint with $p(0,d|z,x)$ does not contribute to the restrictions imposed by (∞-LP3) and $q\in\mathcal{Q}$. $\square$

C.2 Proof of Theorem 5.1

In proving the claim of the theorem, note that Z can be fixed at a certain value, so we fix $Z=z$ here. We first prove with Case (a). To simplify notation, let $q(e_1,...,e_J|u)\equiv\Pr[\epsilon\in\{e_1,...,e_J\}|u]=\sum_{j=1}^{J}q(e_j|u)$. Based on Table (1), we can easily derive

$$\begin{aligned}
p(1,1|z,1)&=\int_0^{P(z)}\sum_{e:g_e(1,1)=1}q(e|u)du=\int_0^{P(z)}q(9,...,16|u)du,\\
p(1,1|z,0)&=\int_0^{P(z)}\sum_{e:g_e(1,0)=1}q(e|u)du=\int_0^{P(z)}q(5,...,8,13,...,16|u)du,\\
p(1,0|z,1)&=\int_{P(z)}^{1}\sum_{e:g_e(0,1)=1}q(e|u)du=\int_{P(z)}^{1}q(3,4,7,8,11,12,15,16|u)du,\\
p(1,0|z,0)&=\int_{P(z)}^{1}\sum_{e:g_e(0,0)=1}q(e|u)du=\int_{P(z)}^{1}q(2,4,6,8,10,12,14,16|u)du.
\end{aligned}$$

Define the operator

$$T^d_z q^e\equiv\int_{\mathcal{U}^d_z}q(e|u)du.$$

Then, for the r.h.s. $(p_{11|z1}, p_{11|z0}, p_{10|z1}, p_{10|z0})'$ of the constraints in (LP$_W$3) that correspond to $Z = z$, the corresponding l.h.s. is

$$\begin{aligned}
&\begin{pmatrix} \int_0^{P(z)} q(9,...,16|u)du \\ \int_0^{P(z)} q(5,...,8,13,...,16|u)du \\ \int_{P(z)}^1 q(3,4,7,8,11,12,15,16|u)du \\ \int_{P(z)}^1 q(2,4,6,8,10,12,14,16|u)du \end{pmatrix} \\
&= \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & T_z^1 & T_z^1 & T_z^1 & T_z^1 & T_z^1 & T_z^1 & T_z^1 & T_z^1 \\ 0 & 0 & 0 & 0 & T_z^1 & T_z^1 & T_z^1 & T_z^1 & 0 & 0 & 0 & 0 & T_z^1 & T_z^1 & T_z^1 & T_z^1 \\ 0 & 0 & T_z^0 & T_z^0 & 0 & 0 & T_z^0 & T_z^0 & 0 & 0 & T_z^0 & T_z^0 & 0 & 0 & T_z^0 & T_z^0 \\ 0 & T_z^0 & 0 & T_z^0 & 0 & T_z^0 & 0 & T_z^0 & 0 & T_z^0 & 0 & T_z^0 & 0 & T_z^0 & 0 & T_z^0 \end{pmatrix} q \\
&\equiv Tq,
\end{aligned}$$

where T is a matrix of operators implicitly defined and $q(u) \equiv (q(1|u), ..., q(16|u))$. Now for $q \in \mathcal{Q}_K$, define a $16K$-vector

$$\theta \equiv \begin{pmatrix} \theta^1 \\ \vdots \\ \theta^{16} \end{pmatrix}$$

where, for each $e \in \{1, ..., 16\}$, $\theta^e \equiv (\theta_1^e, ..., \theta_K^e)'$. Similarly, let $b(u) \equiv (b_1(u), ..., b_K(u))'$. Then, we have $q(e|u) = b(u)'\theta^e$. Let H be a 16×16 diagonal matrix of 1's and 0's that imposes additional identifying assumptions on the outcome data-generating process. In this proof, H is used to incorporate Assumption R(i). Given H, the constraints in (LP$_W$3) (that correspond to $Z = z$) can be written as

$$THq = \{TH \otimes b'\}\,\theta = (p_{11|z1}, p_{11|z0}, p_{10|z1}, p_{10|z0})'.$$

Now, we prove the claim of the theorem. Suppose the claim is not true, i.e., the even rows are linearly dependent to odd rows in TH. Given the form of T, which has full rank under Assumption R(ii)(a), this linear dependence only occurs when H is such that $H_{jj} = 1$ for $j \in \{1, 4, 13, 16\}$ and 0 otherwise. But, according to Table 1, this implies that $\Pr[Y(d,w) \neq Y(d,w')] = 0$ for all d and $w \neq w'$, which contradicts Assumption R(i). This proves the theorem for Case (a).

Now we move to prove the theorem for Case (b), analogous to the previous case. For every z, we can derive

$$
\begin{aligned}
p(1,1|z,1) &= \int_0^{P(z,1)} \sum_{e:g_e(1,1)=1} q(e|u)du = \int_0^{P(z,1)} q(9,...,16|u)du, \\
p(1,1|z,0) &= \int_0^{P(z,0)} \sum_{e:g_e(1,0)=1} q(e|u)du = \int_0^{P(z,0)} q(5,...,8,13,...,16|u)du, \\
p(1,0|z,1) &= \int_{P(z,1)}^1 \sum_{e:g_e(0,1)=1} q(e|u)du = \int_{P(z,1)}^1 q(3,4,7,8,11,12,15,16|u)du, \\
p(1,0|z,0) &= \int_{P(z,0)}^1 \sum_{e:g_e(0,0)=1} q(e|u)du = \int_{P(z,0)}^1 q(2,4,6,8,10,12,14,16|u)du.
\end{aligned}
$$

Define

$$T^d_{z,w} q^e \equiv \int_{\mathcal{U}^d_{z,w}} q(e|u)du$$

where $\mathcal{U}^d_{z,w}$ can be analogously defined. Then,

$$
\begin{aligned}
&\begin{pmatrix} \int_0^{P(z,w)} q(9,...,16|u)du \\ \int_0^{P(z,w')} q(5,...,8,13,...,16|u)du \\ \int_{P(z,w)}^1 q(3,4,7,8,11,12,15,16|u)du \\ \int_{P(z,w')}^1 q(2,4,6,8,10,12,14,16|u)du \end{pmatrix} \\
&= \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & T^1_{z,w} & T^1_{z,w} & T^1_{z,w} & T^1_{z,w} & T^1_{z,w} & T^1_{z,w} & T^1_{z,w} & T^1_{z,w} \\
0 & 0 & 0 & 0 & T^1_{z,w'} & T^1_{z,w'} & T^1_{z,w'} & T^1_{z,w'} & 0 & 0 & 0 & 0 & T^1_{z,w'} & T^1_{z,w'} & T^1_{z,w'} & T^1_{z,w'} \\
0 & 0 & T^0_{z,w} & T^0_{z,w} & 0 & 0 & T^0_{z,w} & T^0_{z,w} & 0 & 0 & T^0_{z,w} & T^0_{z,w} & 0 & 0 & T^0_{z,w} & T^0_{z,w} \\
0 & T^0_{z,w'} & 0 & T^0_{z,w'} & 0 & T^0_{z,w'} & 0 & T^0_{z,w'} & 0 & T^0_{z,w'} & 0 & T^0_{z,w'} & 0 & T^0_{z,w'} & 0 & T^0_{z,w'}
\end{pmatrix} q \\
&\equiv \tilde{T} q,
\end{aligned}
$$

where $\tilde{T}$ is a matrix of operators implicitly defined. Then, inserting H, the constraint becomes

$$\tilde{T}Hq = \left\{ \tilde{T}H \otimes b' \right\} \theta = (p_{11|z1}, p_{11|z0}, p_{10|z1}, p_{10|z0})'.$$

Then the remaining argument is the same as in the previous case, which completes the proof. □

C.3 Proof of Theorem B.1

For any given $\bar{u} \in [0,1]$, $\overline{\tau}(\bar{u}) = \sum_{e \in \mathcal{E}: g_e(1)=1} q^*_{\bar{u}}(e|\bar{u}) - \sum_{e \in \mathcal{E}: g_e(0)=1} q^*_{\bar{u}}(e|\bar{u})$ for some $q^*_{\bar{u}}(\cdot) \equiv \{q^*_{\bar{u}}(e|\cdot)\}_{e \in \mathcal{E}}$ in the feasible set of the LP, (B.1) and (B.3). Therefore, $\overline{\tau}(\bar{u}) = \overline{\tau}_{MTE,\bar{u}}(\bar{u})$ for

$\overline{\tau}_{MTE,\bar{u}}(\bar{u}) = \sum_{e\in\mathcal{E}:g_e(1)=1} q^*_{\bar{u}}(e|\bar{u}) - \sum_{e\in\mathcal{E}:g_e(0)=1} q^*_{\bar{u}}(e|\bar{u})$, which is in $\mathcal{M}$ by definition. We can have a symmetric proof for $\underline{\tau}(\cdot)$. □

C.4 Proof of Theorem B.2

Again, by the fact that $\tau_{MTE}(\cdot) = \sum_{e\in\mathcal{E}:g_e(1)=1} q(e|\cdot) - \sum_{e\in\mathcal{E}:g_e(0)=1} q(e|\cdot)$ in general, $\overline{\tau}(u) = \sum_{e\in\mathcal{E}:g_e(1)=1} q^*(e|u) - \sum_{e\in\mathcal{E}:g_e(0)=1} q^*(e|u)$ for all $u \in [0,1]$ is equivalent to $\overline{\tau}(\cdot)$ being contained in $\mathcal{M}$, and similarly for $\underline{\tau}(\cdot)$. □

C.5 Proof of Theorem B.3

From (∞-LP3), we can write $E[Y|D=0,Z,X]$ in terms of $q(e|u,X)$ as below:

$$\begin{aligned}
E[Y|D=0,Z,X] = \Pr[Y=1|D=0,Z,X] &= \frac{\Pr[Y=1,D=0|Z,X]}{\Pr[D=0|Z,X]} \\
&= \frac{1}{1-P(Z,X)} \sum_{e:g_e(0)=1} \int_{P(Z,X)}^{1} q(e|u,X)du \\
&= \frac{1}{1-P(Z,X)} \int_{P(Z,X)}^{1} \sum_{e:g_e(0)=1} q(e|u,X)du \qquad \text{(C.1)}
\end{aligned}$$

Therefore, for $(m_0, m_1) \in \mathcal{M}_f$

$$E[Y|D=0,Z,X] = \frac{1}{P(Z,X)} \int_{P(Z,X)}^{1} m_0(u,X)du$$

and symmetrically,

$$E[Y|D=1,Z,X] = \frac{1}{P(Z,X)} \int_{0}^{P(Z,X)} m_1(u,X)du$$

We conclude that $\mathcal{M}_f \subset \mathcal{M}_{id}$.

Now suppose $m \in \mathcal{M}_{id}$. By (B.16) and (C.1), for $\forall z, x$

$$\frac{1}{1-P(z,x)} \int_{P(z,x)}^{1} m_0(u,x)du = \frac{1}{1-P(z,x)} \sum_{e:g_e(0)=1} \int_{P(z,x)}^{1} q(e|u,x)du$$

and,

$$\int_{P(z,x)}^{1}\left[m_0(u,x)-\sum_{e:g_e(0)=1}q(e|u,x)\right]du=0$$

This equality holds for all the possible values of $P(z,x)$, we conclude that $m_0(u,x) = \sum_{e:g_e(0)=1} q(e|u,x)$ on the support $u \in [0,1]$, $\forall x$ following the fundamental theorem of calculus. Following the symmetric procedure, we can conclude that $m_1(u,x) = \sum_{e:g_e(1)=1} q(e|u,x)$. And we show that $\mathcal{M}_{id} \subset \mathcal{M}_f$. Thus, $\mathcal{M}_f = \mathcal{M}_{id}$.

References

Angrist, J. and I. Fernandez-Val (2010): "Extrapolate-ing: External validity and overidentification in the late framework," Tech. rep., National Bureau of Economic Research. 1

Balat, J. F. and S. Han (2018): "Multiple treatments with strategic interaction," *Available at SSRN 3182766*. 1, 5

Balke, A. and J. Pearl (1997): "Bounds on treatment effects from studies with imperfect compliance," *Journal of the American Statistical Association*, 92, 1171–1176. 1, 3

Bertanha, M. and G. W. Imbens (2019): "External validity in fuzzy regression discontinuity designs," *Journal of Business & Economic Statistics*, 1–39. 1

Bhattacharya, J., A. M. Shaikh, and E. Vytlacil (2008): "Treatment effect bounds under monotonicity assumptions: an application to Swan-Ganz catheterization," *American Economic Review*, 98, 351–56. 6.1

Brinch, C. N., M. Mogstad, and M. Wiswall (2017): "Beyond LATE with a discrete instrument," *Journal of Political Economy*, 125, 985–1039. 1, 6, 6.3

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of econometrics*, 6, 5549–5632. 7.3

Chen, X. and T. Christensen (2015): "Optimal sup-norm rates, adaptivity and inference in nonparametric instrumental variables estimation," . 7.3

Chen, X., E. T. Tamer, and A. Torgovitsky (2011): "Sensitivity analysis in semiparametric likelihood models," . 4

Chen, X., J. Tan, Z. Liu, and J. Xie (2017): "Approximation of functions by a new family of generalized Bernstein operators," *Journal of Mathematical Analysis and Applications*, 450, 244–261. 4

Chernozhukov, V. and C. Hansen (2005): "An IV model of quantile treatment effects," *Econometrica*, 73, 245–261. 1, 6.1

Coolidge, J. L. (1949): "The story of the binomial theorem," *The American Mathematical Monthly*, 56, 147–157. 4

Cornelissen, T., C. Dustmann, A. Raute, and U. Schönberg (2016): "From LATE to MTE: Alternative methods for the evaluation of policy interventions," *Labour Economics*, 41, 47–60. 6

Deb, R., Y. Kitamura, J. K.-H. Quah, and J. Stoye (2017): "Revealed price preference: Theory and stochastic testing," . 1, B.2

Dehejia, R., C. Pop-Eleches, and C. Samii (2019): "From local to global: External validity in a fertility natural experiment," *Journal of Business & Economic Statistics*, 1–27. 1

Dunlop, D. D., L. M. Manheim, J. Song, and R. W. Chang (2002): "Gender and ethnic/racial disparities in health care utilization among older adults," *The Journals of Gerontology Series B: Psychological sciences and social sciences*, 57, S221–S233. 8

Finkelstein, A., S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen, K. Baicker, and O. H. S. Group (2012): "The Oregon health insurance experiment: evidence from the first year," *The Quarterly journal of economics*, 127, 1057–1106. 8

Firpo, S. and G. Ridder (2019): "Partial identification of the treatment effect distribution and its functionals," *Journal of Econometrics*, 213, 210–234. 9

Gunsilius, F. (2019): "Bounds in continuous instrumental variable models," *arXiv preprint arXiv:1910.09502*. 1

Han, S. (2020a): "Nonparametric estimation of triangular simultaneous equations models under weak identification," *Quantitative Economics*, 11, 161–202. 7.3

——— (2020b): "Optimal Dynamic Treatment Regimes and Partial Welfare Ordering," *arXiv preprint arXiv:1912.10014*. 1, 6.1

Han, S. and S. Lee (2019): "Estimation in a generalization of bivariate probit models with dummy endogenous regressors," *Journal of Applied Econometrics*, 34, 994–1015. 1, 5, 8

Han, S. and E. J. Vytlacil (2017): "Identification in a generalization of bivariate probit models with dummy endogenous regressors," *Journal of Econometrics*, 199, 63–73. 1, 5

Heckman, J. J. and E. Vytlacil (2005): "Structural equations, treatment effects, and econometric policy evaluation1," *Econometrica*, 73, 669–738. 1, 2, 2, B.1

Heckman, J. J. and E. J. Vytlacil (1999): "Local instrumental variables and latent variable models for identifying and bounding treatment effects," *Proceedings of the national Academy of Sciences*, 96, 4730–4734. 1

Hsieh, Y.-W., X. Shi, and M. Shum (2018): "Inference on estimators defined by mathematical programming," *Available at SSRN 3041040*. B.2

Hurd, M. D. and K. McGarry (1997): "Medical insurance and the use of health care services by the elderly," *Journal of Health Economics*, 16, 129–154. 8

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475. 1, 2

Joy, K. I. (2000): "Bernstein polynomials," *On-Line Geometric Modeling Notes*, 13. 4

Kamat, V. (2019): "Identification with latent choice sets: The case of the head start impact study," *arXiv preprint arXiv:1711.02048*. 1

Kitamura, Y. and J. Stoye (2019): "Nonparametric Counterfactuals in Random Utility Models," *arXiv preprint arXiv:1902.08350*. 1

Kowalski, A. E. (2020): "Reconciling Seemingly Contradictory Results from the Oregon Health Insurance Experiment and the Massachusetts Health Reform," Tech. rep., National Bureau of Economic Research. 1, 6, 8

Machado, C., A. Shaikh, and E. Vytlacil (2019): "Instrumental variables and the sign of the average treatment effect," *Journal of Econometrics*, 212, 522–555. 1

Manski, C. F. (1997): "Monotone treatment response," *Econometrica: Journal of the Econometric Society*, 1311–1334. 6.1

——— (2007): "Partial identification of counterfactual choice probabilities," *International Economic Review*, 48, 1393–1410. 1

Manski, C. F. and J. V. Pepper (2000): "Monotone instrumental variables: With an application to the returns to schooling," *Econometrica*, 68, 997–1010. 1, 6.1, 6.2

Masten, M. A. and A. Poirier (2018): "Salvaging falsified instrumental variable models," *arXiv preprint arXiv:1812.11598*. 4

Mogstad, M., A. Santos, and A. Torgovitsky (2017): "Using Instrumental Variables for Inference about Policy Relevant Treatment Effects," Tech. rep., National Bureau of Economic Research. B.2

——— (2018): "Using instrumental variables for inference about policy relevant treatment parameters," *Econometrica*, 86, 1589–1619. 1, 2, 4, 3, 4, 4, 6.3, A, B.2, B.4, B.1, B.4, B.4

Mogstad, M., A. Torgovitsky, and C. R. Walters (2019): "Identification of causal effects with multiple instruments: Problems and some solutions," Tech. rep., National Bureau of Economic Research. 4

Mourifié, I. (2015): "Sharp bounds on treatment effects in a binary triangular system," *Journal of Econometrics*, 187, 74–81. 1, 5

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. 1

Shaikh, A. M. and E. J. Vytlacil (2011): "Partial identification in triangular systems of equations with binary dependent variables," *Econometrica*, 79, 949–955. 1, 5, 6.1

Taubman, S. L., H. L. Allen, B. J. Wright, K. Baicker, and A. N. Finkelstein (2014): "Medicaid increases emergency-department use: evidence from Oregon's Health Insurance Experiment," *Science*, 343, 263–268. 8

Tebaldi, P., A. Torgovitsky, and H. Yang (2019): "Nonparametric estimates of demand in the california health insurance exchange," Tech. rep., National Bureau of Economic Research. 1

Torgovitsky, A. (2019a): "Nonparametric Inference on State Dependence in Unemployment," *Econometrica*, 87, 1475–1505. 1

——— (2019b): "Nonparametric inference on state dependence in unemployment," *Econometrica*, 87, 1475–1505. B.2

Vuong, Q. and H. Xu (2017): "Counterfactual mapping and individual treatment effects in nonseparable models with binary endogeneity," *Quantitative Economics*, 8, 589–610. 1, 5

Vytlacil, E. (2002): "Independence, monotonicity, and latent index models: An equivalence result," *Econometrica*, 70, 331–341. 2

Vytlacil, E. and N. Yildiz (2007): "Dummy endogenous variables in weakly separable models," *Econometrica*, 75, 757–779. 1, 5