

Inference for Treatment Effects Conditional on Generalized Principal Strata using Instrumental Variables *

Yuehao Bai
Department of Economics
University of Southern California
yuehao.bai@usc.edu

Shunzhuang Huang
Booth School of Business
University of Chicago
shunzhuang.huang@chicagobooth.edu

Sarah Moon
Department of Economics
Massachusetts Institute of Technology
sarahmn@mit.edu

Andres Santos
Department of Economics
University of California–Los Angeles
andres@econ.ucla.edu

Azeem M. Shaikh
Department of Economics
University of Chicago
amshaikh@uchicago.edu

Edward J. Vytlacil
Department of Economics
Yale University
edward.vytlacil@yale.edu

September 18, 2026

Abstract

We propose a general approach to inference for a broad class of models that arise in the analysis of treatment effects with discrete-valued treatments and instruments and a general-valued outcome. In addition to instrument exogeneity, the main substantive assumption in our class of models rules out certain response types by assuming that they occur with probability zero. Here, the response type refers to the vector of potential outcomes and potential treatments, and we refer to a set of possible values for the response type as a generalized principal stratum. Through a series of examples, we show that this framework encompasses a wide variety of assumptions that have been considered in the previous literature. Our framework allows inference on any treatment effect parameter that can be expressed as the expectation of a function of the response type conditional on a generalized principal stratum. We develop methods for inference on such parameters under these assumptions, as well as methods for testing the validity of the assumptions themselves. A key result of our analysis is a characterization of the identified set for such parameters under these assumptions and the testable restrictions for the assumptions themselves in terms of existence of a nonnegative solution to linear systems of equations with a special structure. We propose methods for inference exploiting this special structure and recent results in [Fang et al. \(2023\)](#).

KEYWORDS: Multi-valued Treatments, Model Specification, Model Validity, Randomized Controlled Trial, Principal Strata, Instrumental Variables, Partial Identification.

*We thank the editor and three anonymous referees, as well as Jing Cheng, Jiaying Gu, Sukjin Han, Nathan Kallus, Soonwoo Kwon, Sokbae Lee, Derek Neal, Kirill Ponomarev, Vitor Possebom, Bernard Salanié, and Panos Toulis for helpful comments. Shaikh acknowledges support from NSF Grant SES-2419008; Moon acknowledges support from NSF Grant No. 2141064; Vytlacil acknowledges financial support from the Tobin Center for Economic Policy at Yale.

1 Introduction

We propose a general approach to inference for a broad class of models that arise in the analysis of treatment effects with discrete-valued treatments and instruments and a general-valued outcome. In addition to instrument exogeneity, the main substantive restriction imposed in our analysis is that certain values for the response types occur with probability zero. Here, the response type refers to the vector of potential outcomes and potential treatments, and we refer to a set of possible values for the response type as a generalized principal stratum. Through a series of examples, we show that our framework encompasses a wide variety of assumptions that have been considered in the previous literature, including the following: (i) restrictions in the analysis of randomized controlled trials (RCTs) under noncompliance considered, e.g., in [Imbens and Angrist \(1994\)](#) and [Cheng and Small \(2006\)](#); (ii) generalizations of these restrictions considered in [Bai et al. \(2026\)](#); (iii) revealed preference-type restrictions on response types considered in [Kirkeboen et al. \(2016\)](#), [Kline and Walters \(2016\)](#), and [Heckman and Pinto \(2018\)](#); and (iv) restrictions on the ordering of potential treatments or on the ordering of potential outcomes considered in [Manski \(1997\)](#), [Manski and Pepper \(2000\)](#), and [Machado et al. \(2019\)](#).¹

Our framework allows inference on any treatment effect parameter that can be expressed as the expectation of a function of the response type conditional on a generalized principal stratum. In this way, our framework accommodates many parameters that have been considered previously in the literature, including both average and distributional treatment effect parameters, such as the probability of being strictly helped and the probability of not being hurt by the treatment. It further accommodates versions of these parameters conditional on sets of possible values of potential treatments, such as the local average treatment effect in [Imbens and Angrist \(1994\)](#) and average effects conditional on principal strata in [Frangakis and Rubin \(2002\)](#). It also accommodates parameters conditional on sets of possible values of potential outcomes, such as those considered in [Heckman et al. \(1997\)](#) and [Heckman and Smith \(1999\)](#), as well as the population persuasion rate in [Jun and Lee \(2023\)](#). By contrast, each of the papers cited in the preceding paragraph tailors its method to specific examples of such treatment effect parameters. Furthermore, because the types of restrictions considered in these papers are typically insufficient for identification of many parameters of interest, some of these papers focus specifically on those treatment effect parameters that are identified from the two-stage least squares estimand. A characterization of the parameters that are point identified by such restrictions is given by [Navjeevan et al. \(2026\)](#) and [Goff \(2025\)](#).

A key result in our analysis is a characterization of the identified set for such parameters

¹In the context of mediation analysis, [Kwon and Roth \(2026\)](#) show that “full mediation” is equivalent to certain assumptions like those we consider.

under these assumptions and the testable restrictions for the assumptions themselves in terms of existence of a nonnegative solution to linear systems of equations with a special structure. Using this result, we build on results in Fang et al. (2023) to develop a test for the null hypothesis that a pre-specified value for the parameter of interest lies in the identified set as well as the null hypothesis that the testable restrictions are satisfied. Importantly, the resulting test remains well behaved even in “high-dimensional” settings, meaning it controls size uniformly over a large class of distributions satisfying weak assumptions and permitting, among other things, the support of the treatment and instrument or the support of the response types to be “large” relative to the sample size. Through test inversion, we may also construct confidence regions for such parameters that are uniformly consistent in level. We highlight that while our test builds on Fang et al. (2023), it leverages the structure of our problem to obtain novel strong approximations that significantly improve on the coupling rates of Fang et al. (2023). These coupling results for high-dimensional vectors of estimates of conditional probabilities and their bootstrap counterparts may be of independent interest.

Other methods for inference have been proposed in the prior literature for some special cases of our framework. For inference on certain treatment effect parameters, see, e.g., Cheng and Small (2006), Bhattacharya et al. (2008, 2012), and Machado et al. (2019); for inference on the validity of the assumptions, see, e.g., Kitagawa (2015), Heckman and Pinto (2018), Sun (2023), Bai et al. (2025), and Kaido and Ponomarev (2025). These approaches rely upon closed-form expressions for the identified set for the parameter of interest or testable inequalities, which must be proved or computed on a case-by-case basis. When such expressions need to be computed, the methods are feasible only when the support of response types is small. In contrast, our approach does not rely on such closed-form expressions and remains computationally feasible even in high-dimensional settings. A key strength of our approach is its broad applicability, including in settings for which no existing methods apply.

The remainder of the paper is organized as follows. Section 2 introduces our formal setup and notation. In Section 3, we provide several examples of parameters and restrictions previously considered in the literature that are nested by our framework. We present our inference method when the support of the outcome variable is restricted to be finite in Section 4, and without such a restriction in Section 5. We illustrate our method through a simulation study in Section 6, and demonstrate its usefulness by revisiting the empirical study of Blattman et al. (2017) in Section 7. Proofs of our main results can be found in Appendix A. An Online Appendix contains the proofs of auxiliary results. Appendices C, D, E contain additional examples and details of the simulation study and the empirical application, and can be found at <https://github.com/edward-vytlacil/STRATA-ATE/blob/main/>.

2 Setup and Notation

Denote by $Y \in \mathcal{Y}$ an outcome, by $D \in \mathcal{D}$ a discrete-valued endogenous regressor (i.e., the treatment), and by $Z \in \mathcal{Z}$ a discrete-valued instrumental variable. To rule out degenerate cases, we assume throughout that $2 \leq |\mathcal{Y}|$, $2 \leq |\mathcal{D}| < \infty$, and $2 \leq |\mathcal{Z}| < \infty$. We emphasize that neither $\mathcal{D}$ nor $\mathcal{Z}$ needs to be a subset of the real line; in this way, our framework accommodates vector-valued treatments and instruments provided they only take a finite number of values. In what follows, we will consider both settings in which $|\mathcal{Y}|$ is finite and those in which it is not. Further denote by $Y(d) \in \mathcal{Y}$ the potential outcome if $D = d \in \mathcal{D}$ and by $D(z) \in \mathcal{D}$ the potential treatment if $Z = z \in \mathcal{Z}$. As usual, we assume that

$$Y = \sum_{d \in \mathcal{D}} Y(d) I\{D = d\} \quad \text{and} \quad D = \sum_{z \in \mathcal{Z}} D(z) I\{Z = z\} , \quad (1)$$

where, throughout the paper, $I\{E\}$ denotes the indicator function equal to one if the event E holds and zero otherwise.

In our discussion below, it will be convenient to define $R_o \equiv (Y(d) : d \in \mathcal{D})$ and $R_t \equiv (D(z) : z \in \mathcal{Z})$, and set $R \equiv (R_o, R_t)$. Following [Heckman and Pinto \(2018\)](#), we will refer to R_t as the “treatment response type.” By analogy with this terminology, we will also refer to R_o as the “outcome response type” and to R as simply the “response type.” We let Q denote the distribution of (R_o, R_t, Z) and note that by (1) we have

$$(Y, D, Z) = T(R_o, R_t, Z) ,$$

for a transformation T implicitly defined through (1). The distribution Q therefore determines, through the transformation T , the distribution of (Y, D, Z) which we denote by P . We notationally formalize this relation by writing $P = QT^{-1}$, which signifies that P is the pushforward distribution of Q by the function T .

Below we will require that $Q \in \mathbf{Q}$, where $\mathbf{Q}$ is a class of distributions satisfying assumptions that we will specify. Different specifications of $\mathbf{Q}$ represent different assumptions that we impose on the distribution of potential outcomes and potential treatments. In this sense, $\mathbf{Q}$ may be viewed as a model for potential outcomes and potential treatments.

Given a distribution P of (Y, D, Z) and a model $\mathbf{Q}$, we define the identified set for the distribution of (R_o, R_t, Z) as the set of $Q \in \mathbf{Q}$ that rationalizes P . Formally, this set equals

$$\mathbf{Q}_0(P, \mathbf{Q}) \equiv \{Q \in \mathbf{Q} : P = QT^{-1}\} .$$

We say $\mathbf{Q}$ is consistent with P if and only if $\mathbf{Q}_0(P, \mathbf{Q})$ is not empty. For every model $\mathbf{Q}$ considered in this paper, every $Q \in \mathbf{Q}$ is assumed to satisfy the restriction:

Assumption 2.1 (Instrument Exogeneity). $R \perp\!\!\!\perp Z$ under Q .

Our remaining restrictions on $\mathbf{Q}$ will be formulated in terms of restrictions on possible values of the response type R . These restrictions will be expressed by specifying a set $\mathcal{R} \subseteq \mathcal{Y}^{|\mathcal{D}|} \times \mathcal{D}^{|\mathcal{Z}|}$ characterizing the possible values of R . In Section 3, we provide several different choices of $\mathcal{R}$ that have been previously considered in the literature. For a given choice of $\mathcal{R}$, we will impose the following assumption on every $Q \in \mathbf{Q}$:

Assumption 2.2 (Response Type Restrictions). $Q\{R \in \mathcal{R}\} = 1$.

Following Frangakis and Rubin (2002), sets of the form $\{R_t = r_t\}$ for r_t a possible value of R_t are referred to as “principal strata.” By analogy with this terminology, we will refer to sets of the form $\{R \in \mathcal{R}'\}$ for $\mathcal{R}' \subseteq \mathcal{R}$ as “generalized principal strata.” Using this notation, we introduce our parameters of interest, which are those that can be written as

$$\theta(Q) \equiv E_Q[g(R) \mid R \in \mathcal{R}'] \quad (2)$$

for different choices of integrable function $g : \mathcal{R} \rightarrow \mathbf{R}$ and generalized principal strata $\mathcal{R}' \subseteq \mathcal{R}$. To ensure the parameter $\theta(Q)$ is well-defined we require $Q\{R \in \mathcal{R}'\} > 0$. A wide variety of parameters can be accommodated in this way. In Section 3, we show that specific parameters considered in the literature have the structure in (2). We note now, however, that natural choices of g correspond to the effect of one treatment versus another (i.e., $g(R) = Y(d) - Y(d')$ for $d \in \mathcal{D}$ and $d' \in \mathcal{D}$) and the probability that one treatment leads to a larger outcome than another treatment (i.e., $g(R) = I\{Y(d) > Y(d')\}$ for $d \in \mathcal{D}$ and $d' \in \mathcal{D}$). We further note that when $\mathcal{R}' = \mathcal{R}$, the parameter $\theta(Q)$ simplifies to $E_Q[g(R)]$.

For a given distribution P and model $\mathbf{Q}$, the identified set for $\theta(Q)$ is given by

$$\Theta_0(P, \mathbf{Q}) \equiv \{\theta(Q) : Q \in \mathbf{Q}_0(P, \mathbf{Q}) \text{ and } Q\{R \in \mathcal{R}'\} > 0\} . \quad (3)$$

The set $\Theta_0(P, \mathbf{Q})$ is nonempty whenever there exists at least one $Q \in \mathbf{Q}_0(P, \mathbf{Q})$ satisfying $Q\{R \in \mathcal{R}'\} > 0$. By construction, this set is “sharp” in the sense that for any value in the set there exists a distribution Q that is consistent with P , satisfies the restrictions of the model, and for which $\theta(Q)$ equals the prescribed value.

3 Examples

In this section, we show how to accommodate several examples from the literature in our framework. Our discussion focuses in particular on Assumption 2.1 and the specification of $\mathcal{R}$ in Assumption 2.2 but, where the cited literature has emphasized specific parameters of

interest, we additionally describe how those parameters of interest are special cases of $\theta(Q)$, as defined in (2), for suitable choices of g and $\mathcal{R}'$. Throughout, denote by $r_o = (y(d) : d \in \mathcal{D})$ and $r_t = (d(z) : z \in \mathcal{Z})$, respectively, a possible value of an outcome response type and a treatment response type.

Example 3.1 (*RCT with one-sided noncompliance*). Consider a multi-arm randomized controlled trial (RCT) with noncompliance, where $\mathcal{D} = \mathcal{Z} = \{0, \dots, K\}$, $Z = d$ denotes random assignment to treatment d , and $D(d) = d$ denotes that the subject would comply with assignment if assigned to treatment d . In this example, Q satisfies Assumption 2.1 because Z is randomly assigned. Suppose noncompliance to the assignment is one-sided (Bloom, 1984) in the sense that one can always take the control $d = 0$, but, for any other treatment $d \neq 0$, one can only take that treatment if assigned to it. This restriction can be expressed in terms of Assumption 2.2 with $\mathcal{R} = \{(r_o, r_t) : d(j) \in \{0, j\} \text{ for all } j \in \mathcal{D}\}$. Analyses of causal parameters that condition on generalized principal strata in this context follow Imbens and Angrist (1994) in using the Wald estimand to identify local average treatment effect (LATE) parameters of the form $E_Q[Y(j) - Y(0) \mid D(j) = j]$ for $j \in \{1, \dots, K\}$. Note that identification of such parameters does not allow comparison of $Y(j)$ to $Y(k)$ for $j, k \neq 0$ for any subgroup of subjects. A notable exception is Cheng and Small (2006), who provide limited results for $K = 2$ and other treatment effect parameters; further discussion of their approach is provided in Remark 4.3. Our framework nests these parameters as well as others, such as relative treatment effectiveness, for any subgroup defined by a generalized principal stratum. ■

Example 3.2 (*Encouragement design*). Consider a multi-arm RCT with possibly two-sided noncompliance, where $\mathcal{D} = \mathcal{Z} = \{0, \dots, K\}$, $Z = d$ denotes random assignment to treatment d , and $D(d) = d$ denotes that the subject would comply with assignment if assigned to treatment d . More generally, not necessarily in the context of an RCT, one can interpret $Z = d$ as random encouragement to treatment d and interpret $D(d) = d$ as the subject taking treatment d if encouraged to do so. In this example, Q satisfies Assumption 2.1 because Z is randomly assigned. Bai et al. (2026) generalize the “no-defier” restriction of Imbens and Angrist (1994) to the condition

$$Q\{D(d) \neq d, D(d') = d \text{ for some } d' \neq d\} = 0 \text{ for all } d \in \mathcal{D}, \quad (4)$$

i.e., a subject that would not take treatment d if assigned to (encouraged to take) d but who would take d if assigned (encouraged) to some other treatment $d' \neq d$ occurs with probability zero. This restriction can be formulated in terms of Assumption 2.2 with $\mathcal{R} = \{(r_o, r_t) : d(j) \neq j \Rightarrow d(k) \neq j \text{ for all } j, k \in \mathcal{D}\}$. Bai et al. (2026) derive the identified set for unconditional average treatment effects in this context. While the two-stage least squares

(TSLS) estimand in this setting generally does not correspond to any well-defined causal parameter when $K \geq 2$, [Bhuller and Sigstad \(2024\)](#) show that under strong, additional assumptions, it identifies a particular weighted average of strata-specific causal effects that may not be of a priori interest. In contrast, our framework allows inference on a broad class of parameters of substantive interest including those that condition on generalized principal strata and that need not coincide with the TSLS estimand. ■

Example 3.3 (*RCT with close substitute*). [Kline and Walters \(2016\)](#) consider an RCT with a “close substitute” to study the effects of preschooling on educational outcomes. In their setting, $D \in \mathcal{D} = \{c, h, n\}$, where $D = h$ denotes a preschool program called Head Start, $D = c$ denotes a preschool other than Head Start (the close substitute), and $D = n$ denotes home care (no preschool). Let $Z \in \mathcal{Z} = \{0, 1\}$ indicate a randomized offer to attend Head Start. Assumption 2.1 holds because Z is randomly assigned. In evaluating the cost-effectiveness of Head Start, [Kline and Walters \(2016\)](#) impose the restriction

$$Q\{D(1) = h \mid D(0) \neq D(1)\} = 1 \quad (5)$$

and the instrument relevance requirement $Q\{D(0) \neq D(1)\} > 0$. Condition (5) states that if a family’s schooling choice changes upon receiving a Head Start offer, then they must choose Head Start when receiving the offer. Equivalently, it cannot be the case that upon receiving a Head Start offer, a family switches from no preschool to preschools other than Head Start, or the other way around. The restriction can be formulated in terms of Assumption 2.2 with

$$\mathcal{R} = \{(r_o, r_t) : (d(0), d(1)) \in \{(c, c), (h, h), (n, n), (c, h), (n, h)\}\} .$$

[Kline and Walters \(2016\)](#) show that the Wald estimand identifies a weighted combination of “sub-LATEs”:

$$\frac{E[Y \mid Z = 1] - E[Y \mid Z = 0]}{P\{D = h \mid Z = 1\} - P\{D = h \mid Z = 0\}} = S_c \text{SubLATE}_{ch} + (1 - S_c) \text{SubLATE}_{nh} ,$$

where

$$\begin{aligned} \text{SubLATE}_{ch} &= E_Q[Y(h) - Y(c) \mid D(1) = h, D(0) = c] \\ \text{SubLATE}_{nh} &= E_Q[Y(h) - Y(n) \mid D(1) = h, D(0) = n] \end{aligned}$$

and $S_c = Q\{D(1) = h, D(0) = c \mid D(1) = h, D(0) \neq h\}$. Here, S_c and the sub-LATEs can all be written as (2) for appropriate choices of g and $\mathcal{R}'$. [Kline and Walters \(2016\)](#) show that these separate sub-LATE parameters are of substantive interest, and identify them under additional assumptions including imposing a multinomial normal selection model. Our framework allows inference on these sub-LATE parameters without imposing such additional

assumptions sufficient to identify them. ■

Example 3.4 (*Ordered monotonicity with known direction*). Consider a setting in which both the treatment and the instrument admit natural orderings, and write $\mathcal{D} = \{0, \dots, K_d\}$ and $\mathcal{Z} = \{0, \dots, K_z\}$. Suppose further that Z is randomly assigned, so Assumption 2.1 holds. In many applications, it is reasonable to assume the monotonicity restriction $Q\{D(j) \geq D(k)\} = 1$ for any pair of instrument values $j, k \in \mathcal{Z}$ with $j \geq k$. For example, Dupas (2014) reports an experiment in which Z represents different randomly assigned subsidy levels for insecticide-treated bed nets, and D denotes the total number of bed nets purchased over a two-year period. The monotonicity restriction in this context is that one would purchase at least as many insecticide-treated bed nets with a greater subsidy as with a lower subsidy. This restriction can be expressed in terms of Assumption 2.2 by setting $\mathcal{R} = \{(r_o, r_t) : d(K_z) \geq \dots \geq d(0)\}$. The results of Angrist and Imbens (1995) imply that the TSLS estimand in this context identifies a specific weighted average of strata-specific causal effects, whereas our analysis allows inference on a broader class of parameters—including partially identified parameters—that need not coincide with the TSLS estimand. ■

Example 3.5 (*Harmless treatment*). Consider a setting in which $Z \in \mathcal{Z} = \{0, \dots, K_z\}$ is randomly assigned, so Assumption 2.1 holds, and there is a baseline treatment, i.e., the control, corresponding to $0 \in \mathcal{D} = \{0, \dots, K_d\}$. In many applications, it is reasonable to assume that the remaining treatments are harmless relative to the control. For example, in Angrist et al. (2009), the outcome is academic performance, the control is no treatment, and the noncontrol treatments are (i) providing students with academic peer-advising service; (ii) financial incentives for good academic performance; and (iii) both (i) and (ii). It is therefore natural to assume that $Q\{Y(d) \geq Y(0)\} = 1$ for all $d \in \mathcal{D}$. This restriction can be formulated in terms of Assumption 2.2 by setting $\mathcal{R} = \{(r_o, r_t) : y(d) \geq y(0) \text{ for all } d \in \mathcal{D}\}$. This restriction is an example of what Manski (1997) termed a semi-monotone ordering of outcomes. While Manski (1997) develops partial-identification results for unconditional treatment parameters in this context without restrictions on treatment-response types, our analysis additionally allows such restrictions to be incorporated and enables inference on parameters that condition on generalized principal strata. ■

Appendix C contains additional examples that fit our framework, including RCTs with restrictions on noncompliance considered in Cheng and Small (2006), revealed preference-type restrictions on response types considered in Kirkeboen et al. (2016) and Heckman and Pinto (2018), and monotone treatment response considered in Manski (1997).

In many of the examples discussed above, Y may be an ordinal outcome. In such instances, average treatment effects may not be interpretable. Nonetheless, researchers may consider other parameters that fit our framework, such as: (i) the conditional probability of benefit

of treatment j versus treatment k (i.e. $Q\{Y(j) > Y(k) \mid R \in \mathcal{R}'\}$); (ii) the conditional probability of no harm of treatment j versus treatment k (i.e. $Q\{Y(j) \geq Y(k) \mid R \in \mathcal{R}'\}$); or (iii) the conditional relative treatment effect of treatment j versus k (i.e. $Q\{Y(j) > Y(k) \mid R \in \mathcal{R}'\} - Q\{Y(k) > Y(j) \mid R \in \mathcal{R}'\}$). Each of these parameters has the structure in (2) for appropriate choices of g and therefore fits our framework. These parameters have been previously studied for the special case in which $\mathcal{R}' = \mathcal{R}$ and $|\mathcal{D}|$ equals two or three by Lu et al. (2018), Huang et al. (2019), and Gabriel et al. (2024).

It is also instructive to discuss examples of restrictions that *do not* fit our framework. Angrist and Imbens (1995), for instance, consider a generalization of the monotonicity assumption in Example 3.4 in which the direction of the monotonicity is not known *a priori*; i.e., they consider the restriction that, for any pair of values for the instrument $j, k \in \mathcal{Z}$, either $Q\{D(j) \geq D(k)\} = 1$ or $Q\{D(j) \leq D(k)\} = 1$. Such a restriction is also analyzed in Vytlačil (2006) and in Heckman and Pinto (2018), where the latter paper terms it “ordered monotonicity.” These ordered monotonicity conditions fall outside our framework because they correspond to specifying a collection of restrictions $\mathcal{R}_\pi$ indexed by $\pi \in \Pi$, and requiring that $Q\{R \in \mathcal{R}_\pi\} = 1$ for some $\pi \in \Pi$ (instead of $Q\{R \in \mathcal{R}\} = 1$ for some *known* $\mathcal{R}$).

4 Inference with Discrete Outcomes

In this section, we propose a test for conducting inference on the parameter of interest $\theta(Q)$. To this end, recall that P denotes the distribution of (Y, D, Z) . In this section only, we suppose $|\mathcal{Y}| < \infty$. Doing so allows us to introduce our method succinctly. This assumption will be modified in Section 5, where we allow $\mathcal{Y}$ to be a compact interval in $\mathbf{R}$. In what follows, we assume that $P \in \mathbf{P}$, where $\mathbf{P}$ is a “large” class of distributions that we specify below.

We define $\mathcal{M} \subseteq \mathcal{Y} \times \mathcal{D} \times \mathcal{Z}$ as the set of possible realizations of (Y, D, Z) that is consistent with $\mathcal{R}$ being the support of the response types R . Formally, recall that $r_o = (y(d) : d \in \mathcal{D})$ and $r_t = (d(z) : z \in \mathcal{Z})$, and let $r = (r_o, r_t)$. For any $z \in \mathcal{Z}$, let

$$\mathcal{M}_z \equiv \{(y, d) \in \mathcal{Y} \times \mathcal{D} : (y(d), d(z)) = (y, d) \text{ for some } r = (r_o, r_t) \in \mathcal{R}\} \quad (6)$$

and

$$\mathcal{M} \equiv \bigcup_{z \in \mathcal{Z}} (\mathcal{M}_z \times \{z\}) . \quad (7)$$

The set $\mathcal{M}$ need not equal $\mathcal{Y} \times \mathcal{D} \times \mathcal{Z}$ — as in, for example, RCTs with one-sided noncompliance in Example 3.1. For $(y, d, z) \in \mathcal{M}$, we further define $P_{ydz} \equiv P\{Y = y, D = d, Z = z\}$, $P_z \equiv P\{Z = z\}$, and $P_{ydz} \equiv P\{Y = y, D = d \mid Z = z\}$. The class $\mathbf{P}$ can depend on the

sample size n , but we suppress the dependence from our notation. We consider testing

$$H_0 : P \in \mathbf{P}_0 \text{ versus } H_1 : P \in \mathbf{P} \setminus \mathbf{P}_0 \quad (8)$$

at level $\alpha \in (0, 1)$, where, for a pre-specified value of θ_0 , the null hypothesis we consider is

$$\mathbf{P}_0 = \{P \in \mathbf{P} : \theta_0 \in \Theta_0(P, \mathbf{Q})\} . \quad (9)$$

A test of this null hypothesis can be used to construct confidence regions for $\theta(Q)$ through test inversion over the pre-specified value θ_0 . Moreover, as we show below, a test of this null hypothesis can also be used to conduct a specification test by appropriately selecting $\theta(Q)$.

The key insight underlying the construction of our test is the following theorem, which provides a convenient reformulation of $\mathbf{P}_0$ in terms of existence of a nonnegative solution to a (possibly under-determined) system of linear equations in which the “slope coefficients” are known. The statement of the theorem involves the parameter

$$\beta(P) \equiv ((P_{y|d|z} : (y, d, z) \in \mathcal{M}), 1, 0)' . \quad (10)$$

We also need to define a matrix A with $|\mathcal{M}|+2$ rows and $|\mathcal{R}|$ columns. Indexing the first $|\mathcal{M}|$ rows of A by $(y, d, z) \in \mathcal{M}$ and the columns by $r = (r_o, r_t) \in \mathcal{R}$, recalling $r_o = (y(d) : d \in \mathcal{D})$ and $r_t = (d(z) : z \in \mathcal{Z})$, we then set the coordinates of A as

$$\begin{aligned} A_{(y,d,z),r} &= I\{y(d) = y, d(z) = d\} \\ A_{(|\mathcal{M}|+1),r} &= 1 \\ A_{(|\mathcal{M}|+2),r} &= (g(r) - \theta_0)I\{r \in \mathcal{R}'\} . \end{aligned} \quad (11)$$

Table 1 illustrates the construction of A and $\beta(P)$ when the parameter of interest is the ATE, the model assumes one-sided noncompliance, and the outcome, treatment, and instrument are binary. Following Example 3.1 with $K = 1$, here $\mathcal{R} = \{r = (r_o, r_t) : r_o \in \{0, 1\} \times \{0, 1\}, r_t \in \{(0, 0), (0, 1)\}\}$ with $\mathcal{R}' = \mathcal{R}$ and $g(r) = y(1) - y(0)$.

Our first theorem employs $\beta(P)$ and A to characterize the null hypothesis.

Theorem 4.1. *Suppose $|\mathcal{Y}| < \infty$. Let $\mathbf{Q}$ be the set of all distributions of (R, Z) satisfying Assumptions 2.1 and 2.2, and $\mathbf{P}$ be the set of all distributions of (Y, D, Z) satisfying $P\{(Y, D, Z) \in \mathcal{M}\} = 1$ and $P\{Z = z\} > 0$ for all $z \in \mathcal{Z}$. Then, for $\beta(P)$ as defined in (10) and A as defined in (11), it follows that*

$$\mathbf{P}_0 \subseteq \{P \in \mathbf{P} : Ax = \beta(P) \text{ for some } x \geq 0\} . \quad (12)$$

Furthermore, if $\mathbf{P}_0 \neq \emptyset$, then $\{P \in \mathbf{P} : Ax = \beta(P) \text{ for some } x \geq 0\} \subseteq \text{cl}_{\mathbf{P}}(\mathbf{P}_0)$, where

$(y(0), y(1), d(0), d(1))$								$\beta(P)$
$(0, 0, 0, 0)$	$(0, 0, 0, 1)$	$(1, 0, 0, 0)$	$(1, 0, 0, 1)$	$(0, 1, 0, 0)$	$(0, 1, 0, 1)$	$(1, 1, 0, 0)$	$(1, 1, 0, 1)$	
1	1			1	1			$P_{00 0}$
		1	1			1	1	$P_{10 0}$
1				1				$P_{00 1}$
		1				1		$P_{10 1}$
	1		1					$P_{01 1}$
					1		1	$P_{11 1}$
1	1	1	1	1	1	1	1	1
$-\theta_0$	$-\theta_0$	$-1 - \theta_0$	$-1 - \theta_0$	$1 - \theta_0$	$1 - \theta_0$	$-\theta_0$	$-\theta_0$	0

Table 1: Matrix A and $\beta(P)$ in a one-sided noncompliance model, as in Example 3.1, with $\mathcal{Y} = \mathcal{D} = \mathcal{Z} = \{0, 1\}$, and the parameter of interest being the ATE so that $\mathcal{R}' = \mathcal{R}$, and $g(r) = y(1) - y(0)$. All omitted entries are zero.

$\text{cl}_{\mathbf{P}}(\mathbf{P}_0)$ denotes the closure of $\mathbf{P}_0$ in $\mathbf{P}$ (understood as a subset of $\mathbf{R}^{|\mathcal{M}|}$).

The key implication of Theorem 4.1 is that we may test the null hypothesis of interest by examining whether there exists a nonnegative vector x satisfying $Ax = \beta(P)$. Because $\theta(Q)$ is not well-defined under Q assigning zero probability to $\mathcal{R}'$, result (12) involves an inclusion rather than an equality. From a testing perspective, however, $\mathbf{P}_0$ and its closure are indistinguishable, in the sense that any level α test of whether P belongs to $\mathbf{P}_0$ necessarily has power no larger than α against any P in the closure $\text{cl}_{\mathbf{P}}(\mathbf{P}_0)$ (Romano, 2004). Therefore, the second part of Theorem 4.1 may be interpreted as showing that testing whether $P \in \mathbf{P}_0$ is in fact equivalent to testing whether $\beta(P) = Ax$ for some $x \geq 0$. This conclusion holds under the condition that $\mathbf{P}_0$ is not empty. However, by definition, $\mathbf{P}_0$ is empty whenever there is no possible distribution of the data $P \in \mathbf{P}$ that is compatible with the null hypothesis. In other words, we may focus on testing whether $\beta(P) = Ax$ for some $x \geq 0$ *unless* the null hypothesis specifies restrictions that are incompatible with any P and, in this sense, may be interpreted as being logically inconsistent with each other.

Remark 4.1. The system of equations $Ax = \beta(P)$ contains linearly dependent restrictions. This is seen in Table 1, for example, where summing the first two restrictions yields restriction number seven (the adding up constraint). More generally, the set of solutions to $Ax = \beta(P)$ is unchanged if, for each $z \in \mathcal{Z}$, we select a point $(y_z, d_z) \in \mathcal{M}_z$ and remove the rows $((y_z, d_z, z) : z \in \mathcal{Z})$ from both A and $\beta(P)$. In particular, it follows that Theorem 4.1 continues to hold if we redefine A and $\beta(P)$ in this manner; see Lemma A.1 in the Appendix. As we discuss below, this observation can be helpful when constructing our test. ■

Theorem 4.1 can be easily extended to test a joint hypothesis on multiple parameters with the structure in (2). Adding an additional parameter simply corresponds to including an additional row in A (with the same structure as the last equation in (11)) and adding a zero to $\beta(P)$. Through test inversion, this observation provides the basis for constructing

joint confidence regions for multiple parameters. Moreover, Theorem 4.1 also allows us to test whether the model is correctly specified in the sense that $\mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset$. To see this, consider setting $g = 1$, $\theta_0 = 1$, and $\mathcal{R}' = \mathcal{R}$. For such values of these quantities, it follows

$$\Theta_0(P, \mathbf{Q}) = \begin{cases} \{1\} & \text{if } \mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset \\ \emptyset & \text{otherwise} \end{cases}.$$

Therefore, $\mathbf{P}_0 = \{P \in \mathbf{P} : 1 \in \Theta_0(P, \mathbf{Q})\} = \{P \in \mathbf{P} : \mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset\}$. Under these choices of g and θ_0 , the equality in the last row of the equation $Ax = \beta(P)$ in (12) is always satisfied, and we therefore let $\tilde{A}$ and $\tilde{\beta}(P)$ denote all but the last rows of A and $\beta(P)$. Using this notation, the next result formalizes the implications of Theorem 4.1 for specification testing.

Corollary 4.1. *Suppose $|\mathcal{Y}| < \infty$. Let $\mathbf{Q}$ be the set of all distributions of (R, Z) satisfying Assumptions 2.1 and 2.2, and $\mathbf{P}$ be the set of all distributions of (Y, D, Z) satisfying $P\{(Y, D, Z) \in \mathcal{M}\} = 1$ and $P\{Z = z\} > 0$ for all $z \in \mathcal{Z}$. Then, for $g = 1$, $\theta_0 = 1$, and $\mathcal{R}' = \mathcal{R}$,*

$$\mathbf{P}_0 = \{P \in \mathbf{P} : \mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset\} = \{P \in \mathbf{P} : \tilde{A}x = \tilde{\beta}(P) \text{ for some } x \geq 0\}. \quad (13)$$

Unlike Theorem 4.1, the conclusion of Corollary 4.1 holds without needing to close $\mathbf{P}_0$. This contrast is due to $\mathcal{R}'$ equaling $\mathcal{R}$ in the context of Corollary 4.1, which ensures that any Q in our model assigns probability one to $\mathcal{R}'$. The main implication of Corollary 4.1 is a general characterization of correct model specification that can be used for specification testing in general settings. As we show in Appendix D, Corollary 4.1 can also be employed to derive a set of analytical inequalities that characterize correct model specification. We emphasize, however, that our inference methods below do not rely on these analytical inequalities and instead directly leverage the characterization of correct specification in Corollary 4.1.

By Theorem 4.1, we may test the null hypothesis of interest by testing whether P satisfies

$$\beta(P) = Ax \text{ for some } x \geq 0 \quad (14)$$

or, equivalently, the smaller set of restrictions in Remark 4.1. In recent work, Fang et al. (2023) derived a test for whether P satisfies the restriction in (14) for a general class of problems; see also Bai et al. (2022) and Cox and Shi (2023) for other approaches. In what follows, we build on Fang et al. (2023) by employing the structure of our problem to sharpen rate conditions and establish the validity of their test under weaker requirements on the dimension of $\beta(P)$ (relative to the sample size). Following Corollary 4.1, it is straightforward to modify the test described below to test the null hypothesis that the model is correctly specified in the sense that $\mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset$.

In order to describe the test statistic, we require some additional notation. We will assume the availability of an i.i.d. sample $\{V_i\}_{i=1}^n$ with $V_i = (Y_i, D_i, Z_i)$, $V_i \sim P$, and let $\hat{P}_n$ denote the empirical distribution corresponding to the sample $\{V_i\}_{i=1}^n$. We further set $\hat{\beta}_n \equiv \beta(\hat{P}_n)$ to denote the plug-in estimator for the parameter $\beta(P)$ and let $\hat{\Omega}_n \equiv \Omega(\hat{P}_n)$, where $\Omega(P)$ is a diagonal weighting matrix whose (i, i) -entry equals the asymptotic standard deviation of the i^{th} row of $\sqrt{n}(\hat{\beta}_n - \beta(P))$. Specifically, the first $|\mathcal{M}|$ diagonal entries of $\Omega(P)$ equal

$$\left(\frac{P_{yd|z}(1 - P_{yd|z})}{P_z} \right)^{1/2}$$

for the corresponding $(y, d, z) \in \mathcal{M}$, and the final two diagonal entries of $\Omega(P)$ are equal to zero. Finally, we let $A^\dagger$ denote the Moore-Penrose pseudoinverse of A .

Given the introduced notation, we define $\hat{\mathcal{V}}_n \equiv \{s : A^\dagger s \leq 0, \|\hat{\Omega}_n(AA')^\dagger s\|_1 \leq 1\}$ and set

$$T_n \equiv \sup_{s \in \hat{\mathcal{V}}_n} \sqrt{n} \langle A^\dagger s, A^\dagger \hat{\beta}_n \rangle \quad (15)$$

as our test-statistic — here, $\|a\|_1 = \sum_{i=1}^d |a_i|$ for any vector $a = (a_1, \dots, a_d) \in \mathbf{R}^d$ and $\langle a, b \rangle = a'b$. As shown in Fang et al. (2023), the test statistic T_n represents a sample analogue to an equivalent formulation of the restriction in (14) obtained from Farkas' lemma. In particular, the population analogue to T_n equals zero whenever (14) indeed holds and, as a result, under the null hypothesis T_n should not be “too large.” We highlight that the test statistic can be computed through linear programming. As a result, T_n can be reliably and quickly computed even in high-dimensional applications.

Unfortunately, the asymptotic distribution of T_n is not pivotal. In particular, its asymptotic distribution depends on the “directions” $s \in \hat{\mathcal{V}}_n$ at which the population analogue $\langle A^\dagger s, A^\dagger \beta(P) \rangle$ is “close” to zero. In order to obtain a valid critical value, we therefore rely on the nonparametric bootstrap and a construction that asymptotically excludes directions s that do not play a role in the distribution of T_n ; i.e., directions s for which $\langle A^\dagger s, A^\dagger \beta(P) \rangle$ is “far” from zero. Specifically, for the latter purpose we introduce

$$\begin{aligned} \hat{\beta}_n^r &\in \underset{b}{\operatorname{argmin}} \sup_{s \in \hat{\mathcal{V}}_n} |\langle A^\dagger s, A^\dagger (\hat{\beta}_n - b) \rangle| \\ \text{s.t. } Ax &= b \text{ for some } x \geq 0 \text{ and } b = (b'_u, 1, 0)' \text{ for } b_u \in \mathbf{R}^{|\mathcal{M}|}, \end{aligned}$$

which represents an estimator for $\beta(P)$ that is restricted to satisfy the null hypothesis. Importantly, the restricted estimator $\hat{\beta}_n^r$ can be computed through linear programming, which preserves the computational scalability of the test; see Appendix M.1 of Fang et al. (2021) for details. Letting $\{V_i^*\}_{i=1}^n$ denote a bootstrap sample (i.e., a sample drawn i.i.d.

from the empirical distribution $\hat{P}_n$), and $\hat{\beta}_n^*$ the bootstrap analogue to $\hat{\beta}_n$, we set

$$T_n^* \equiv \sup_{s \in \hat{\mathcal{V}}_n} \left\{ \sqrt{n} \langle A^\dagger s, A^\dagger (\hat{\beta}_n^* - \hat{\beta}_n) \rangle + \lambda_n \sqrt{n} \langle A^\dagger s, A^\dagger \hat{\beta}_n^r \rangle \right\} \quad (16)$$

as our bootstrap statistic. Here, $\lambda_n \in [0, 1]$ represents a penalty converging to zero with the sample size, and whose role is to ensure that directions s that do not impact the finite-sample distribution of T_n do not impact the bootstrap statistic either. Following Fang et al. (2023) and the default option in their `lpinfer` package in R, we use the data-dependent choice of λ_n given by $\min\{1, 1/\hat{\tau}_n\}$, where $\hat{\tau}_n$ is the $1 - 1/\sqrt{\log(e \vee \log(e \vee n))}$ bootstrap quantile of

$$\sup_{s \in \hat{\mathcal{V}}_n} \langle A^\dagger s, \sqrt{n} A^\dagger (\hat{\beta}_n^* - \hat{\beta}_n) \rangle.$$

Importantly, just like T_n , the bootstrap statistic T_n^* can also be computed through linear programming and as a result remains computationally tractable in high-dimensional problems. We also consider the other recommended choice of λ_n in Fang et al. (2023), motivated by the law of the iterated logarithm, and find that our empirical results are largely robust to this choice; see Section 7 and Table 10 in Appendix E for additional details.

The critical value for a level α test is then the $1 - \alpha$ quantile of the distribution of T_n^* conditional on the sample $\{V_i\}_{i=1}^n$. Formally, the critical value for our test is given by

$$\hat{c}_n(1 - \alpha) \equiv \inf \left\{ u \in \mathbf{R} : \hat{P}_n \{T_n^* \leq u\} \geq 1 - \alpha \right\}. \quad (17)$$

As usual, the critical value $\hat{c}_n(1 - \alpha)$ can be computed through simulation by drawing multiple i.i.d. samples $\{V_i^*\}_{i=1}^n$ from the empirical distribution $\hat{P}_n$ and computing the $1 - \alpha$ quantile of the corresponding sample of bootstrap statistics. Alternative sampling schemes, such as clustered (instead of i.i.d.) data, can be accommodated by appropriately modifying the bootstrap sampling — e.g., by drawing clusters (instead of observations) with replacement to construct the bootstrap sample. Given the critical value, we next formally introduce our test, which rejects whenever the test statistic T_n exceeds the critical value $\hat{c}_n(1 - \alpha)$:

$$\phi_n(\theta_0) \equiv I\{T_n > \hat{c}_n(1 - \alpha)\}. \quad (18)$$

We establish the asymptotic validity of our test under the following main assumption.

Assumption 4.1. (i) $V_i \equiv (Y_i, D_i, Z_i), i = 1, \dots, n$ is an i.i.d. sample with $V_i \sim P \in \mathbf{P}$; (ii) $(\hat{\beta}_n - \beta(P)) \in \text{range}(A)$ with probability tending to one uniformly in $P \in \mathbf{P}$; (iii) there exists a constant $c_0 > 0$ such that $\sup_{P \in \mathbf{P}} P_{ydz} \leq 1 - c_0$ for all $(y, d, z) \in \mathcal{M}$, and there exist

constants c_1, c_2 , $0 < c_1 < c_2$, satisfying

$$\begin{aligned} c_1 &\leq \inf_{P \in \mathbf{P}} \min_{(y,d,z) \in \mathcal{M}} |\mathcal{M}| P_{ydz} \leq \sup_{P \in \mathbf{P}} \max_{(y,d,z) \in \mathcal{M}} |\mathcal{M}| P_{ydz} \leq c_2 \\ c_1 &\leq \inf_{P \in \mathbf{P}} \min_{z \in \mathcal{Z}} |\mathcal{Z}| P_z \leq \sup_{P \in \mathbf{P}} \max_{z \in \mathcal{Z}} |\mathcal{Z}| P_z \leq c_2 . \end{aligned}$$

Assumption 4.1(i) formalizes our requirement that the sample be independent and identically distributed. In Assumption 4.1(ii) we impose the additional condition that $(\hat{\beta}_n - \beta(P))$ belong to the range of the matrix A with probability tending to one. This requirement is satisfied in our examples, where $(\hat{\beta}_n - \beta(P))$ belongs to the range of A whenever $\hat{P}_z > 0$ for all $z \in \mathcal{Z}$ — i.e., whenever $\hat{\beta}_n$ is well-defined. More generally, Assumption 4.1(ii) can be verified numerically once the matrix A is constructed; see Remark 4.4. We note that Assumption 4.1(ii) is not imposed in Fang et al. (2023), but we deploy it here due to its wide applicability and its importance in obtaining simple primitive conditions for establishing the validity of our test. In applications in which Assumption 4.1(ii) is violated, it is still possible to conduct inference by using the results in Fang et al. (2023). Finally, Assumption 4.1(iii) represents our main regularity condition on the set $\mathbf{P}$. Assumption 4.1(iii) requires that all support points have probabilities of the same order ($1/|\mathcal{M}|$) and that no conditional probability is dominant, in the sense that P_{ydz} is bounded away from one uniformly in the $(y, d, z) \in \mathcal{M}$ and $P \in \mathbf{P}$.

We next establish the validity of our test under Assumption 4.1 and an additional regularity condition that we formally state in the appendix and discuss after the theorem.

Theorem 4.2. *Suppose Assumption 4.1 holds, $\alpha \in (0, 0.5)$, $\log^3(n)|\mathcal{M}|/n = o(1)$, and $0 \leq \lambda_n \leq 1$ satisfies $\lambda_n \sqrt{\log(|\mathcal{M}|)} = o(1)$. Further suppose Assumption A.1 in the appendix is satisfied. Then, it follows that the test defined in (18) satisfies*

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathbf{P}_0} E_P[\phi_n(\theta_0)] \leq \alpha.$$

Theorem 4.2 establishes the validity of our test under weak conditions on the number of support points for (Y, D, Z) . In the appendix, we also show that an analogous result holds if we instead drop linearly dependent restrictions from A when constructing our test, as discussed in Remark 4.1; see Theorem A.1. The main rate requirement imposed by Theorem 4.2 is that $|\mathcal{M}|/n$ tend to zero (up to logs). We view this rate condition as close to minimal, in the sense that it is necessary for the empirical distribution $\hat{P}_n$ to assign positive probability to each support point $(y, d, z) \in \mathcal{M}$ with probability tending to one.² Our rate restriction is in addition a significant improvement over Fang et al. (2023), whose best attainable rate is

²Formally, it can be shown using the negative association of the multinomial distribution (Joag-Dev and Proschan, 1983) and Assumption 4.1(iii) that in order for $P\{\hat{P}_n\{(Y, D, Z) = (y, d, z)\} > 0 \text{ for all } (y, d, z) \in \mathcal{M}\} \rightarrow 1$ we must have $|\mathcal{M}|/n = o(1)$.

that $|\mathcal{M}|^2/n$ tend to zero (up to logs). We improve on their results by employing the specific structure of our problem to derive better coupling rates for our test and bootstrap statistics. In particular, by relying on coupling results in [Massart \(1989\)](#) for the empirical process, we are able to couple our test statistic to a Gaussian counterpart at a rate r_n given by

$$r_n \equiv \frac{\log^{3/2}(n)\sqrt{|\mathcal{M}|}}{\sqrt{n}}.$$

The bootstrap statistic is in turn shown to be coupled to a Gaussian distribution at the same rate r_n , which yields the asymptotic validity of our test under the condition that r_n tend to zero; see [Lemma B.2](#) in the Online Appendix for additional details.

Showing weak convergence of a test statistic and its bootstrap counterpart to a common distribution is in general not sufficient for establishing consistency of the corresponding critical values. In high-dimensional settings, consistency of the critical values requires a condition termed *anti-concentration* by [Chernozhukov et al. \(2015\)](#). Intuitively, anti-concentration ensures that the c.d.f. of the bootstrap statistic is suitably well behaved at the quantile of interest. [Assumption A.1](#), stated in the appendix for ease of exposition, is a sufficient condition for anti-concentration to hold in our setting. Alternatively, [Assumption A.1](#) can be dispensed with by modifying our critical value to include a “correction factor” introduced by [Andrews and Shi \(2013\)](#). We opt to impose [Assumption A.1](#) instead, however, to avoid introducing an additional tuning parameter.

Finally, we emphasize that while we have focused our discussion on hypothesis testing, our results readily deliver confidence regions for the parameter of interest as well. In particular, through test inversion it is straightforward to show that the confidence region

$$C_n = \{\theta_0 \in \mathbf{R} : \phi_n(\theta_0) = 0\} \tag{19}$$

covers the parameter of interest with asymptotic probability of at least $1 - \alpha$ uniformly in $P \in \mathbf{P}$. Implementing test inversion is facilitated by the observation that only the last row of the matrix A depends (linearly) on θ_0 . As a result, the computational cost of obtaining A for test inversion is comparable to the computational cost of computing A for a single test.

Remark 4.2. [Assumption 2.2](#) imposes that response types take on certain values with probability zero. Our framework can be modified to accommodate the restriction that response types belong to $\mathcal{R}$ with at most (or at least) some pre-specified probability. This modification is useful for conducting sensitivity analysis, as in [Kline and Santos \(2013\)](#) and [Masten and Poirier \(2020\)](#). For instance, in [Example 3.2](#), one may relax [restriction \(4\)](#) so that it instead specifies that the left-hand side (the probability of defiers) is at most some pre-specified $\epsilon > 0$, and explore how inferences on $\theta(Q)$ change as one varies ϵ . Similarly, if we maintain

Assumption 2.1 and reject correct model specification by following Corollary 4.1, then our framework also allows us to quantify the violation of the response type restrictions encoded in $\mathcal{R}$. In particular, we may follow Goff (2025) in specifying a richer set of response types $\tilde{\mathcal{R}}$ satisfying $\mathcal{R} \subsetneq \tilde{\mathcal{R}}$ and employ our procedure to obtain a confidence region for $Q\{R \in \tilde{\mathcal{R}} \setminus \mathcal{R}\}$ under the assumption that $Q\{R \in \tilde{\mathcal{R}}\} = 1$. ■

Remark 4.3. In applications with closed-form expressions for the identified set, $\Theta_0(P, \mathbf{Q}) = [L(P), U(P)]$, the plug-in estimator $[L(\hat{P}_n), U(\hat{P}_n)]$ provides a natural approach for estimating the identified set. Such approaches are employed by Cheng and Small (2006), who rely on bounds for $L(P)$ and $U(P)$ that are necessarily sharp. However, the confidence regions based on the plug-in bootstrap in general do not provide correct coverage. Formally, the endpoint maps $L(P)$ and $U(P)$ are typically maxima or minima of finitely many smooth functionals of the observable conditional-probability vector: affine functionals for unconditional parameters and linear-fractional functionals for conditional parameters. Consequently, they are directionally differentiable but not fully differentiable at distributions where multiple component functionals attain the relevant maximum or minimum with different derivatives. Hence, it follows from results in Fang and Santos (2019) that plug-in bootstrap estimates of the quantiles of $L(\hat{P}_n)$ and $U(\hat{P}_n)$ are not consistent. While one could use the methods in Fang and Santos (2019) to consistently estimate these quantiles pointwise in $P \in \mathbf{P}$, the resulting confidence regions would not necessarily be valid uniformly in $P \in \mathbf{P}$. Moreover, we emphasize that our framework provides valid inference when a closed-form expression of the identified set is not available, which is the case for most of the parameters of interest considered in this paper. ■

Remark 4.4. Assumption 4.1(ii) can, under the remaining conditions of Theorem 4.2, be shown to be implied by a requirement on A that is simple to verify. Specifically, let $e_j \in \mathbf{R}^{|\mathcal{M}|+2}$ be the unit vector with its j^{th} coordinate equal to one and $v_{\tilde{z}} \in \mathbf{R}^{|\mathcal{M}|+2}$ be such that its first $|\mathcal{M}|$ coordinates (indexed by $(y, d, z) \in \mathcal{M}$) equal one when $z = \tilde{z}$, and all other coordinates equal zero. Assumption 4.1(ii) then holds whenever the null space of A' is contained in the span of $\{v_z : z \in \mathcal{Z}\} \cup \{e_{|\mathcal{M}|+1}, e_{|\mathcal{M}|+2}\}$. In low dimensional settings, this condition is also necessary for Assumption 4.1(ii) to hold; see Lemma B.8 in the Online Appendix. ■

4.1 Inference under Conditional Instrument Exogeneity

The framework of Theorem 4.1 can be extended to allow instrument exogeneity to hold conditional on an observable discrete-valued covariate $W \in \mathcal{W}$ with $|\mathcal{W}| < \infty$. To this end, we continue to assume that (Y, D) are determined by (R, Z) according to (1) and denote this relation by $(Y, D, Z, W) = T^w(R, Z, W)$. With a slight abuse of notation, in this subsection

only we let Q and P denote the distributions of (R, Z, W) and (Y, D, Z, W) . Therefore, for a given model $\mathbf{Q}^w$, the set of distributions Q compatible with a given P equals

$$\mathbf{Q}_0^w(P, \mathbf{Q}^w) \equiv \{Q \in \mathbf{Q}^w : P = Q(T^w)^{-1}\}.$$

We continue to consider models $\mathbf{Q}^w$ that, for a given choice of $\mathcal{R} \subseteq \mathcal{Y}^{|\mathcal{D}|} \times \mathcal{D}^{|\mathcal{Z}|}$, satisfy Assumption 2.2 for every $Q \in \mathbf{Q}^w$. However, we now replace the instrument exogeneity requirement imposed in Assumption 2.1 with the condition that every $Q \in \mathbf{Q}^w$ satisfies:

Assumption 4.2 (Conditional Instrument Exogeneity). $R \perp\!\!\!\perp Z \mid W$ under Q .

We extend the parameters of interest we consider by allowing them to depend on the covariates W . Formally, we focus on parameters that can be written as

$$\theta^w(Q) = E_Q[g(R, W) | R \in \mathcal{R}', W \in \mathcal{W}'] \quad (20)$$

for different choices of function $g : \mathcal{R} \times \mathcal{W} \rightarrow \mathbf{R}$, and sets $\mathcal{R}' \subseteq \mathcal{R}$ and $\mathcal{W}' \subseteq \mathcal{W}$. To ensure the parameter $\theta^w(Q)$ is well-defined we require $Q\{R \in \mathcal{R}', W \in \mathcal{W}'\} > 0$. Hence, for a given distribution P and model $\mathbf{Q}^w$, the identified set for $\theta^w(Q)$ is given by

$$\Theta_0^w(P, \mathbf{Q}^w) \equiv \{\theta^w(Q) : Q \in \mathbf{Q}_0^w(P, \mathbf{Q}^w) \text{ and } Q\{R \in \mathcal{R}', W \in \mathcal{W}'\} > 0\}.$$

In parallel with Theorem 4.1, we consider testing the null hypothesis that $P \in \mathbf{P}_0^w$, where

$$\mathbf{P}_0^w = \{P \in \mathbf{P}^w : \theta_0 \in \Theta_0^w(P, \mathbf{Q}^w)\}$$

and $\mathbf{P}^w$ is a “large” class of distributions satisfying conditions that we specify below.

We will show that the relationship established in Theorem 4.1 between the null hypothesis and the existence of a nonnegative solution to a linear system extends to this setting. The principal challenge in doing so is to ensure that the “slopes” in the linear system (e.g., the matrix A in Theorem 4.1) do not depend on P so that the inference framework of Theorem 4.2, based on Fang et al. (2023), continues to apply. To this end, we let $\mathcal{M}$ be as defined in (7), set $\mathcal{M}^w \equiv \mathcal{M} \times \mathcal{W}$, and for any $(y, d, z, w) \in \mathcal{M}^w$ we let $P_{yd|zw} \equiv P\{Y = y, D = d | Z = z, W = w\}$ and $P_w \equiv P\{W = w\}$. Given this notation, we replace the vector $\beta(P)$, defined in (10), with a vector $\beta^w(P)$ of dimension $|\mathcal{M}^w| + |\mathcal{W}| + 1$ given by

$$\beta^w(P) \equiv ((P_w \times P_{yd|zw} : (y, d, z, w) \in \mathcal{M}^w), (P_w : w \in \mathcal{W}), 0)'. \quad (21)$$

Similarly, we replace the matrix A , defined in (11), with a matrix A^w having $|\mathcal{M}^w| + |\mathcal{W}| + 1$ rows and $|\mathcal{R}| \times |\mathcal{W}|$ columns. Writing the support of the covariates as $\mathcal{W} = \{w_1, \dots, w_{|\mathcal{W}|}\}$ and indexing the first $|\mathcal{M}^w|$ rows of A^w by $(y, d, z, w) \in \mathcal{M}^w$ and the columns of A^w by

$(r, \tilde{w}) \in \mathcal{R} \times \mathcal{W}$, we set the coordinates of A^w to be given by

$$\begin{aligned} A_{(y,d,z,w),(r,\tilde{w})}^w &= I\{y(d) = y, d(z) = d, \tilde{w} = w\} \\ A_{|\mathcal{M}^w|+j,(r,\tilde{w})}^w &= I\{\tilde{w} = w_j\} \quad \text{for } 1 \leq j \leq |\mathcal{W}| \\ A_{|\mathcal{M}^w|+|\mathcal{W}|+1,(r,\tilde{w})}^w &= (g(r, \tilde{w}) - \theta_0)I\{r \in \mathcal{R}', \tilde{w} \in \mathcal{W}'\}, \end{aligned} \tag{22}$$

where recall that $r_o = (y(d) : d \in \mathcal{D})$ and $r_t = (d(z) : z \in \mathcal{Z})$ for any $(r_o, r_t) = r \in \mathcal{R}$.

The next result shows that the conclusion of Theorem 4.1 continues to hold in the presence of discrete-valued covariates, provided we replace $\beta(P)$ and A by $\beta^w(P)$ and A^w .

Theorem 4.3. *Suppose $|\mathcal{Y}| < \infty$ and $|\mathcal{W}| < \infty$. Let $\mathbf{Q}^w$ be the set of all distributions satisfying Assumptions 4.2 and 2.2, and $\mathbf{P}^w$ be the set of all distributions of (Y, D, Z, W) satisfying $P\{(Y, D, Z, W) \in \mathcal{M}^w\} = 1$ and $P\{Z = z, W = w\} > 0$ for all $(z, w) \in \mathcal{Z} \times \mathcal{W}$. Then, for $\beta^w(P)$ as defined in (21) and A^w as defined in (22), it follows that*

$$\mathbf{P}_0^w \subseteq \{P \in \mathbf{P}^w : A^w x = \beta^w(P) \text{ for some } x \geq 0\}.$$

Furthermore, if $\mathbf{P}_0^w \neq \emptyset$, then $\{P \in \mathbf{P}^w : A^w x = \beta^w(P) \text{ for some } x \geq 0\} \subseteq \text{cl}_{\mathbf{P}^w}(\mathbf{P}_0^w)$, where $\text{cl}_{\mathbf{P}^w}(\mathbf{P}_0^w)$ denotes the closure of $\mathbf{P}_0^w$ in $\mathbf{P}^w$ (understood as a subset of $\mathbf{R}^{|\mathcal{M}^w|}$).

Theorem 4.3 justifies testing the null hypothesis that $P \in \mathbf{P}_0^w$ by instead examining whether there exists a (weakly positive) vector x solving the linear system $A^w x = \beta^w(P)$. Because A^w is constructed to not depend on P , the latter null hypothesis can again be tested by following the approach in Theorem 4.2 for the case in which there are no covariates.

While we have focused on scalar parameters, the characterization of Theorem 4.3 readily extends to vector-valued parameters. For example, consider the conditional parameter curve

$$\boldsymbol{\theta}(Q) \equiv (\theta_w(Q) : w \in \mathcal{W})', \quad \theta_w(Q) \equiv E_Q[g(R, W) \mid R \in \mathcal{R}', W = w],$$

where the curve is defined for Q satisfying $Q\{R \in \mathcal{R}', W = w\} > 0$ for every $w \in \mathcal{W}$. To test the null hypothesis that a candidate curve $\boldsymbol{\theta}_0 = (\theta_{w,0} : w \in \mathcal{W})'$ belongs to the identified set for $\boldsymbol{\theta}(Q)$ we then replace the final row of A^w (as in (22)) with $|\mathcal{W}|$ rows with the structure

$$A_{|\mathcal{M}^w|+|\mathcal{W}|+j,(r,\tilde{w})}^w = (g(r, \tilde{w}) - \theta_{w_j,0})I\{r \in \mathcal{R}', \tilde{w} = w_j\}$$

for each $w_j \in \mathcal{W} = \{w_1, \dots, w_{|\mathcal{W}|}\}$, and similarly replace the final row of $\beta^w(P)$ (as in (21)) with a vector of zeros of dimension $|\mathcal{W}|$. By conducting test inversion we may then obtain a joint confidence region for the conditional parameter curve. Similarly, Theorem 4.3 can be extended to study vector-valued parameters defined by a vector-valued g .

Remark 4.5. A simpler linear system can be employed in applications in which the param-

eter of interest $\theta^w(Q)$ (as in (20)) is specified for a singleton $\mathcal{W}' = \{w'\}$. In such cases, it is possible to replace $\beta(P)$ in (10) by $((P_{ydz|w'} : (y, d, z) \text{ such that } (y, d, z) \in \mathcal{M}), 1, 0)'$ and replace the last row of A in (11) with $(g(r, w') - \theta_0)I\{r \in \mathcal{R}'\}$. Then, Theorem 4.1 applies to the distribution of (Y, D, Z) given $W = w'$ and, through test inversion, we may obtain a confidence region for the parameter of interest. ■

5 Inference with General-Valued Outcomes

Section 4 proposed an inference framework in which the support $\mathcal{Y}$ of Y is taken to be finite. We now modify that framework to allow for continuous outcomes. In this section, we take $\mathcal{Y}$ to be a subset of $\mathbf{R}$ that is no longer restricted to be finite. The treatment D and instrument Z remain discrete, with supports $\mathcal{D}$ and $\mathcal{Z}$ satisfying $2 \leq |\mathcal{D}| < \infty$ and $2 \leq |\mathcal{Z}| < \infty$.

Our approach partitions the outcome space $\mathcal{Y}$, which is now a subset of $\mathbf{R}$, into a finite number of subsets $\{B_{\ell,L}\}_{\ell=1}^L$ and reduces the problem into a sequence of discrete ones indexed by L . To this end, we set

$$\mathbb{L} \equiv \sum_{\ell=1}^L \ell I\{Y \in B_{\ell,L}\}$$

to indicate the subset $B_{\ell,L}$ to which Y belongs. The random variable $\mathbb{L}$ takes values in the set $\mathcal{L} \equiv \{1, \dots, L\}$ and its distribution therefore depends on L but, for simplicity, we omit the dependence from the notation. Paralleling Section 4, we define $\mathcal{M}_L$ to be the set of values of $(\mathbb{L}, D, Z)$ that is consistent with response types R belonging to $\mathcal{R}$. Specifically, recalling that $r = (r_o, r_t)$ with $r_o = (y(d) : d \in \mathcal{D})$ and $r_t = (d(z) : z \in \mathcal{Z})$, we let

$$\mathcal{M}_{L,z} \equiv \{(\ell, d) \in \mathcal{L} \times \mathcal{D} : (y(d), d(z)) \in B_{\ell,L} \times \{d\} \text{ for some } r = (r_o, r_t) \in \mathcal{R}\} \quad (23)$$

and

$$\mathcal{M}_L \equiv \bigcup_{z \in \mathcal{Z}} (\mathcal{M}_{L,z} \times \{z\}). \quad (24)$$

Our analysis mimics the discrete case studied in Theorem 4.1, but replaces $\beta(P)$ with

$$\beta_L(P) \equiv ((P\{Y \in B_{\ell,L}, D = d | Z = z\} : (\ell, d, z) \in \mathcal{M}_L), 1, 0, 0)' ; \quad (25)$$

i.e., the first $|\mathcal{M}_L|$ elements of $\beta_L(P)$ match our construction of $\beta(P)$ in Section 4 for the discrete case when the observations are given by $(\mathbb{L}, D, Z)$ (instead of (Y, D, Z)). Reflecting Theorem 4.1, in what follows we relate our inference problem with general Y to the problem of testing whether $\beta_L(P)$ equals $A_L x$ for some $x \geq 0$ and a suitably defined matrix A_L .

In order to formally state our results, it is helpful to note that because the support of the outcome response type R_o is contained in $\mathcal{Y}^{|\mathcal{D}|}$, any partition $\{B_{\ell,L}\}_{\ell=1}^L$ of $\mathcal{Y}$ induces a

partition for the support of R_o as well. Specifically, writing any $c_o \in \mathcal{L}^{|\mathcal{D}|}$ as $c_o \equiv (c_o(d) : d \in \mathcal{D})$, and setting $B_{c_o} \equiv \prod_{d \in \mathcal{D}} B_{c_o(d), L}$ we have that $\{B_{c_o}\}_{c_o \in \mathcal{L}^{|\mathcal{D}|}}$ forms a partition for $\mathcal{Y}^{|\mathcal{D}|}$. Our next assumption uses this notation in introducing the regularity conditions we require for the main result in this section.

Assumption 5.1. (i) $\mathcal{Y}$ is a compact subset of $\mathbf{R}$; (ii) g is continuous; (iii) $\{B_{\ell, L}\}_{\ell=1}^L$ is a partition of $\mathcal{Y}$ satisfying $\max_{1 \leq \ell \leq L} \text{diam}\{B_{\ell, L}\} = o(1)$ as $L \rightarrow \infty$; (iv) $\mathcal{R}'_o(r_t) \equiv \{r_o \in \mathcal{Y}^{|\mathcal{D}|} : (r_o, r_t) \in \mathcal{R}'\}$ is closed and $B_{c_o} \cap \mathcal{R}'_o(r_t)$ is connected for all $c_o \in \mathcal{L}^{|\mathcal{D}|}$ and $r_t \in \mathcal{D}^{|\mathcal{Z}|}$.

Assumption 5.1(i) imposes that $\mathcal{Y}$ be closed and bounded, which ensures that the partition $\{B_{\ell, L}\}_{\ell=1}^L$ can be chosen so that each set $B_{\ell, L}$ is bounded. In Assumption 5.1(ii), we further restrict attention to the case in which g is continuous. As we discuss in Remark 5.1 below, some forms of discontinuity can be accommodated, though they can require specific choices of partitions $\{B_{\ell, L}\}_{\ell=1}^L$ that satisfy restrictions beyond those imposed by Assumption 5.1. Assumption 5.1(iii) imposes that the partition $\{B_{\ell, L}\}_{\ell=1}^L$ become finer as L becomes larger. Finally, for every treatment response type $r_t \in \mathcal{D}^{|\mathcal{Z}|}$, Assumption 5.1(iv) imposes two restrictions on the corresponding set $\mathcal{R}'_o(r_t)$ of outcome response types consistent with $\mathcal{R}'$: first, $\mathcal{R}'_o(r_t)$ must be closed; second, its intersection with B_{c_o} must be connected for every $c_o \in \mathcal{L}^{|\mathcal{D}|}$. When $\mathcal{R}'_o(r_t)$ is closed and convex for all response types r_t , Assumption 5.1(iv) is satisfied whenever the sets $\{B_{\ell, L}\}_{\ell=1}^L$ are chosen to be convex. In applications in which $\mathcal{Y}$ is a compact interval in $\mathbf{R}$, examples of $\mathcal{R}'_o(r_t)$ being closed and convex include, for instance, applications in which $\mathcal{R}'$ does not restrict R_o (i.e., $\mathcal{R}'_o(r_t) = \mathcal{Y}^{|\mathcal{D}|}$) or $\mathcal{R}'$ imposes (weak) monotonicity restrictions on R_o (i.e., $\mathcal{R}'_o(r_t) \equiv \{r_o \in \mathcal{Y}^{|\mathcal{D}|} : y(d) \geq y(d') \text{ when } d \geq d'\}$). In contrast, Assumption 5.1(iv) fails when we impose (strong) monotonicity restrictions on R_o (i.e., $\mathcal{R}'_o(r_t) \equiv \{r_o \in \mathcal{Y}^{|\mathcal{D}|} : y(d) > y(d') \text{ when } d > d'\}$) because $\mathcal{R}'_o(r_t)$ then fails to be closed.

Assumption 5.1 enables us to relate the inference problem with general outcomes to the problem of testing whether $\beta_L(P) = A_L x$ for some $x \geq 0$ and matrix A_L . For ease of exposition, and to focus on the main ideas, we describe A_L in the special case in which the outcome response types are unrestricted so that $\mathcal{R} = \mathcal{Y}^{|\mathcal{D}|} \times \mathcal{R}_t$ and $\mathcal{R}' = \mathcal{Y}^{|\mathcal{D}|} \times \mathcal{R}'_t$. The more general case in which outcome response types are restricted is discussed in the appendix. The matrix A_L has $(|\mathcal{M}_L| + 3)$ rows and $(L^{|\mathcal{D}|} \times |\mathcal{R}_t| + 2)$ columns. Indexing the first $|\mathcal{M}_L|$ rows by $(\ell, d, z) \in \mathcal{M}_L$ and first $L^{|\mathcal{D}|} \times |\mathcal{R}_t|$ columns by (c_o, r_t) we then set

$$A_{L, (\ell, d, z), (c_o, r_t)} = I\{c_o(d) = \ell, d(z) = d\}$$

$$A_{L, |\mathcal{M}_L|+1, (c_o, r_t)} = 1$$

and let the final two entries of the j^{th} row of A_L be zero for any $j \leq |\mathcal{M}_L| + 1$. The upper $(|\mathcal{M}_L| + 1) \times (L^{|\mathcal{D}|} |\mathcal{R}_t|)$ submatrix of A_L therefore matches the structure of the first

$|\mathcal{M}| + 1$ rows of the matrix A in Theorem 4.1 when the data is given by $(\mathbb{L}, D, Z)$ (instead of (Y, D, Z)). The final two rows of A_L impose the hypothesized value on $\theta(Q)$ by setting

$$\begin{aligned} A_{L, |\mathcal{M}_L|+2, (c_o, r_t)} &= \left(\sup_{r_o \in B_{c_o}} g(r_o, r_t) - \theta_0 \right) I\{r_t \in \mathcal{R}'_t\} \\ A_{L, |\mathcal{M}_L|+3, (c_o, r_t)} &= \left(\inf_{r_o \in B_{c_o}} g(r_o, r_t) - \theta_0 \right) I\{r_t \in \mathcal{R}'_t\} \end{aligned} \quad (26)$$

and letting the last two entries of the $|\mathcal{M}_L| + 2$ row be $(-1, 0)$ and the last two entries of the $|\mathcal{M}_L| + 3$ row be $(0, 1)$. Intuitively, this construction approximates $E_Q[(g(R) - \theta_0)I\{R \in \mathcal{R}'\}]$, which equals zero under the null hypothesis, from above and below with the approximation becoming more accurate as L becomes larger and the sets B_{c_o} smaller.

The next theorem may be seen as a generalization of Theorem 4.1. We emphasize that the result relies on a general formulation of A_L that allows for restrictions on R_o .

Theorem 5.1. *Let $\mathbf{Q}$ be the set of all distributions of (R, Z) satisfying Assumptions 2.1 and 2.2 and $\mathbf{P}$ be the set of all distributions of (Y, D, Z) satisfying $P\{(\mathbb{L}, D, Z) \in \mathcal{M}_L\} = 1$ for all L , and $P\{Z = z\} > 0$ for all $z \in \mathcal{Z}$. Further suppose $\{B_{\ell, L}\}_{\ell=1}^L$ is a partition of $\mathcal{Y}$, and g in the definition of $\theta(Q)$ in (2) is bounded. Then, for $\beta_L(P)$ as in (25) and a matrix A_L defined in Appendix A.5, it follows*

$$\mathbf{P}_0 \subseteq \{P \in \mathbf{P} : A_L x = \beta_L(P) \text{ for some } x \geq 0\} \equiv \mathbf{P}_{0,L} . \quad (27)$$

If in addition Assumption 5.1 is satisfied and $\mathbf{P}_0$ is not empty, then it follows that

$$\limsup_{L \rightarrow \infty} \mathbf{P}_{0,L} \equiv \bigcap_{M=1}^{\infty} \bigcup_{L \geq M} \mathbf{P}_{0,L} \subseteq \text{cl}_{\mathbf{P}}(\mathbf{P}_0) ,$$

where $\text{cl}_{\mathbf{P}}(\mathbf{P}_0)$ denotes the closure of $\mathbf{P}_0$ in $\mathbf{P}$ under weak convergence.

The first part of Theorem 5.1 shows that $\mathbf{P}_0 \subseteq \mathbf{P}_{0,L}$ for $\mathbf{P}_{0,L}$ as defined in (27). Therefore, any (asymptotically) α -level test of the null hypothesis that $P \in \mathbf{P}_{0,L}$ is also an (asymptotically) α -level test of the null hypothesis that $P \in \mathbf{P}_0$. Importantly, this conclusion does not depend on any of the regularity conditions imposed in Assumption 5.1. Any such test of the null hypothesis that $P \in \mathbf{P}_{0,L}$, however, may fail to reject for certain $P \in \mathbf{P} \setminus \mathbf{P}_0$ — i.e., it is not consistent against alternatives $P \in \mathbf{P}_{0,L} \setminus \mathbf{P}_0$. The second part of Theorem 5.1 addresses this concern. In particular, it establishes that, under Assumption 5.1, the limit superior of $\mathbf{P}_{0,L}$ is included in the closure of the set $\mathbf{P}_0$. The second part of Theorem 5.1 therefore shows that, as our discretization becomes finer, we are able to detect whether P belongs to $\mathbf{P}_0$ or not (up to closure). We also note that, in analogy to Remark 4.1, these conclusions continue to hold if we remove linearly dependent restrictions from A_L and $\beta_L(P)$ — see Lemma A.3

in the Appendix for a formal statement.

The main implication of Theorem 5.1 is that, with a general outcome Y , we still can test the null hypothesis of interest by testing whether P is such that $A_L x = \beta_L(P)$ for some $x \geq 0$. The latter null hypothesis can be tested by using the test defined in (18) for the finite support case but replacing $\hat{\beta}_n$ with $\hat{\beta}_{L,n} \equiv \beta_L(\hat{P}_n)$ and A with A_L . Implementing such a test requires selecting the partition $\{B_{\ell,L}\}_{\ell=1}^L$. In our simulations, we set the partition to consist of L equal-length intervals of the support of Y . Such a choice is consistent with Assumption 5.1(iii) and, under regularity conditions on the conditional density of Y given (D, Z) , with the requirements of Theorem 4.2 provided $(L/n) \times |\mathcal{D}| \times |\mathcal{Z}|$ tends to zero (up to logs). In particular, if $\mathcal{D}$ and $\mathcal{Z}$ are fixed, then we require L/n to tend to zero (up to logs), allowing the partition to quickly become finer with the sample size. In this regard, Theorem 5.1 suggests that, for power considerations, it is desirable to set L to be large. Two concerns about setting L to be large are the increased computational cost and the potential for our Gaussian approximations to become inaccurate. Because $\hat{\beta}_{L,n}$ is a vector of (conditional) empirical probabilities and the Gaussian approximation to a Binomial distribution fails for “small” probabilities, a simple rule of thumb for ensuring that L is not “too large” is to check that all but the last three coordinates of the bootstrap draws of $\hat{\beta}_{L,n}^*$ are strictly positive.

While we have focused our discussion on hypothesis testing, we again emphasize that our results readily deliver confidence regions. Specifically, test inversion arguments imply that the confidence region in (19), implemented with A_L and $\hat{\beta}_{L,n}$ in place of A and $\hat{\beta}_n$, is valid in the sense that it covers the parameter of interest with asymptotic probability of at least $1 - \alpha$ uniformly in $P \in \mathbf{P}$.

Remark 5.1. An important example of a parameter of interest that corresponds to a discontinuous g , and hence violates Assumption 5.1(ii), is the probability of an event conditional on generalized principal strata. However, in this case, g is piecewise constant and the conclusion of the second part of Theorem 5.1 continues to hold if we redefine A_L suitably. As an illustration, we continue to consider the special case in which the outcome response types are unrestricted so that $\mathcal{R} = \mathcal{Y}^{|\mathcal{D}|} \times \mathcal{R}_t$ and $\mathcal{R}' = \mathcal{Y}^{|\mathcal{D}|} \times \mathcal{R}'_t$. The general case can be handled similarly. Note by an inspection of the proof of Theorem 5.1 that a solution to the linear system in (27) is a vector whose coordinates are the probabilities of the cells $B_{c_o} \times \{r_t\}$. If $\mathcal{Y}^{|\mathcal{D}|} = \bigcup_{1 \leq j \leq K} \mathcal{S}_j$ and $g(\cdot, r_t)$ is constant on $\mathcal{S}_j$ for each r_t and $1 \leq j \leq K$, then we may redefine A_L so that the solution to the linear system is instead a vector of probabilities of $(B_{c_o} \cap \mathcal{S}_j) \times \{r_t\}$ for $1 \leq j \leq K$. ■

Remark 5.2. In a binary treatment setting with $R_o = (Y(0), Y(1))$, for example, we may define the neighborhood $B_\epsilon(q) \equiv \{y : |y - q| \leq \epsilon\}$ for a given point q and $\epsilon > 0$, and set

$\mathcal{R}' = \{(r_o, r_t) \in \mathcal{R} : y(0) \in B_\epsilon(q)\}$ and $g(r) = y(1)$ to conduct inference on

$$E_Q[Y(1) \mid Y(0) \in B_\epsilon(q)] . \quad (28)$$

Our framework applies to (28) provided $Q\{Y(0) \in B_\epsilon(q)\}$ is positive. Conducting inference on $E_Q[Y(1) \mid Y(0) = q]$, however, additionally requires conditions such as smoothness restrictions even if a sample $\{R_i\}_{i=1}^n$ is directly observed (Bertanha and Moreira, 2020). An explicit bound K on the derivative of the conditional mean, for example, allows us to obtain a confidence region for $E_Q[Y(1) \mid Y(0) = q]$ by simply expanding a confidence region for (28), obtained through our framework, by $K\epsilon$. ■

Finally, we note that Theorem 5.1 provides a useful characterization of whether the model is correctly specified in the sense that $\mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset$. As in Section 4, we may set $g = 1$, $\theta_0 = 1$, and $\mathcal{R}' = \mathcal{R}$, which implies $\mathbf{P}_0$ then equals $\{P \in \mathbf{P} : \mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset\}$. Under these choices, the equalities in the last two rows of the equation $A_L x = \beta_L(P)$ in (27) are always satisfied and we therefore let $\tilde{A}_L$ denote all but the last two rows and last two columns of A_L and $\tilde{\beta}_L(P)$ all but the last two rows of $\beta_L(P)$, respectively. Our next corollary to Theorem 5.1 formalizes these observations and provides the basis for conducting a specification test with general outcomes Y .

Corollary 5.1. *Let $\mathbf{Q}$ be the set of all distributions of (R, Z) satisfying Assumptions 2.1 and 2.2 and $\mathbf{P}$ be the set of all distributions of (Y, D, Z) satisfying $P\{(\mathbb{L}, D, Z) \in \mathcal{M}_L\} = 1$ for all L , and $P\{Z = z\} > 0$ for all $z \in \mathcal{Z}$. Then, for $g = 1$, $\theta_0 = 1$, and $\mathcal{R}' = \mathcal{R}$,*

$$\mathbf{P}_0 = \{P \in \mathbf{P} : \mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset\} \subseteq \{P \in \mathbf{P} : \tilde{A}_L x = \tilde{\beta}_L(P) \text{ for some } x \geq 0\} \equiv \tilde{\mathbf{P}}_{0,L} .$$

If in addition Assumption 5.1 holds and $\mathbf{P}_0$ is not empty, then $\limsup_{L \rightarrow \infty} \tilde{\mathbf{P}}_{0,L} \subseteq \text{cl}_{\mathbf{P}}(\mathbf{P}_0)$, where $\text{cl}_{\mathbf{P}}(\mathbf{P}_0)$ denotes the closure of $\mathbf{P}_0$ in $\mathbf{P}$ under weak convergence.

6 Simulations

In this section, we present simulation results that illustrate the finite-sample performance of our proposed inference methods. Throughout, we let $\mathcal{D} = \mathcal{Z} = \{0, 1, 2\}$ while setting $\mathcal{Y}$ to be discrete in Sections 6.1 and 6.3 and an interval in Section 6.2. The parameters of interest in Sections 6.1 and 6.2 are

$$\theta_{\text{ATE}}(Q) = E_Q[Y(2) - Y(1) \mid R \in \mathcal{R}'] \quad \text{and} \quad \theta_{\text{Prob}}(Q) = Q\{Y(2) \geq Y(1) \mid R \in \mathcal{R}'\} , \quad (29)$$

where $\mathcal{R}' = \{r \in \mathcal{R} : (d(0), d(1), d(2)) = (0, 1, 2)\}$ and recall $r = (r_o, r_t)$ with $r_o = (y(d) : d \in \mathcal{D})$ and $r_t = (d(z) : z \in \mathcal{Z})$. We consider the following two specifications of $\mathcal{R}$:

One-sided noncompliance: $\mathcal{R}$ is as specified in Example 3.1 and denoted by $\mathcal{R}_{1s}$, with the corresponding model $\mathbf{Q}_{1s} = \{Q : Q \text{ satisfies Assumptions 2.1 and 2.2 with } \mathcal{R} = \mathcal{R}_{1s}\}$;

Encouragement design: $\mathcal{R}$ is as specified in Example 3.2 and denoted by $\mathcal{R}_{\text{enc}}$, with the corresponding model $\mathbf{Q}_{\text{enc}} = \{Q : Q \text{ satisfies Assumptions 2.1 and 2.2 with } \mathcal{R} = \mathcal{R}_{\text{enc}}\}$.

In Section 6.3, we consider $\mathcal{R}' = \mathcal{R} = \mathcal{R}_{\text{enc}}$ and $\theta_{\text{ATE}}(Q)$, so closed-form expressions of the identified set are readily available, and compare our method with existing approaches based on such closed-form expressions.

Throughout, we implement our test using the `lpinfer` package in R with the default choice of penalty λ_n (discussed in Section 4), five hundred bootstrap draws, and using the specification of A that removes linearly dependent restrictions (as described in Remark 4.1). In unreported simulations we find that the results are nearly unaffected by the exact identity of the removed restrictions, and that a procedure based on A without removing such restrictions also controls size but delivers slightly larger confidence regions. All results are averaged over three thousand replications, which were run in parallel using 5 cores of an AMD EPYC 9335 processor and 20 GB of memory. The details of the distributions employed in the simulation designs can be found in the online repository at <https://github.com/edward-vytlacil/STRATA-ATE/blob/main/replication.pdf>.

6.1 Discrete Outcomes

We first present simulation results when $\mathcal{Y} = \{-1, 0, 1\}$. For each model, $\mathbf{Q}_{1s}$ and $\mathbf{Q}_{\text{enc}}$, we select a distribution Q . The chosen distributions Q are described in detail in the online repository. Employing these selected distributions Q , we then repeatedly draw i.i.d. samples (Y_i, D_i, Z_i) , $i = 1, \dots, n$ from $P = QT^{-1}$ (i.e., by using relation (1)). For each replication, we obtain the confidence intervals for the parameters $\theta_{\text{ATE}}(Q)$ and $\theta_{\text{Prob}}(Q)$ by inverting the test of Theorem 4.2 implemented at the five percent level.

Table 2 reports results for sample sizes of $n = 3000$ and $n = 12000$. For each parameter and each model, we report the average length of the 95% level confidence intervals and the coverage rates for the true parameter and the endpoints of the identified set. We also report the lengths of the identified sets, $|\Theta_0(P, \mathbf{Q}_{1s})|$ and $|\Theta_0(P, \mathbf{Q}_{\text{enc}})|$. The difference between the average length of the confidence intervals and the length of the identified set can be attributed to sampling uncertainty. To illustrate the computational cost, we also report the average runtime required to compute one confidence region, as well as the dimension of the matrix A , which corresponds to the number of constraints and number of decision variables in each of the linear systems being tested. For clarity, we report $\dim(A)$ without removing linearly dependent restrictions.

One-Sided Noncompliance Design								
	$\dim(A)$	$ \Theta_0(P, \mathbf{Q}_{1s}) $	n	95% CI length	Cov. θ_0	Cov. min Θ_0	Cov. max Θ_0	Runtime (s.d.)
$\theta_{\text{ATE}}(Q)$	$(15+2) \times 108$	0.881	3000	1.035	1.000	0.948	0.947	97.75 (18.62)
			12000	0.961	1.000	0.955	0.954	109.15 (21.66)
$\theta_{\text{Prob}}(Q)$	$(15+2) \times 108$	0.381	3000	0.414	1.000	0.945	1.000	67.85 (12.85)
			12000	0.397	1.000	0.950	1.000	75.85 (15.15)

Encouragement Design								
	$\dim(A)$	$ \Theta_0(P, \mathbf{Q}_{\text{enc}}) $	n	95% CI length	Cov. θ_0	Cov. min Θ_0	Cov. max Θ_0	Runtime (s.d.)
$\theta_{\text{ATE}}(Q)$	$(27+2) \times 270$	2.836	3000	3.340	1.000	0.958	0.945	85.35 (21.41)
			12000	3.083	1.000	0.945	0.950	97.91 (24.89)
$\theta_{\text{Prob}}(Q)$	$(27+2) \times 270$	0.894	3000	0.984	1.000	0.951	1.000	63.97 (15.89)
			12000	0.969	1.000	0.947	1.000	73.28 (18.65)

Table 2: Summary results for confidence intervals under $Q \in \mathbf{Q}_{1s}$ and $Q \in \mathbf{Q}_{\text{enc}}$ as specified in the online repository. Each row reports length of the identified set, average length of the 95% confidence intervals, coverage rate for the true parameter, coverage rates for the lower and upper endpoints of the identified set, and average runtime in seconds for obtaining one confidence region (with standard deviation of runtime in parentheses).

In these designs, the confidence interval covers the true parameter with probability one in all cases. This phenomenon is to be expected because in each case the true value of the parameter lies in the interior of the identified set. For example, in the one-sided noncompliance design, the identified set $\Theta_0(P, \mathbf{Q}_{1s})$ for $\theta_{\text{ATE}}(Q)$ equals $[0.148, 1.030]$, while $\theta_{\text{ATE}}(Q)$ itself equals 0.492 (see Table 8 in Appendix E). On the other hand, we obtain nominal coverage for the lower and upper endpoints of the identified set in all cases except for the upper endpoint of the identified set for $\theta_{\text{Prob}}(Q)$ in both models. In the latter cases, the upper endpoints of both the identified set and the confidence interval equal the logical bound of one for probabilities. Computational costs for obtaining a single confidence interval are low and range between one and two minutes for these designs. While we generally expect the computational cost to increase with the dimension of A , we see in Table 2 that it is also influenced by the design.

6.2 General-Valued Outcomes

To examine the performance of our procedure when Y is continuous, we next set $\mathcal{Y} = [-5, 5]$ while retaining $\mathcal{D} = \mathcal{Z} = \{0, 1, 2\}$. We focus on the one-sided noncompliance design (i.e., $\mathcal{R} = \mathcal{R}_{1s}$) and the parameter $\theta_{\text{ATE}}(Q)$ as defined in (29). As in Section 6.1, we select a distribution $Q \in \mathbf{Q}_{1s}$, specified in detail in the online repository, and repeatedly draw i.i.d. samples (Y_i, D_i, Z_i) , $i = 1, \dots, n$ from $P = QT^{-1}$.

Partition	$\dim(A_L)$	$ \Theta_{0,L}(P, \mathbf{Q}_{1s}) $	n	95% CI length	Cov. θ_0	Cov. min $\Theta_{0,L}$	Cov. max $\Theta_{0,L}$	Runtime (s.d.)
bin1 ($L = 2$)	$(10 + 3) \times$ $(32 + 2)$	12.385	3000	12.966	1.000	0.946	0.955	103.91 (29.14)
			12000	12.675	1.000	0.959	0.956	115.60 (33.48)
bin2 ($L = 5$)	$(25 + 3) \times$ $(500 + 2)$	5.788	3000	6.100	1.000	0.987	0.987	119.80 (33.97)
			12000	5.920	1.000	0.982	0.979	136.95 (39.52)
bin3 ($L = 10$)	$(50 + 3) \times$ $(4000 + 2)$	3.668	3000	4.080	1.000	0.996	0.997	206.50 (56.13)
			12000	3.848	1.000	0.995	0.995	238.83 (65.61)

Table 3: Summary results for confidence intervals under $Q \in \mathbf{Q}_{1s}$ as specified in the online repository. Each row includes the value of L , dimension of A_L , length of $\Theta_{0,L}(P, \mathbf{Q}_{1s})$, average length of the 95% confidence intervals, coverage rate for the true $\theta_{ATE}(Q)$, coverage rates for the lower and upper endpoints of $\Theta_{0,L}(P, \mathbf{Q}_{1s})$, and average runtime in seconds for obtaining one confidence region (with standard deviation of runtime in parentheses).

Implementing our inference approach with continuous outcomes requires a choice of partition of the support of Y . We first consider the following three partitions of $\mathcal{Y}$:

$$\text{bin1: } \mathcal{Y} = [-5, 0) \cup [0, 5];$$

$$\text{bin2: } \mathcal{Y} = [-5, -3) \cup [-3, -1) \cup \dots \cup [3, 5];$$

$$\text{bin3: } \mathcal{Y} = [-5, -4) \cup [-4, -3) \cup \dots \cup [4, 5].$$

In Table 3, we report results for sample sizes $n = 3000$ and $n = 12000$. For each level of discretization, we report the average length of the 95% level confidence intervals and the coverage rates of the true parameter and the endpoints of $\Theta_{0,L}(P, \mathbf{Q}_{1s})$, which is the set of values that are compatible with P satisfying the restrictions corresponding to a partition of size L . In order to examine computational cost, we again report average runtime and the dimension of the matrix A_L in Theorem 5.1.

As in Table 2, the confidence intervals cover the parameter with probability one for each level of discretization. This phenomenon is again to be expected because in each case the true value of the parameter lies strictly in the interior of $\Theta_{0,L}(P, \mathbf{Q}_{1s})$ (see Table 8 in Appendix E). For $L = 2$, we continue to obtain nominal coverage for the lower and upper endpoints of the identified sets. For $L = 5$ and $L = 10$, however, the coverage rates of the lower and upper endpoints of $\Theta_{0,L}(P, \mathbf{Q}_{1s})$ exceed the nominal level. We cannot calculate the identified set $\Theta_0(P, \mathbf{Q}_{1s})$ explicitly, but because $\Theta_0(P, \mathbf{Q}_{1s}) \subseteq \Theta_{0,L}(P, \mathbf{Q}_{1s})$ for each L by Theorem 5.1, the coverage rates for the endpoints of $\Theta_0(P, \mathbf{Q}_{1s})$ must also exceed the nominal level for $L = 5$ and $L = 10$. Increasing L decreases the length of the identified set and the average length of our confidence intervals. However, because the number of constraints and decision variables grows as the partition becomes finer, the computational cost also increases with L .

To further understand the tradeoff between computation time and the length of the confidence intervals, we increase L to 20 (**bin4**) and 50 (**bin5**). Repeated simulation draws

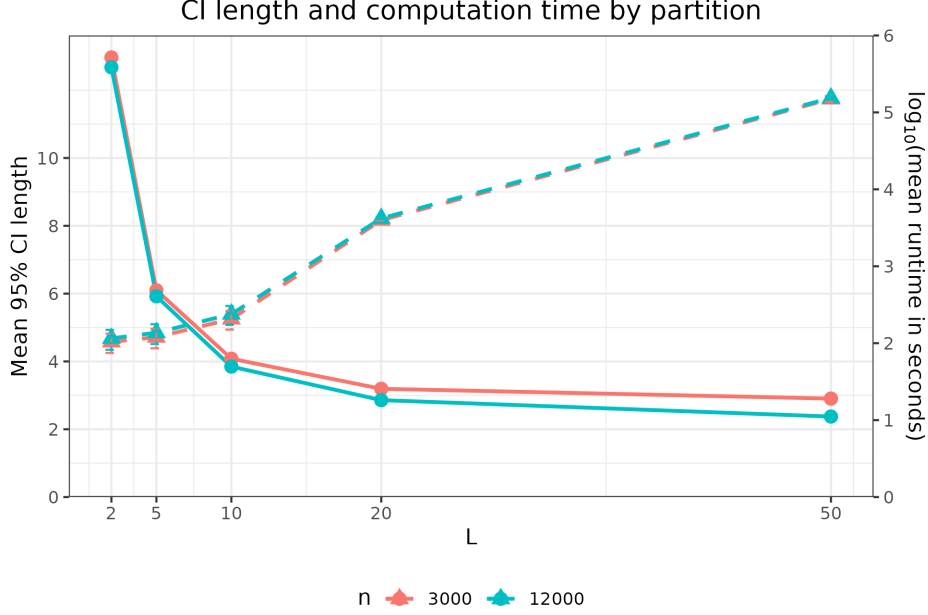


Figure 1: Average length and computation time of a confidence interval for $\theta_{\text{ATE}}(Q)$ for different choices of partition size L . Results for $L = 2, 5, 10$ are averaged over 3000 replications (also reported in Table 3), while results for $L = 20, 50$ correspond to two replications.

in these designs are computationally prohibitive, so we instead report results based on two replications in Figure 1. As expected, when L further increases, the length of the confidence interval further decreases (with the caveat that we only compute the confidence interval in one replication). However, while the marginal decrease in confidence interval length quickly diminishes with L , the corresponding computational cost increases rapidly in L . For $n = 12000$, for example, the computational cost for $L = 20$ and $L = 50$ is approximately eleven and five hundred times larger than for $L = 10$. This significant increment in computational cost reflects the important change in the dimension of A_L , which equals $103 \times 32,002$ and $253 \times 500,002$ for $L = 20$ and $L = 50$ (vs. $53 \times 4,002$ for $L = 10$).

6.3 Comparison with Existing Approaches

We conclude our simulation study by comparing our methods to existing approaches based on closed-form expressions of the identified sets for the parameters of interest. Before proceeding, we emphasize that closed-form expressions are not available for most parameters considered in this paper. The comparison in this subsection serves to illustrate the advantages and disadvantages of our methods when closed-form expressions happen to be available.

We retain $\mathcal{D} = \mathcal{Z} = \{0, 1, 2\}$ and consider either $\mathcal{Y} = \{-1, 0, 1\}$ or $\mathcal{Y} = \{-1, 0, 1, 2\}$. We focus on the encouragement design (i.e., $\mathcal{R} = \mathcal{R}_{\text{enc}}$) in Section 6.1 and set $\mathcal{R}' = \mathcal{R}$, so that

the parameter of interest is the average treatment effect

$$\theta(Q) = E_Q[Y(2) - Y(1)].$$

Following our analysis of Sections 6.1 and 6.2, for each $\mathcal{Y}$, we again fix a distribution $Q \in \mathbf{Q}_{\text{enc}}$, and draw i.i.d. samples (Y_i, D_i, Z_i) , $i = 1, \dots, n$ through the relation (1) — i.e., by drawing from $P = QT^{-1}$. The Q for $\mathcal{Y} = \{-1, 0, 1\}$ is the same $Q \in \mathbf{Q}_{\text{enc}}$ that we used in Section 6.1. Details of Q for both $\mathcal{Y}$ can be found in the online repository.

In this case, closed-form expressions of the identified set are readily available. Letting y^L and y^U denote the minimum and maximum of the elements of the support $\mathcal{Y}$, it follows from Corollary 3.1 in Bai et al. (2026) that $\Theta_0(P, \mathbf{Q}_{\text{enc}}) = [L(P), U(P)]$, where

$$\begin{aligned} L(P) &= \sum_{y \in \mathcal{Y}} yP\{Y = y, D = 2|Z = 2\} + y^L P\{D \neq 2|Z = 2\} \\ &\quad - \sum_{y \in \mathcal{Y}} yP\{Y = y, D = 1|Z = 1\} - y^U P\{D \neq 1|Z = 1\}, \\ U(P) &= \sum_{y \in \mathcal{Y}} yP\{Y = y, D = 2|Z = 2\} + y^U P\{D \neq 2|Z = 2\} \\ &\quad - \sum_{y \in \mathcal{Y}} yP\{Y = y, D = 1|Z = 1\} - y^L P\{D \neq 1|Z = 1\}. \end{aligned}$$

We compare our proposed method in Section 4 based on the implicit characterization of the identified set (labeled “FSST latent”) with a procedure that exploits this closed-form expression for the identified set. Specifically, we compare our method to a confidence region obtained through test inversion of a test that the moment inequalities (i) $\theta_0 - L(P) \geq 0$ and (ii) $U(P) - \theta_0 \geq 0$ hold. To help isolate the effect of having a closed-form solution for the identified set in our comparison, we again employ the procedure in Fang et al. (2023) (labeled “FSST ineq”) to test these moment inequalities. For this purpose, we note that the moment inequalities can be expressed as the restriction that $\beta(P) = Ax$ for some $x \geq 0$ by setting $\beta(P) = (\theta_0 - L(P), U(P) - \theta_0)'$ and A to be the 2×2 identity matrix.

Table 4 reports the properties of 95% confidence intervals obtained through “FSST latent” and “FSST ineq” for samples of size $n = 3000$ and $n = 12000$. For both the design with $\mathcal{Y} = \{-1, 0, 1\}$ and the design with $\mathcal{Y} = \{-1, 0, 1, 2\}$, the confidence regions for both procedures exhibit remarkably similar average lengths and coverage properties. However, the computational cost of “FSST latent” is larger than that of “FSST ineq,” reflecting the larger dimension of A in the former procedure. In Table 9 in Appendix E, we consider a related design and find the “FSST ineq” confidence intervals are sometimes shorter than those based on “FSST latent” while retaining a lower computational cost. These results illustrate that, when available, methods based on closed-form expressions can be preferable,

Encouragement Design with $\mathcal{Y} = \{-1, 0, 1\}$								
Method	$\dim(A)$	$ \Theta_0(P, \mathbf{Q}_{\text{enc}}) $	n	95% CI length	Cov. θ_0	Cov. min Θ_0	Cov. max Θ_0	Runtime (s.d.)
FSST latent	$(27 + 2) \times 270$	1.061	3000	1.180	1.000	0.953	0.953	91.58 (30.03)
			12000	1.120	1.000	0.953	0.956	108.20 (36.17)
FSST ineq	2×2	1.061	3000	1.178	1.000	0.953	0.952	83.66 (27.49)
			12000	1.119	1.000	0.951	0.955	101.27 (33.64)
Encouragement Design with $\mathcal{Y} = \{-1, 0, 1, 2\}$								
Method	$\dim(A)$	$ \Theta_0(P, \mathbf{Q}_{\text{enc}}) $	n	95% CI length	Cov. θ_0	Cov. min Θ_0	Cov. max Θ_0	Runtime (s.d.)
FSST latent	$(36 + 2) \times 640$	2.392	3000	2.562	1.000	0.949	0.952	122.92 (31.69)
			12000	2.476	1.000	0.951	0.960	147.48 (37.59)
FSST ineq	2×2	2.392	3000	2.562	1.000	0.952	0.952	101.97 (24.86)
			12000	2.476	1.000	0.952	0.960	125.77 (30.95)

Table 4: Summary results for confidence intervals under $Q \in \mathbf{Q}_{\text{enc}}$ for $\mathcal{Y} = \{-1, 0, 1\}$ and $\mathcal{Y} = \{-1, 0, 1, 2\}$. Each row reports the length of the identified set, average length of the 95% confidence intervals, coverage rate of the true parameter, coverage rates of the lower and upper endpoints of the identified set, and average runtime in seconds for obtaining one confidence region (with standard deviation of runtime in parentheses).

especially for larger specifications of A . We again emphasize, however, that closed-form expressions are not available for most of the parameters studied in this paper. We also note that “FSST latent” is a valid test jointly for the value of the parameter as well as correct model specification, whereas “FSST ineq” is only for testing the value of the parameter and only when the model is correctly specified.

7 Impact of Therapy on Reducing Crime

In an important study, [Blattman et al. \(2017\)](#) examined the impact of cognitive behavioral therapy and cash transfers on criminal behavior through an RCT conducted in Liberia. Their analysis showed that intent-to-treat effects of cash transfers or therapy alone led to reductions in criminal behavior in the short term (2–5 weeks following randomization), but that these effects disappeared over the long term. In contrast, offering both cash transfers and therapy led to intent-to-treat effects that persisted in the long term (12–13 months following randomization), which suggests an unexpected and important long-term complementarity between these treatments in reducing criminal behavior.

We employ our theoretical results to go beyond intent-to-treat effects in [Blattman et al. \(2017\)](#) and examine whether the long-term importance of receiving both cash and therapy (vs. only cash or only therapy) is present in average and distributional causal effects of treatment receipt as well, and, if so, for which subpopulations. The experiment in [Blattman](#)

D	Z			
	0	t	c	tc
0	220	28	6	4
t	0	252	0	0
c	0	0	244	18
tc	0	0	0	227

Table 5: Cell counts for each pair of values of the treatment and the instrument.

et al. (2017) employed a 2×2 factorial design with the factors being cognitive behavioral therapy and cash transfer. We use $Z \in \mathcal{Z} = \{0, t, c, tc\}$ to denote the assignment to control, therapy, cash, and both therapy and cash, and $D \in \mathcal{D} = \{0, t, c, tc\}$ to denote the take-up³ of the corresponding treatment. The sample size is $n = 999$. Table 5 presents the cell count of each pair of values of the treatment and the instrument.

In order to focus on the benefits of receiving both treatments, we examine the incremental effect of receiving therapy in addition to cash and the incremental effect of receiving cash in addition to therapy. Among a wide variety of outcomes related to crime and violence considered by Blattman et al. (2017), we focus on the binary indicator of whether they engaged in any drug-selling activities in the short term and the long term.

Throughout our analysis, we maintain the assumption that noncompliance for either therapy or cash is one-sided. Because the design is factorial, this means that if the subject is assigned both therapy and cash, then they can choose neither, one of them, or both. To formally map this restriction into our notation, we set $\mathcal{R}$ to equal

$$\mathcal{R}_{1s} = \{(r_o, r_t) : d(0) = 0, d(t) \in \{0, t\}, d(c) \in \{0, c\}, d(tc) \in \{0, t, c, tc\}\},$$

where recall $r_o = (y(d) : d \in \mathcal{D})$ and $r_t = (d(z) : z \in \mathcal{Z})$. We also consider complementing the one-sided noncompliance restriction with the assumption of harmless treatment — i.e., that receiving cash transfers, therapy, or both, does not induce individuals into selling drugs. Since $Y(d)$ is here an indicator for whether the individual sold drugs, imposing the no-harm restriction in addition to one-sided noncompliance corresponds to setting $\mathcal{R}$ to equal

$$\mathcal{R}_{1s,nh} = \{(r_o, r_t) \in \mathcal{R}_{1s} : y(0) \geq y(d) \geq y(tc) \text{ for } d \in \{t, c\}\}.$$

Specification tests based on Corollary 4.1 for both of these restrictions fail to reject the null hypothesis that the restrictions are valid at conventional significance levels. In the case of one-sided noncompliance, we obtain p -values of one when Y is the short-term outcome, and also of one when Y is the long-term outcome. When we additionally impose the no-harm

³We define compliance to the therapy assignment to be “attending at least 8 days of therapy,” which is the first of the two alternative definitions of compliance used in (Blattman et al., 2017, Appendix A.4).

One-Sided Noncompliance ($\mathcal{R} = \mathcal{R}_{1s}$)					
Strata ($\mathcal{R}'$)	Outcome	$\theta_{ATE}(tc, t \mathcal{R}')$	$\theta_{ATE}(tc, c \mathcal{R}')$	$\theta_{Pr}(tc, t \mathcal{R}')$	$\theta_{Pr}(tc, c \mathcal{R}')$
Unrestricted ($\mathcal{R}' = \mathcal{R}$)	Short Term	[-0.122, 0.158]	[-0.098, 0.100]	[0.000, 0.243]	[0.000, 0.213]
	Long Term	[-0.159, 0.116]	[-0.198, 0.020]	[0.000, 0.258]	[0.048, 0.298]
Pairwise Compliers ($\mathcal{R}' = \mathcal{R}_{j,k}$)	Short Term	[-0.161, 0.208]	[-0.118, 0.044]	[0.000, 0.165]	[0.000, 0.230]
	Long Term	[-0.198, 0.157]	[-0.232, -0.047]	[0.000, 0.196]	[0.052, 0.305]
Full Compliers ($\mathcal{R}' = \mathcal{R}_{full}$)	Short Term	[-0.165, 0.213]	[-0.253, 0.204]	[0.000, 0.164]	[0.000, 0.257]
	Long Term	[-0.205, 0.181]	[-0.338, 0.058]	[0.000, 0.202]	[0.000, 0.340]

One-sided Noncompliance + No-Harm Restriction ($\mathcal{R} = \mathcal{R}_{1s,nh}$)					
Strata ($\mathcal{R}'$)	Outcome	$\theta_{ATE}(tc, t \mathcal{R}')$	$\theta_{ATE}(tc, c \mathcal{R}')$	$\theta_{Pr}(tc, t \mathcal{R}')$	$\theta_{Pr}(tc, c \mathcal{R}')$
Unrestricted ($\mathcal{R}' = \mathcal{R}$)	Short Term	[-0.069, 0.000]	[-0.109, 0.000]	[0.000, 0.069]	[0.000, 0.122]
	Long Term	[-0.116, 0.000]	[-0.148, -0.024]	[0.000, 0.112]	[0.029, 0.147]
Pairwise Compliers ($\mathcal{R}' = \mathcal{R}_{j,k}$)	Short Term	[-0.078, 0.000]	[-0.131, 0.000]	[0.000, 0.091]	[0.000, 0.130]
	Long Term	[-0.145, 0.000]	[-0.169, -0.025]	[0.000, 0.142]	[0.028, 0.165]
Full Compliers ($\mathcal{R}' = \mathcal{R}_{full}$)	Short Term	[-0.062, 0.000]	[-0.139, 0.000]	[0.000, 0.063]	[0.000, 0.142]
	Long Term	[-0.146, 0.000]	[-0.183, -0.011]	[0.000, 0.147]	[0.007, 0.180]

Table 6: 95% confidence intervals for causal parameters $\theta_{ATE}(tc, d|\mathcal{R}')$ and $\theta_{Pr}(tc, d|\mathcal{R}')$ (as defined in (30)) for different subsets of generalized principal strata $\mathcal{R}'$. Pairwise complier subset $\mathcal{R}_{j,k}$ corresponds to the (j, k) in the corresponding causal parameter — e.g., $\mathcal{R}_{j,k} = \mathcal{R}_{tc,t}$ in column 3 and $\mathcal{R}_{j,k} = \mathcal{R}_{tc,c}$ in column 4. Confidence regions were obtained by inverting the test in Theorem 4.2 after removing linearly dependent rows (as in Remark 4.1), and using two thousand bootstrap draws, the default choice of penalty λ_n in the `lpinfer` R package, and the same machine configuration as in the simulations.

restriction, the p -values decrease to 0.91 when Y is the short-term outcome, and to 0.18 when Y is the long-term outcome.

For each of these sets of restrictions, and different subsets of generalized principal strata $\mathcal{R}'$, we study average and distributional causal effects by conducting inference on

$$\begin{aligned}\theta_{ATE}(tc, d|\mathcal{R}') &= E_Q[Y(tc) - Y(d)|R \in \mathcal{R}'] \\ \theta_{Pr}(tc, d|\mathcal{R}') &= Q\{Y(tc) < Y(d)|R \in \mathcal{R}'\}\end{aligned}\tag{30}$$

for $d \in \{t, c\}$. Thus, $\theta_{ATE}(tc, d|\mathcal{R}')$ reveals the average incremental effect of receiving both cash and therapy over just one of them, while $\theta_{Pr}(tc, d|\mathcal{R}')$ informs us about the corresponding distributional effects. As subsets of generalized principal strata $\mathcal{R}'$ we consider:

Unrestricted: $\mathcal{R}' = \mathcal{R}$.

Pairwise compliers: $\mathcal{R}' = \mathcal{R}_{j,k}$ for $\mathcal{R}_{j,k} = \{(r_o, r_t) \in \mathcal{R} : d(j) = j, d(k) = k\}$.

Full compliers: $\mathcal{R}' = \mathcal{R}_{full}$ for $\mathcal{R}_{full} = \{(r_o, r_t) \in \mathcal{R} : d(j) = j \text{ for all } j \in \mathcal{Z}\}$.

The first panel in Table 6 reports 95% confidence intervals for the different causal parameters while imposing only the one-sided noncompliance restriction (i.e., $\mathcal{R} = \mathcal{R}_{1s}$). We find

no evidence that receiving both cash and therapy is better in the short term than receiving cash alone or therapy alone: all our confidence regions for causal effects based on pairwise comparisons between receiving both therapy and cash against receiving cash alone or therapy alone cover zero. However, we find evidence of long-term benefits of receiving therapy in addition to cash, relative to receiving cash alone. In particular, we find a close to significant average treatment effect in the population (the p -value for testing $\theta_{\text{ATE}}(tc, c|\mathcal{R}) = 0$ is 0.126; see Table 11 in Appendix E) and that at least 5% of the population would not sell drugs in the long term when receiving both cash and therapy, but would if receiving cash alone (i.e., $\theta_{\text{Pr}}(tc, c|\mathcal{R})$). These effects are accentuated among individuals that are both tc -compliers and c -compliers (i.e., $\mathcal{R}' = \mathcal{R}_{tc,c}$), but not statistically significant among the group of full compliers (i.e., $\mathcal{R}' = \mathcal{R}_{\text{full}}$). The results for the long-term causal effects of receiving both therapy and cash, relative to receiving therapy alone, are inconclusive. While these results are compatible with the long-term positive effects of receiving both therapy and cash over receiving cash alone being driven by the benefits of therapy itself, in unreported results we find no statistically significant evidence that receiving therapy alone is beneficial over control. Our analysis therefore suggests an important complementarity in offering both cash and therapy together, in particular over offering cash alone and for the subgroup of individuals that are both tc -compliers and c -compliers.

In the second panel of Table 6, we strengthen our restrictions by imposing both one-sided noncompliance and the no-harm assumption — i.e., we set $\mathcal{R} = \mathcal{R}_{1s,nh}$. The conclusions remain largely in line with those obtained under one-sided noncompliance only, though the additional no-harm restriction proves informative in shortening our confidence intervals. For example, the confidence interval for the long-term average treatment effect of receiving both cash and therapy instead of cash alone now excludes zero both for the entire population (i.e., $\mathcal{R}' = \mathcal{R}$) and for the set of full compliers (i.e., $\mathcal{R}' = \mathcal{R}_{\text{full}}$). We caution, however, that the data are less supportive of imposing both one-sided noncompliance and no-harm than of imposing one-sided noncompliance only — recall the specification test p -values were 0.18 and 1 respectively when employing long-term outcomes.

As an alternative to the no-harm restriction, we may maintain one-sided noncompliance but instead examine causal effects conditional on not being harmed by the treatment. These parameters correspond to conditioning on generalized principal strata groups $\mathcal{R}'$ that incorporate a restriction on the outcome response type $r_o = (y(d) : d \in \mathcal{D})$. Formally, set

$$\mathcal{R}_{tc,k}^+ = \{(r_o, r_t) \in \mathcal{R} : y(tc) \leq y(k)\}$$

for $k \in \{t, c\}$ to indicate the set of individuals who see a (weakly) better outcome under treatment tc than under alternative treatment k . Notice this generalized principal stratum

One-Sided Noncompliance ($\mathcal{R} = \mathcal{R}_{1s}$)					
Strata ($\mathcal{R}'$)	Outcome	$\theta_{ATE}(tc, t \mathcal{R}')$	$\theta_{ATE}(tc, c \mathcal{R}')$	$\theta_{Pr}(tc, t \mathcal{R}')$	$\theta_{Pr}(tc, c \mathcal{R}')$
Unrestricted + No Harm ($\mathcal{R}' = \mathcal{R}_{tc,k}^+$)	Short Term	[-0.309, 0.000]	[-0.272, 0.000]	[0.000, 0.308]	[0.000, 0.273]
	Long Term	[-0.322, 0.000]	[-0.362, -0.045]	[0.000, 0.322]	[0.047, 0.361]
Pairwise Compliers + No Harm ($\mathcal{R}' = \mathcal{R}_{j,k} \cap \mathcal{R}_{tc,k}^+$)	Short Term	[-0.198, 0.000]	[-0.270, 0.000]	[0.000, 0.198]	[0.000, 0.268]
	Long Term	[-0.227, 0.000]	[-0.349, -0.048]	[0.000, 0.228]	[0.048, 0.349]
Full Compliers + No Harm ($\mathcal{R}' = \mathcal{R}_{full} \cap \mathcal{R}_{tc,k}^+$)	Short Term	[-0.200, 0.000]	[-0.312, 0.000]	[0.000, 0.197]	[0.000, 0.306]
	Long Term	[-0.238, 0.000]	[-0.395, 0.000]	[0.000, 0.234]	[0.000, 0.391]

Table 7: 95% confidence intervals for causal parameters $\theta_{ATE}(tc, d|\mathcal{R}')$ and $\theta_{Pr}(tc, d|\mathcal{R}')$. Pairwise complier subset $\mathcal{R}_{j,k}$ corresponds to the (j, k) in the corresponding causal parameter — e.g., $\mathcal{R}_{j,k} = \mathcal{R}_{tc,t}$ in column 3 and $\mathcal{R}_{j,k} = \mathcal{R}_{tc,c}$ in column 4. Restriction $\mathcal{R}_{tc,k}^+$ corresponds to the k in the corresponding causal parameter — e.g., $\mathcal{R}_{tc,k}^+ = \mathcal{R}_{tc,t}^+$ in column 3 and $\mathcal{R}_{tc,k}^+ = \mathcal{R}_{tc,c}^+$ in column 4. Confidence regions were obtained by inverting the test in Theorem 4.2 after removing linearly dependent rows (as in Remark 4.1) and using two thousand bootstrap draws, the default choice of the penalty parameter λ_n in the `lpinfer` R package, and the same machine configuration as in the simulations.

has probability one under the no-harm restriction, but not necessarily under only one-sided noncompliance. Table 7 reports causal effects for this generalized principal stratum under the one-sided noncompliance restriction only. Interestingly, the confidence interval for the average treatment effect of receiving both cash and therapy rather than cash alone, among individuals who would not be harmed by receiving both rather than cash alone (i.e., $\mathcal{R}' = \mathcal{R}_{tc,c}^+$) now includes only strictly negative (beneficial) values. With an upper endpoint of -4.5% , this confidence interval is more informative than the confidence region obtained under the no-harm assumption, and, perhaps more surprisingly, almost identical to the upper end for the average treatment effect among individuals that are tc -compliers and c -compliers. The similarity suggests that the group of individuals who are both tc -compliers and c -compliers may heavily overlap with the group of individuals who benefit from receiving therapy and cash instead of only cash. Indeed, the 95% confidence interval for $Q\{R \in \mathcal{R}_{tc,c} \mid Y(tc) \leq Y(c)\}$ has a lower endpoint of 85.1%, providing evidence of substantial overlap between individuals who would not be harmed by receiving therapy in addition to cash and individuals who are both tc -compliers and c -compliers.

References

- ALIPRANTIS, C. D. and BORDER, K. C. (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide*. 3rd ed. Springer, Berlin.
- ANDREWS, D. W. K. and SHI, X. (2013). Inference based on conditional moment inequalities. *Econometrica*, **81** 609–666.
- ANGRIST, J., LANG, D. and OREOPOULOS, P. (2009). Incentives and services for college

- achievement: Evidence from a randomized trial. *American Economic Journal: Applied Economics*, **1** 136–163.
- ANGRIST, J. D. and IMBENS, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, **90** 431–442.
- AVIS, D. and FUKUDA, K. (1991). A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra. In *Proceedings of the Seventh Annual Symposium on Computational Geometry*. Association for Computing Machinery, New York, 98–104.
- BAI, Y., HUANG, S., MOON, S., SHAIKH, A. M. and VYTLACIL, E. J. (2026). On the identifying power of generalized monotonicity for average treatment effects. *Journal of the American Statistical Association* 1–8. Advance online publication.
- BAI, Y., HUANG, S. and TABORD-MEEHAN, M. (2025). Sharp testable implications of encouragement designs. arXiv:2411.09808 [econ.EM], URL <https://arxiv.org/abs/2411.09808>.
- BAI, Y., SANTOS, A. and SHAIKH, A. M. (2022). On testing systems of linear inequalities with known coefficients. Working paper.
- BERTANHA, M. and MOREIRA, M. J. (2020). Impossible inference in econometrics: Theory and applications. *Journal of Econometrics*, **218** 247–270.
- BERTSIMAS, D. and TSITSIKLIS, J. N. (1997). *Introduction to Linear Optimization*, vol. 6 of *Athena Scientific Optimization and Computation Series*. Athena Scientific, Belmont, MA.
- BHATIA, R. (2013). *Matrix Analysis*, vol. 169 of *Graduate Texts in Mathematics*. Springer, New York.
- BHATTACHARYA, J., SHAIKH, A. M. and VYTLACIL, E. J. (2008). Treatment effect bounds under monotonicity assumptions: An application to Swan–Ganz catheterization. *American Economic Review*, **98** 351–356.
- BHATTACHARYA, J., SHAIKH, A. M. and VYTLACIL, E. J. (2012). Treatment effect bounds: An application to Swan–Ganz catheterization. *Journal of Econometrics*, **168** 223–243.
- BHULLER, M. and SIGSTAD, H. (2024). 2SLS with multiple treatments. *Journal of Econometrics*, **242** 105785.

- BLATTMAN, C., JAMISON, J. C. and SHERIDAN, M. (2017). Reducing crime and violence: Experimental evidence from cognitive behavioral therapy in Liberia. *American Economic Review*, **107** 1165–1206.
- BLOOM, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review*, **8** 225–246.
- BOGACHEV, V. I. (1998). *Gaussian Measures*, vol. 62 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI.
- CHENG, J. and SMALL, D. S. (2006). Bounds on causal effects in three-arm trials with non-compliance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68** 815–836.
- CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probability Theory and Related Fields*, **162** 47–70.
- COX, G. and SHI, X. (2023). Simple adaptive size-exact testing for full-vector and subvector inference in moment inequality models. *The Review of Economic Studies*, **90** 201–228.
- DUPAS, P. (2014). Short-run subsidies and long-run adoption of new health products: Evidence from a field experiment. *Econometrica*, **82** 197–228.
- FANG, Z. and SANTOS, A. (2019). Inference on directionally differentiable functions. *The Review of Economic Studies*, **86** 377–412.
- FANG, Z., SANTOS, A., SHAIKH, A. M. and TORGOVITSKY, A. (2021). Inference for large-scale linear systems with known coefficients. arXiv:2009.08568 [econ.EM], URL <https://arxiv.org/abs/2009.08568>.
- FANG, Z., SANTOS, A., SHAIKH, A. M. and TORGOVITSKY, A. (2023). Inference for large-scale linear systems with known coefficients. *Econometrica*, **91** 299–327.
- FRANGAKIS, C. E. and RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics*, **58** 21–29.
- GABRIEL, E. E., SACHS, M. C. and JENSEN, A. K. (2024). Sharp symbolic nonparametric bounds for measures of benefit in observational and imperfect randomized studies with ordinal outcomes. *Biometrika*, **111** 1429–1436.
- GOFF, L. (2025). When does IV identification not restrict outcomes? arXiv:2406.02835 [econ.EM], URL <https://arxiv.org/abs/2406.02835>.
- HECKMAN, J. J. and PINTO, R. (2018). Unordered monotonicity. *Econometrica*, **86** 1–35.

- HECKMAN, J. J. and SMITH, J. A. (1999). Evaluating the welfare state. In *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial Symposium* (S. Strøm, ed.). Cambridge University Press, Cambridge, 241–318.
- HECKMAN, J. J., SMITH, J. A. and CLEMENTS, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, **64** 487–535.
- HUANG, E. J., FANG, E. X., HANLEY, D. F. and ROSENBLUM, M. (2019). Constructing a confidence interval for the fraction who benefit from treatment, using randomized trial data. *Biometrics*, **75** 1228–1239.
- IMBENS, G. W. and ANGRIST, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, **62** 467–475.
- JOAG-DEV, K. and PROSCHAN, F. (1983). Negative association of random variables with applications. *The Annals of Statistics*, **11** 286–295.
- JUN, S. J. and LEE, S. (2023). Identifying the effect of persuasion. *Journal of Political Economy*, **131** 2032–2058.
- KAIDO, H. and PONOMAREV, K. (2025). Testing exclusion and shape restrictions in potential outcomes models. arXiv:2512.20851 [econ.EM], URL <https://arxiv.org/abs/2512.20851>.
- KIRKEBOEN, L. J., LEUVEN, E. and MOGSTAD, M. (2016). Field of study, earnings, and self-selection. *The Quarterly Journal of Economics*, **131** 1057–1111.
- KITAGAWA, T. (2015). A test for instrument validity. *Econometrica*, **83** 2043–2063.
- KLINE, P. and SANTOS, A. (2013). Sensitivity to missing data assumptions: Theory and an evaluation of the US wage structure. *Quantitative Economics*, **4** 231–267.
- KLINE, P. and WALTERS, C. R. (2016). Evaluating public programs with close substitutes: The case of Head Start. *The Quarterly Journal of Economics*, **131** 1795–1848.
- KWON, S. and ROTH, J. (2026). Testing mechanisms. *The Review of Economic Studies* rdag028.
- LEE, S. and SALANIÉ, B. (2026). Treatment effects with targeting instruments. *Journal of Econometrics*, **256** 106253.
- LU, J., DING, P. and DASGUPTA, T. (2018). Treatment effects on ordinal outcomes: Causal estimands and sharp bounds. *Journal of Educational and Behavioral Statistics*, **43** 540–567.

- LUENBERGER, D. G. (1997). *Optimization by Vector Space Methods*. John Wiley & Sons, New York.
- MACHADO, C., SHAIKH, A. M. and VYTLACIL, E. J. (2019). Instrumental variables and the sign of the average treatment effect. *Journal of Econometrics*, **212** 522–555.
- MANSKI, C. F. (1997). Monotone treatment response. *Econometrica*, **65** 1311–1334.
- MANSKI, C. F. and PEPPER, J. V. (2000). Monotone instrumental variables: With an application to the returns to schooling. *Econometrica*, **68** 997–1010.
- MASSART, P. (1989). Strong approximation for multivariate empirical and related processes, via KMT constructions. *The Annals of Probability*, **17** 266–291.
- MASTEN, M. A. and POIRIER, A. (2020). Inference on breakdown frontiers. *Quantitative Economics*, **11** 41–111.
- MONRAD, D. and PHILIPP, W. (1991). Nearby variables with nearby conditional laws and a strong approximation theorem for Hilbert space valued martingales. *Probability Theory and Related Fields*, **88** 381–404.
- MUNKRES, J. R. (2000). *Topology*. 2nd ed. Prentice Hall, Upper Saddle River, NJ.
- NAVJEEVAN, M., PINTO, R. and SANTOS, A. (2026). Identification in instrumental variables models: The central role of Abadie’s kappa. *Econometrica*, **94** 1095–1133.
- POLLARD, D. (2002). *A User’s Guide to Measure Theoretic Probability*, vol. 8 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.
- ROMANO, J. P. (2004). On non-parametric testing, the uniform behaviour of the t -test, and related problems. *Scandinavian Journal of Statistics*, **31** 567–584.
- SUN, Z. (2023). Instrument validity for heterogeneous causal effects. *Journal of Econometrics*, **237** 105523.
- TROPP, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, **12** 389–434.
- VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics, Springer, New York.
- VYTLACIL, E. J. (2006). Ordered discrete-choice selection models and local average treatment effect assumptions: Equivalence, nonequivalence, and representation results. *The Review of Economics and Statistics*, **88** 578–581.

Appendix A Proof of Main Results

To reduce notational clutter, throughout the appendix we suppress the dependence of $\hat{P}_n$ on n and write $\hat{P}$ instead.

A.1 Proof of Theorem 4.1

We first establish Theorem 4.1 and subsequently show that the result continues to hold if certain rows of A and $\beta(P)$ are removed, as stated in Remark 4.1.

PROOF OF THEOREM 4.1. Suppose $P \in \mathbf{P}_0$. Then, it follows that $\theta_0 \in \Theta_0(P, \mathbf{Q})$, so there exists a $Q \in \mathbf{Q}_0(P, \mathbf{Q})$ such that $\theta_0 = \theta(Q)$ and $Q\{R \in \mathcal{R}'\} > 0$. Recall $Q \in \mathbf{Q}_0(P, \mathbf{Q})$ if and only if $Q \in \mathbf{Q}$ and $P = QT^{-1}$. Since $Q \in \mathbf{Q}$, it satisfies both Assumptions 2.1 and 2.2. Because Q satisfies Assumption 2.1 and $P = QT^{-1}$, we can conclude that

$$P\{Y = y, D = d | Z = z\} = Q\{Y(d) = y, D(z) = d | Z = z\} = Q\{Y(d) = y, D(z) = d\}. \quad (31)$$

Let q be the column vector with $|\mathcal{R}|$ elements indexed by r , where the r element is $q(r) = Q\{R = r\}$. It follows from result (31) and the definition of A that the first $|\mathcal{M}|$ rows of Aq equal the first $|\mathcal{M}|$ rows of $\beta(P)$. Next, note that because $Q \in \mathbf{Q}_0(P, \mathbf{Q})$ implies Q satisfies Assumption 2.2 and hence $Q\{R \in \mathcal{R}\} = 1$, it follows that

$$\sum_{r \in \mathcal{R}} q(r) = 1,$$

and therefore the $|\mathcal{M}| + 1$ row of Aq equals the $|\mathcal{M}| + 1$ row of $\beta(P)$. Finally, since $\theta_0 = \theta(Q)$ by hypothesis, the definition of the conditional expectation yields that

$$\theta_0 = \frac{E_Q[g(R)I\{R \in \mathcal{R}'\}]}{Q\{R \in \mathcal{R}'\}}, \quad (32)$$

where recall Q satisfies $Q\{R \in \mathcal{R}'\} > 0$. Rearranging the expression in (32) we obtain that $E_Q[(g(R) - \theta_0)I\{R \in \mathcal{R}'\}] = 0$ or, equivalently,

$$\sum_{r \in \mathcal{R}'} (g(r) - \theta_0)q(r) = 0.$$

It follows from the definition of A that the $|\mathcal{M}| + 2$ row of Aq equals the $|\mathcal{M}| + 2$ row of $\beta(P)$. Hence, $\mathbf{P}_0 \subseteq \tilde{\mathbf{P}}_0$ for $\tilde{\mathbf{P}}_0 \equiv \{P \in \mathbf{P} : Ax = \beta(P) \text{ for some } x \geq 0\}$.

To conclude, it only remains to show that $\tilde{\mathbf{P}}_0 \subseteq \text{cl}_{\mathbf{P}}(\mathbf{P}_0)$ whenever $\mathbf{P}_0 \neq \emptyset$. To this end, fix any $\tilde{P} \in \tilde{\mathbf{P}}_0$. Since $Aq = \beta(\tilde{P})$ for some $q \geq 0$, $\tilde{P}\{(Y, D, Z) \in \mathcal{M}\} = 1$ and the definition

of A and $\beta(\tilde{P})$ imply there exists a $\tilde{Q} \in \mathbf{Q}$ such that $\tilde{P} = \tilde{Q}T^{-1}$ and $E_{\tilde{Q}}[g(R)I\{R \in \mathcal{R}'\}] = \theta_0 \tilde{Q}\{R \in \mathcal{R}'\}$. By assumption, $\mathbf{P}_0$ is nonempty and therefore Lemma B.14 implies $\mathbf{P}_0(\tilde{P}) \equiv \{P \in \mathbf{P}_0 : P_Z = \tilde{P}_Z\}$ is nonempty as well, where P_Z denotes the marginal distribution of Z under P . Hence, there exists a $P \in \mathbf{P}_0$ such that $P_Z = \tilde{P}_Z$, $P = QT^{-1}$ for some $Q \in \mathbf{Q}$, $Q\{R \in \mathcal{R}'\} > 0$, and $\theta(Q) = \theta_0$. Define $P_n = n^{-1}P + (1 - n^{-1})\tilde{P}$ and $Q_n = n^{-1}Q + (1 - n^{-1})\tilde{Q}$. It then follows that $P_n = Q_n T^{-1}$, $Q_n\{R \in \mathcal{R}'\} > 0$, and $E_{Q_n}[g(R)I\{R \in \mathcal{R}'\}] = \theta_0 Q_n\{R \in \mathcal{R}'\}$, which implies that $\theta(Q_n) = \theta_0$. Moreover, since $Q = Q_R \times P_Z$ and $\tilde{Q} = \tilde{Q}_R \times P_Z$ for some distributions Q_R and $\tilde{Q}_R$, it follows that $Q_n = (n^{-1}Q_R + (1 - n^{-1})\tilde{Q}_R) \times P_Z$ and therefore that $R \perp\!\!\!\perp Z$ under Q_n . We conclude that $P_n \in \mathbf{P}_0$ for each n and, since $P_n \rightarrow \tilde{P}$ as $n \rightarrow \infty$, that $\tilde{P} \in \text{cl}_{\mathbf{P}}(\mathbf{P}_0)$ implying $\tilde{\mathbf{P}}_0 \subseteq \text{cl}_{\mathbf{P}}(\mathbf{P}_0)$. ■

We next show that the conclusion of Theorem 4.1 continues to hold if we employ a subset of the restrictions implied by A . To formalize this result, recall that the first $|\mathcal{M}|$ rows of A and $\beta(P)$ are indexed by $(y, d, z) \in \mathcal{M}$. For each $z \in \mathcal{Z}$ we select an arbitrary $(y_z, d_z) \in \mathcal{M}_z$ and define A^s and $\beta^s(P)$ to be the matrix and vector obtained by removing the rows $((y_z, d_z, z) : z \in \mathcal{Z})$ from A and $\beta(P)$ respectively. Further define

$$\begin{aligned}\tilde{\mathbf{P}}_0 &\equiv \{P \in \mathbf{P} : Ax = \beta(P) \text{ for some } x \geq 0\} \\ \tilde{\mathbf{P}}_0^s &\equiv \{P \in \mathbf{P} : A^s x = \beta^s(P) \text{ for some } x \geq 0\},\end{aligned}$$

and note that Theorem 4.1 shows $\mathbf{P}_0 \subseteq \tilde{\mathbf{P}}_0$ and that $\tilde{\mathbf{P}}_0 \subseteq \text{cl}_{\mathbf{P}}(\mathbf{P}_0)$ whenever $\mathbf{P}_0 \neq \emptyset$. Our next lemma establishes that $\tilde{\mathbf{P}}_0 = \tilde{\mathbf{P}}_0^s$, which implies that the conclusion of Theorem 4.1 continues to hold if we replace A and $\beta(P)$ with A^s and β^s respectively.

Lemma A.1. *Let $\mathbf{P}$ denote the set of all distributions of (Y, D, Z) satisfying $P\{(Y, D, Z) \in \mathcal{M}\} = 1$ and $P\{Z = z\} > 0$ for all $z \in \mathcal{Z}$. Then, it follows that $\tilde{\mathbf{P}}_0 = \tilde{\mathbf{P}}_0^s$.*

PROOF OF LEMMA A.1. Since any x satisfying $Ax = \beta(P)$ must also satisfy $A^s x = \beta^s(P)$, it immediately follows that $\tilde{\mathbf{P}}_0 \subseteq \tilde{\mathbf{P}}_0^s$. For the reverse inclusion note that, by definition of $\mathcal{M}_z$ (as in (6)), for any $r = (r_o, r_t) \in \mathcal{R}$ there exists a $(y, d) \in \mathcal{M}_z$ such that $r_o = (y(d) : d \in \mathcal{D})$ and $r_t = (d(z) : z \in \mathcal{Z})$ satisfy $I\{y(d) = y, d(z) = d\} = 1$. Since such $(y, d) \in \mathcal{M}_z$ is necessarily unique, it follows from the definition of A that for any $r \in \mathcal{R}$ and $z \in \mathcal{Z}$ we have

$$\sum_{(y,d) \in \mathcal{M}_z} A_{(y,d,z),r} = 1. \quad (33)$$

In particular, letting $A_{(y,d,z),\cdot}$ denote the (y, d, z) row of the matrix A it then follows that

$$\sum_{(y,d) \in \mathcal{M}_z} A_{(y,d,z),\cdot} = A_{|\mathcal{M}|+1,\cdot} \quad (34)$$

where the equality follows from the $|\mathcal{M}| + 1$ row of the matrix A having all entries equal to one. Next select an arbitrary $P \in \tilde{\mathbf{P}}_0^s$ and note that by definition there exists an $x_0 \geq 0$ satisfying $A^s x_0 = \beta^s(P)$. Since the system $A^s x_0 = \beta^s(P)$ still contains the adding up constraint $A_{|\mathcal{M}|+1, \cdot} x_0 = 1$, result (34) allows us to conclude that

$$1 = \sum_{(y,d) \in \mathcal{M}_z} A_{(y,d,z), \cdot} x_0 = \sum_{(y,d) \in \mathcal{M}_z \setminus \{(y_z, d_z)\}} \beta_{(y,d,z)}(P) + A_{(y_z, d_z, z), \cdot} x_0, \quad (35)$$

where in the second equality we used that $A^s x_0 = \beta^s(P)$ implies that $A_{(y,d,z), \cdot} x_0 = \beta_{(y,d,z)}(P)$ for any $(y, d) \in \mathcal{M}_z \setminus \{(y_z, d_z)\}$ because those constraints remain in A^s . Moreover, since $P\{Z = z\} > 0$ for all $z \in \mathcal{Z}$ and $P\{(Y, D, Z) \in \mathcal{M}\} = 1$ due to $P \in \mathbf{P}$, we also have

$$\sum_{(y,d) \in \mathcal{M}_z} \beta_{(y,d,z)}(P) = \sum_{(y,d) \in \mathcal{M}_z} P\{Y = y, D = d | Z = z\} = 1. \quad (36)$$

Combining (35) and (36) we obtain that $A_{(y_z, d_z, z), \cdot} x_0 = \beta_{(y_z, d_z, z)}(P)$ for any $z \in \mathcal{Z}$. It follows that $Ax_0 = \beta(P)$ and, since $x_0 \geq 0$, that $P \in \tilde{\mathbf{P}}_0$ as desired. ■

A.2 Proof of Corollary 4.1

PROOF OF COROLLARY 4.1. As discussed in the paragraph before the statement of Corollary 4.1, setting $g = 1$, $\theta_0 = 1$, and $\mathcal{R}' = \mathcal{R}$ implies that

$$\Theta_0(P, \mathbf{Q}) = \begin{cases} \{1\} & \text{if } \mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset \\ \emptyset & \text{otherwise} \end{cases}.$$

Therefore,

$$\mathbf{P}_0 = \{P \in \mathbf{P} : 1 \in \Theta_0(P, \mathbf{Q})\} = \{P \in \mathbf{P} : \mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset\}. \quad (37)$$

Next, note that for this choice of g , θ_0 , and $\mathcal{R}'$, we have

$$\{P \in \mathbf{P} : Ax = \beta(P) \text{ for some } x \geq 0\} = \{P \in \mathbf{P} : \tilde{A}x = \tilde{\beta}(P) \text{ for some } x \geq 0\}, \quad (38)$$

because the equality in the last row of the equation $Ax = \beta(P)$ is always satisfied. Theorem 4.1 implies that

$$\mathbf{P}_0 \subseteq \{P \in \mathbf{P} : Ax = \beta(P) \text{ for some } x \geq 0\}. \quad (39)$$

On the other hand, fix some $P \in \mathbf{P}$ such that $\tilde{A}q = \tilde{\beta}(P)$ for some $q \geq 0$. It follows from $P\{(Y, D, Z) \in \mathcal{M}\} = 1$ and the definition of $\tilde{A}$ and $\tilde{\beta}(P)$ that there exists a $Q \in \mathbf{Q}$ such that $P = QT^{-1}$, so $\mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset$. As a result,

$$\{P \in \mathbf{P} : Ax = \beta(P) \text{ for some } x \geq 0\} \subseteq \{P \in \mathbf{P} : \mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset\}. \quad (40)$$

The claim of the corollary then follows from (37)–(40). ■

A.3 Proof of Theorem 4.2

In this Appendix, we provide the proof of Theorem 4.2 and additionally justify the validity of a test based on employing a subset of the restrictions in A , as discussed in Remark 4.1. For the latter purpose, we follow the notation of Lemma A.1 and for each $z \in \mathcal{Z}$ we select a point $(y_z, d_z) \in \mathcal{M}_z$. Recalling that the first $|\mathcal{M}|$ rows of A and $\beta(P)$ are indexed by $(y, d, z) \in \mathcal{M}$, we let A^s and $\beta^s(P)$ denote the matrix and vector obtained by removing the rows $((y_z, d_z, z) : z \in \mathcal{Z})$ from A and $\beta(P)$ respectively. An “s” superscript on other quantities, such as $\hat{\beta}_n^s$ in place of $\hat{\beta}_n$, indicates an analogous modification to the case in which the restrictions indexed by $((y_z, d_z, z) : z \in \mathcal{Z})$ are dropped.

Throughout, we let $\psi(V, P)$ be a vector of dimension $|\mathcal{M}| + 2$. Indexing the first $|\mathcal{M}|$ coordinates of $\psi(V, P)$ by $(y, d, z) \in \mathcal{M}$, we define $\psi_{ydz}(V, P)$ to be given by

$$\psi_{ydz}(V, P) \equiv \frac{1}{P_z} I\{Y = y, D = d, Z = z\} - \frac{P_{ydz}}{P_z^2} I\{Z = z\}, \quad (41)$$

and let the final two coordinates of $\psi(V, P)$ equal zero. The random variable $\psi(V, P)$ is the influence function of $\sqrt{n}\{\hat{\beta}_n - \beta(P)\}$ and will play an important role in our analysis. Similarly note $\psi^s(V, P)$ is the influence function of $\sqrt{n}\{\hat{\beta}_n^s - \beta^s(P)\}$.

The proof of Theorem 4.2 requires a final assumption that ensures that our consistent approximation for the distribution of our test statistic translates into consistency of the critical values. To formally state this assumption, we define the set

$$\mathcal{V}(P) \equiv \{s \in \mathbf{R}^{|\mathcal{M}|+2} : A^\dagger s \leq 0, \|\Omega(P)(AA')^\dagger s\|_1 \leq 1\},$$

which may be understood as the population analogue to $\hat{\mathcal{V}}_n$. We further let $\mathcal{E}(P)$ denote the extreme points of $(AA')^\dagger \mathcal{V}(P)$, set $\sigma^2(s, P) \equiv \text{Var}_P\{s' \psi(V, P)\}$ for any $s \in \mathcal{E}(P)$, and define

$$\bar{\sigma}(P) \equiv \max_{s \in \mathcal{E}(P)} \sigma(s, P), \quad \underline{\sigma}(P) \equiv \min_{s \in \mathcal{E}(P) : \sigma(s, P) > 0} \sigma(s, P),$$

where we let $\underline{\sigma}(P) = +\infty$ if $\sigma(s, P) = 0$ for all $s \in \mathcal{E}(P)$. Finally, for any $P \in \mathbf{P}$, we let $\Sigma(P) \equiv E_P[\psi(V, P)\psi(V, P)']$ and $\mathbb{G}(P) \sim N(0, \Sigma(P))$, and define

$$m(P) \equiv \text{med} \left\{ \sup_{s \in \mathcal{V}(P)} \langle A^\dagger s, A^\dagger \mathbb{G}(P) \rangle \right\},$$

where for any random variable V , $\text{med}\{V\}$ denotes its median. As discussed in the main text, in high-dimensional settings, consistency of the critical values requires a condition termed

anti-concentration by Chernozhukov et al. (2015). Given the introduced notation, we state a final assumption that ensures there is sufficient anti-concentration in our problem.

Assumption A.1. The sequence $\xi_n \equiv \log^{3/2}(n)\sqrt{|\mathcal{M}|}/\sqrt{n} \vee \lambda_n\sqrt{\log(|\mathcal{M}|)}$ satisfies

$$\sup_{P \in \mathbf{P}} \frac{m(P) + \bar{\sigma}(P)}{\underline{\sigma}^2(P)} = o(\xi_n^{-1}).$$

Similarly, establishing the validity of a test based on A^s in place of A requires us to impose a condition in which the quantities $\bar{\sigma}^s(P)$, $\underline{\sigma}^s(P)$, and $m^s(P)$ are understood as parallel modifications to the parameters $\bar{\sigma}(P)$, $\underline{\sigma}(P)$ and $m(P)$.

Assumption A.2. The sequence $\xi_n \equiv \log^{3/2}(n)\sqrt{|\mathcal{M}|}/\sqrt{n} \vee \lambda_n\sqrt{\log(|\mathcal{M}|)}$ satisfies

$$\sup_{P \in \mathbf{P}} \frac{m^s(P) + \bar{\sigma}^s(P)}{(\underline{\sigma}^s(P))^2} = o(\xi_n^{-1}).$$

By dropping redundant rows, the asymptotic distribution of the test based on A^s (in place of A) depends on the asymptotic distribution of $\sqrt{n}(\hat{\beta}_n^s - \beta^s(P))$ (in place of $\sqrt{n}(\hat{\beta}_n - \beta(P))$). Unlike the asymptotic variance of $\sqrt{n}(\hat{\beta}_n - \beta(P))$, the asymptotic variance of $\sqrt{n}(\hat{\beta}_n^s - \beta^s(P))$ is invertible once we exclude the final two coordinates. This property is useful for obtaining lower level, but potentially conservative, sufficient conditions for the anti-concentration condition in Assumption A.2. Our next lemma formalizes this observation.

Lemma A.2. *If Assumptions 4.1(i) and 4.1(iii) hold, then it follows that Assumption A.2 is satisfied provided $\log^2(n)|\mathcal{M}|^{5/2} = o(|\mathcal{Z}|\sqrt{n})$ and $\lambda_n \log(|\mathcal{M}|)|\mathcal{M}|^2 = o(|\mathcal{Z}|)$.*

The conclusion of Lemma A.2 implies that anti-concentration holds for the test based on A^s provided that $|\mathcal{M}|^{5/2} = o(|\mathcal{Z}|\sqrt{n})$ (up to logs) and λ_n tends to zero sufficiently fast. An important implication of Lemma A.2 is that Assumption A.2 is automatically satisfied for the test based on A^s in low dimensional approximations in which $|\mathcal{M}|$ is fixed provided λ_n tends to zero. Moreover, we note that the universal (in A^s) bounds used by Lemma A.2 may be quite conservative for specific A^s , leading to the condition $|\mathcal{M}|^{5/2} = o(|\mathcal{Z}|\sqrt{n})$ (up to logs) being overly stringent. Alternatively, Assumptions A.1 and A.2 can be dispensed with by modifying our critical value to include a “correction factor” introduced by Andrews and Shi (2013). We opt to introduce Assumptions A.1 and A.2 instead, however, to avoid introducing an additional tuning parameter.

For completeness, we state the validity of a test based on A^s , which we denote by $\phi_n^s(\theta_0)$.

Theorem A.1. *Suppose Assumptions 4.1 and A.2 hold, $\alpha \in (0, 0.5)$, $\log^3(n)|\mathcal{M}|/n = o(1)$,*

and $0 \leq \lambda_n \leq 1$ satisfies $\lambda_n \sqrt{\log(|\mathcal{M}|)} = o(1)$. Then it follows that

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathbf{P}_0} E_P[\phi_n^s(\theta_0)] \leq \alpha.$$

We next provide the proofs of Theorems 4.2, A.1 and Lemma A.2. The proofs rely on numerous auxiliary lemmas that are stated and proven in Online Appendix Section B.

PROOF OF THEOREM 4.2. The proof follows the same arguments as those in the proof of Theorem 4.2 in Fang et al. (2023) but with their coupling results (Lemmas A.4 and A.5) replaced by our improved coupling rates from Lemma B.2. Specifically, we note that Lemma B.1 implies Assumptions 4.1, 4.2, 4.3, 4.4(v) and A.1 of Fang et al. (2023) hold. Assumptions 4.4(i)–(iv) are not verified directly, because their only role in the proof of Theorem 4.2 in Fang et al. (2023) is to ensure a coupling of the bootstrap statistic — a condition that we instead establish directly in Lemma B.2. The arguments in Fang et al. (2023) can therefore be applied with $r_n = \log^{3/2}(n) \sqrt{|\mathcal{M}|}/\sqrt{n}$ (the coupling rate of Lemma B.2(i)) and $b_n = r_n$ (the coupling rate of Lemma B.2(ii)). ■

PROOF OF THEOREM A.1. The proof follows from identical arguments to those in the proof of Theorem 4.2 after noting that the assumptions in Fang et al. (2023) holding for $\phi_n(\theta_0)$ imply they hold for $\phi_n^s(\theta_0)$ as well. In particular, note that the coupling results in Lemma B.2 holding for $\sqrt{n}(\hat{\beta}_n - \beta(P))$ and $\sqrt{n}(\hat{\beta}_n^* - \hat{\beta}_n)$ imply that they continue to hold when applied to the subvectors $\sqrt{n}(\hat{\beta}_n^s - \beta^s(P))$ and $\sqrt{n}(\hat{\beta}_n^{s*} - \hat{\beta}_n^s)$. Similarly, Assumption 4.1 in Fang et al. (2023) is satisfied with $a_n = \log(|\mathcal{M}|) \sqrt{|\mathcal{M}|}/n$ by Assumption 4.1(i), Assumption 4.1(ii) implying $(\hat{\beta}_n^s - \beta^s(P)) \in \text{range}(A^s)$ with probability tending to one uniformly in $P \in \mathbf{P}$, and Lemma B.3. Furthermore, Assumptions 4.2(i)(iii)(iv), 4.3, 4.4(v), and A.1 in Fang et al. (2023) holding for $\phi_n(\theta_0)$, as shown in Lemma B.1, immediately imply they hold for $\phi_n^s(\theta_0)$ as well. Finally, Assumption 4.2(ii) in Fang et al. (2023) holds by Lemma B.9(i). Hence, all the assumptions of Fang et al. (2023) verified to hold for $\phi_n(\theta_0)$ in the proof of Theorem 4.2 continue to hold for $\phi_n^s(\theta_0)$ as well, and the theorem can be shown using the same arguments in the proof of Theorem 4.2. ■

PROOF OF LEMMA A.2. First select an arbitrary $s \in \mathcal{E}^s(P)$, which we note by definition must satisfy $s = (A^s(A^s)')^\dagger \tilde{s}$ for some $\tilde{s} \in \mathcal{V}^s(P)$. Note that if $\Omega^s(P)(A^s(A^s)')^\dagger \tilde{s} = 0$, we have

$$(\sigma^s(s, P))^2 = \text{Var}_P\{\langle (A^s(A^s)')^\dagger \tilde{s}, \mathbb{G}^s(P) \rangle\} = \text{Var}_P\{\langle \Omega^s(P)(A^s(A^s)')^\dagger \tilde{s}, \Omega^s(P)^\dagger \mathbb{G}^s(P) \rangle\} = 0 \quad (42)$$

where the second equality follows from $\Omega^s(P)\Omega^s(P)^\dagger \mathbb{G}^s(P) = \mathbb{G}^s(P)$ and the third equality

from $\Omega^s(P)(A^s(A^s)')^\dagger \tilde{s} = 0$. On the other hand, if $\Omega^s(P)(A^s(A^s)')^\dagger \tilde{s} \neq 0$, then

$$\|\Omega^s(P)(A^s(A^s)')^\dagger \tilde{s}\|_1 = 1, \quad (43)$$

because otherwise it would be possible to write s as a convex combination of zero and $(A^s(A^s)')^\dagger \tilde{s} / \|\Omega^s(P)(A^s(A^s)')^\dagger \tilde{s}\|_1$, which contradicts s being an extreme point of the set $(A^s(A^s)')^\dagger \mathcal{V}^s(P)$. Therefore, again using that $\Omega^s(P)\Omega^s(P)^\dagger \mathbb{G}^s(P) = \mathbb{G}^s(P)$ we obtain

$$\begin{aligned} (\underline{\sigma}^s(P))^2 &\geq \inf_{\tilde{s} \in \mathcal{V}^s(P): \|\Omega^s(P)(A^s(A^s)')^\dagger \tilde{s}\|_1=1} \text{Var}_P\{\langle (A^s(A^s)')^\dagger \tilde{s}, \mathbb{G}^s(P) \rangle\} \\ &= \inf_{\tilde{s} \in \mathcal{V}^s(P): \|\Omega^s(P)(A^s(A^s)')^\dagger \tilde{s}\|_1=1} \text{Var}_P\{\langle \Omega^s(P)(A^s(A^s)')^\dagger \tilde{s}, \Omega^s(P)^\dagger \mathbb{G}^s(P) \rangle\}. \end{aligned} \quad (44)$$

Next note that the range of $\Omega^s(P)^\dagger \Sigma^s(P) \Omega^s(P)^\dagger$ includes all vectors whose last two coordinates equal zero by Lemma B.9(iii). Hence, it follows that

$$\Omega^s(P)(A^s(A^s)')^\dagger \tilde{s} \in \text{range}(\Omega^s(P)^\dagger \Sigma^s(P) \Omega^s(P)^\dagger) \quad (45)$$

for all $\tilde{s}$ and $P \in \mathbf{P}$ because the final two diagonal entries of $\Omega^s(P)$ equal zero. Therefore, Theorem 6.6.3(ii) in Luenberger (1997) and (45) imply $\Omega^s(P)(A^s(A^s)')^\dagger \tilde{s}$ is in the span of the eigenvectors corresponding to the nonzero eigenvalues of $\Omega^s(P)^\dagger \Sigma^s(P) \Omega^s(P)^\dagger$. Hence, result (44), the inequality $\|a\|_1 \leq \sqrt{d}\|a\|_2$ for any $a \in \mathbf{R}^d$, and Lemma B.9(ii) yield

$$(\underline{\sigma}^s(P))^2 \geq \pi \times \frac{|\mathcal{Z}|}{|\mathcal{M}|^2} \quad (46)$$

for some $\pi > 0$. Moreover, the definition of $\mathcal{V}^s(P)$, $\Omega^s(P)\Omega^s(P)^\dagger \mathbb{G}^s(P) = \mathbb{G}^s(P)$, Lemma B.9(i) and $\|a\|_2 \leq \|a\|_1$ for any $a \in \mathbf{R}^d$ imply uniformly in $P \in \mathbf{P}$ that

$$\begin{aligned} (\bar{\sigma}^s(P))^2 &\leq \sup_{\tilde{s} \in \mathcal{V}^s(P)} \text{Var}_P\{\langle (A^s(A^s)')^\dagger \tilde{s}, \mathbb{G}^s(P) \rangle\} \\ &\leq \sup_{\|\Omega^s(P)(A^s(A^s)')^\dagger \tilde{s}\|_1 \leq 1} \text{Var}_P\{\langle \Omega^s(P)(A^s(A^s)')^\dagger \tilde{s}, \Omega^s(P)^\dagger \mathbb{G}^s(P) \rangle\} = O(1). \end{aligned} \quad (47)$$

Finally, note similar manipulations, $((A^s)^\dagger)'(A^s)^\dagger = (A^s(A^s)')^\dagger$, and Hölder's inequality yield

$$\mathbf{m}^s(P) \leq \text{med} \left\{ \|\Omega^s(P)^\dagger \mathbb{G}^s(P)\|_\infty \right\} = O(\sqrt{\log(|\mathcal{M}|)}), \quad (48)$$

where the equality holds uniformly in $P \in \mathbf{P}$ by a standard maximal inequality (see, e.g., Lemma 2.2.2 in van der Vaart and Wellner (1996)). The claim of the lemma therefore follows from combining results (46), (47), and (48) with the definition of ξ_n . ■