On the Identifying Power of Monotonicity for Average Treatment Effects*

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

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

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

May 22, 2024

Abstract

In the context of a binary outcome, treatment, and instrument, Balke and Pearl (1993, 1997) establish that adding monotonicity to the instrument exogeneity assumption does not decrease the identified sets for average potential outcomes and average treatment effect parameters when those assumptions are consistent with the distribution of the observable data. We show that the same results hold in the broader context of multi-valued outcome, treatment, and instrument. An important example of such a setting is a multi-arm randomized controlled trial with noncompliance.

KEYWORDS: Multi-valued Treatments, Average Treatment Effects, Endogeneity, Instrumental Variables.

JEL classification codes: C12, C31, C35, C36

*Vytlacil acknowledges financial support from the Tobin Center for Economic Policy at Yale.

1 Introduction

In their analysis of a setting with a binary treatment and instrument, Imbens and Angrist (1994) and Angrist et al. (1996) introduce a restriction on individual-level responses to the instrument that they refer to as monotonicity. They show that imposing this condition in addition to instrument exogeneity allows identification of the Local Average Treatment Effect (LATE) (also called Complier Average Causal Effect) parameter, while that parameter would not be identified under instrument exogeneity alone.

In contrast, Balke and Pearl (1993, 1997) consider the identifying power of monotonicity in addition to instrument exogeneity for the average treatment effect (ATE). In the context of a binary outcome, treatment, and instrument, they establish the suprising result that whenever the distribution of the observable data is consistent with instrument exogeneity and monotonicity, imposing monotonicity does not reduce the identified set for the ATE compared to imposing instrument exogeneity alone. In other words, in the context of a binary outcome, treatment, and instrument, imposing monotonicity does not provide any additional identifying power for the ATE versus imposing instrument exogeneity alone whenever the distribution of the observable data is consistent with those assumptions.

In this paper, we show that this phenomenon holds much more generally for a suitable generalization of their monotonicity condition, which, for simplicity, we hereafter continue to refer to as monotonicity. In particular, we generalize their analysis to discrete but possibly non-binary outcomes, treatments, and instruments. In this way, our framework nests the important example of a multi-arm randomized controlled trial (RCT) with noncompliance. We first derive the identified sets for average potential outcomes under instrument exogeneity and monotonicity when these assumptions are consistent with the distribution of the observable data. Our expressions parallel those found by Balke and Pearl (1993, 1997) for a binary outcome, treatment and instrument and permit us to generalize their result that adding monotonicity to the instrument exogeneity does not reduce the identified sets for average potential outcomes and ATEs whenever the assumptions are consistent with the distribution of the observable data.

In the terminology of Frangakis and Rubin (2002), monotonicty is a restriction on principal strata, and the LATE is the average effect for a particular principal stratum. This paper is related to the literature that extends the analysis of Imbens and Angrist (1994) and Angrist et al. (1996) by considering the identifying power of restrictions on principal strata to identify average treatment effects within principal strata in the context of multi-valued treatments and multi-valued instruments as arises for multi-arm RCTs with noncompliance, see, e.g., Cheng and Small (2006), and Blackwell and Pashley (2023). In contrast to those papers, we consider average potential outcomes and ATEs as our parameters of interest as opposed to the average effect within a principal stratum.

This paper is also related to papers that, in the context of a binary outcome, treatment, and instrument, consider the identifying power of monotonicity for the distribution (as opposed to the average) of potential outcomes, or consider the identifying power of monotonicity when combined with additional restrictions. In particular, Kamat (2019) shows imposing monotonicity in addition to instrument exgoeneity does have additional identifying power for the distribution of potential outcomes versus imposing instrument exogeneity alone. Machado et al. (2019) show that monotonicity does have additional identifying power for the ATE

beyond instrument exogeneity if one additionally imposes a monotonicity assumption on how the outcome relates to the treatment. Thus, while monotonicity has no identifying power beyond instrument exogeneity for the ATEs when imposing no other assumptions, this phenomenon is sensitive to both the choice of parameter and to whether one imposes additional assumptions. In addition, the phenomenon is sensitive to the particular restriction on individual-level responses to the instrument. We show in Remark 3.6 an example of an alternative such restriction that results in identification of average potential outcomes when combined with instrument exogeneity.

In the context of a binary treatment, Vytlacil (2002) establishes that imposing a nonparametric selection model is equivalent to imposing instrument exogeneity and monotonicity. The analysis of this paper is thus related to Heckman and Vytlacil (2001), who establish the identified set for the ATE if imposing a nonparametric selection model without restricting outcomes or instruments to be binary, while, unlike this paper, still restricting the treatment to be binary.

The rest of the paper is organized as follows. Section 2 introduces our formal setup, notation and assumptions. Our main results are presented in Section 3. Proofs of all results can be found in the Appendix.

2 Setup and Notation

Let $Y \in \mathcal{Y}$ be a discrete outcome of interest, $D \in \mathcal{D} = \{0, \ldots, |\mathcal{D}|-1\}$ with $|\mathcal{D}| \geq 2$ be a discrete endogenous regressor (treatment variable), and $Z \in \mathcal{Z} = \{0, \ldots, |\mathcal{Z}|-1\}$ with $|\mathcal{Z}| \geq 2$ be a discrete instrument. Further define Y_d as the potential outcome if $D = d \in \mathcal{D}$ and D_z as the potential treatment choice if $Z = z \in \mathcal{Z}$. We have the following relationship among variables:

$$Y = \sum_{d \in \mathcal{D}} Y_d 1\{D = d\}, \quad D = \sum_{z \in \mathcal{Z}} D_z 1\{Z = z\} \ . \tag{1}$$

In the context of a multi-arm RCT with noncompliance, $|\mathcal{D}| = |\mathcal{Z}|$, $Z = z$ denotes random assignment to treatment arm z, and $D_z = z$ denotes that the subject would comply with treatment assignment if assigned to treatment arm z.

Let P denote the distribution of (Y, D, Z) and Q denote the distribution of $(\{Y_d\}_{d \in \mathcal{D}}, \{D_z\}_{z \in \mathcal{Z}}, Z)$. For T defined by (1), we have

$$(Y, D, Z) = T(\{Y_d\}_{d \in \mathcal{D}}, \{D_z\}_{z \in \mathcal{Z}}, Z) \ ,$$

and therefore $P = QT^{-1}$.

Because D and Z are both discrete, we can characterize any P and Q with a finite number of probabilities. Formally, P is defined by a vector of $|\mathcal{Y}| \times |\mathcal{D}| \times |\mathcal{Z}|$ probabilities:

$$(p_{ydz} : y \in \mathcal{Y}, d \in \mathcal{D}, z \in \mathcal{Z}) \ ,$$

where $p_{ydz} = P\{Y = y, D = d, Z = z\}$. Q is defined by a vector of $|\mathcal{Y}|^{|\mathcal{D}|} \times |\mathcal{D}|^{|\mathcal{Z}|} \times |\mathcal{Z}|$ probabilities:

$$(q(y_0, \dots y_{|\mathcal{D}|-1}, d_0, \dots, d_{|\mathcal{Z}|-1}, z) : y_d \in \mathcal{Y} \text{ for } d \in \mathcal{D}, d_{z'} \in \mathcal{D} \text{ for } z' \in \mathcal{Z}, z \in \mathcal{Z}) \ ,$$

where

$$q(y_0, \dots, y_{|\mathcal{D}|-1}, d_0, \dots, d_{|\mathcal{Z}|-1}, z) = Q\{Y_d = y_d \text{ for } d \in \mathcal{D}, D_{z'} = d_{z'} \text{ for } z' \in \mathcal{Z}, z \in \mathcal{Z}\} \ .$$

We restrict $Q \in \mathbf{Q}$. We will maintain the following exogeneity assumption on $\mathbf{Q}$ throughout:

Assumption 2.1 (Instrument Exogeneity)**.** For all $Q \in \mathbf{Q}$, $(\{Y_d\}_{d\in\mathcal{D}}, \{D_z\}_{z\in\mathcal{Z}}) \perp\!\!\!\perp Z$ under Q.

Under instrument exogeneity, it is convenient to represent P and Q by the conditional distributions given Z instead. This permits us to follow Balke and Pearl (1993, 1997) in writing restrictions on the model as linear equalities. To see this, first note the marginal distribution of instruments under P and Q are the same, that is, for all $z \in \mathcal{Z}$,

$$P\{Z = z\} = Q\{Z = z\} \ .$$

If Q satisfies Assumption 2.1,

$$P\{Y = y, D = d|Z = z\} = Q\{Y_d = y, D_z = d|Z = z\} = Q\{Y_d = y, D_z = d\} \ .$$

Therefore, we will disregard the marginal distributions of Z in P and Q and represent P by conditional probabilities and Q by marginal distributions of $(Y_0, \dots, Y_{|\mathcal{D}|-1}, D_0, \dots, D_{|\mathcal{Z}|-1})$. Formally, P is defined by a vector of $|\mathcal{Y}| \times |\mathcal{D}| \times |\mathcal{Z}|$ conditional probabilities:

$$(p_{yd|z} : y \in \mathcal{Y}, d \in \mathcal{D}, z \in \mathcal{Z}) \ ,$$

where $p_{yd|z} = P\{Y = y, D = d|Z = z\}$. Q is defined by a vector of $|\mathcal{Y}|^{|\mathcal{D}|} \times |\mathcal{D}|^{|\mathcal{Z}|}$ probabilities:

$$(q(y_0, \dots y_{|\mathcal{D}|-1}, d_0, \dots, d_{|\mathcal{Z}|-1}) : y_d \in \mathcal{Y} \text{ for } d \in \mathcal{D}, d_z \in \mathcal{D} \text{ for } z \in \mathcal{Z}) \ ,$$

where

$$q(y_0, \dots, y_{|\mathcal{D}|-1}, d_0, \dots, d_{|\mathcal{Z}|-1}) = Q\{Y_d = y_d \text{ for } d \in \mathcal{D}, D_z = d_z \text{ for } z \in \mathcal{Z}\} \ .$$

It follows from Assumption 2.1 that

$$\begin{aligned} p_{yd|z} &= P\{Y = y, D = d|Z = z\} \\ &= Q\{Y_d = y, D_z = d\} = \sum_{y_d = y, d_z = d} q(y_0, \dots y_{|\mathcal{D}|-1}, d_0, \dots, d_{|\mathcal{Z}|-1}) \ , \end{aligned} \tag{2}$$

and therefore Q is consistent with P if P and Q are related by the finite number of linear equalities defined in equation (2).

Following Heckman and Pinto (2018), we define a treatment response type as a vector $r^t \in \mathcal{D}^{|\mathcal{Z}|}$,

$$r^t = (d_0, \ldots, d_{|\mathcal{Z}|-1}) \in \mathcal{D}^{|\mathcal{Z}|} \ .$$

Treatment response types are also called principal strata (Frangakis and Rubin, 2002). We analogously define an outcome response type as a vector $r^o \in \mathcal{Y}^{|\mathcal{D}|}$,

$$r^o = (y_0, \ldots, y_{|\mathcal{D}|-1}) \in \mathcal{Y}^{|\mathcal{D}|} \ .$$

Let r^o_j denote the $(j+1)$th entry of r^o and r^t_j denote the $(j+1)$th entry of r^t.

Given P and a choice of $\mathbf{Q}$, we define the set of $Q \in \mathbf{Q}$ that can rationalize P as

$$\Delta(P|\mathbf{Q}) = \left\{ Q \in \mathbf{Q} : \ p_{yd|z} = \sum_{(r^o, r^t) \in \mathcal{Y}^{|\mathcal{D}|} \times \mathcal{D}^{|\mathcal{Z}|} : r^o_d = y, r^t_z = d} q(r^o, r^t), \ \forall (y, d, z) \in \mathcal{Y} \times \mathcal{D} \times \mathcal{Z} \right\} , \tag{3}$$

where the equalities linking $p_{yd|z}$ and $q(r^o, r^t)$ come from (2). P is consistent with the restrictions on $\mathbf{Q}$ if $\Delta(P|\mathbf{Q}) \neq \emptyset$.

Let $\theta_Q = (\mathbb{E}_Q[Y_j] : j \in \mathcal{D})$ denote the vector of average potential outcomes. For fixed P and $\mathbf{Q}$, the identified set for θ_Q is given by

$$I(P, \mathbf{Q}) \equiv \{(\mathbb{E}_Q[Y_j] : j \in \mathcal{D}) : Q \in \Delta(P|\mathbf{Q})\} \ . \tag{4}$$

$I(P, \mathbf{Q})$ is nonempty when $\Delta(P|\mathbf{Q})$ is nonempty. By construction, this set is "sharp" in the sense that for any value in the set there exists $Q \in \Delta(P|\mathbf{Q})$ for which $(\mathbb{E}_Q[Y_j] : j \in \mathcal{D})$ equals the prescribed value. The identified set for θ_Q immediately implies identified sets for functions of θ_Q, such as the average treatment effect for treatment j versus k, $\mathbb{E}_Q[Y_j] - \mathbb{E}_Q[Y_k]$.

In addition to the maintained instrument exogeneity assumption on $\mathbf{Q}$, we also consider the following generalization of the monotonicity assumption:

Assumption 2.2. For all $Q \in \mathbf{Q}$,

$$Q\{D_j \neq j, \ D_z = j \text{ for some } z \neq j\} = 0$$

for any $j \in \mathcal{Z}$.

In the context of a RCT with non-compliance, Assumption 2.2 imposes that there is zero probability that an individual would not comply with treatment assignment if assigned to treatment j but would take treatment j if assigned to some other treatment arm. More generally, interpreting $Z = z$ as encouragement to take treatment z, Assumption 2.2 imposes that there is zero probability that an individual would not take treatment j when encouraged to take j but would take treatment j for any other value of the instrument. Adapting the terminology of Angrist et al. (1996), Assumption 2.2 imposes that there are no defiers in the sense that those who would not take treatment j even when encouraged towards treatment j are never-takers

for treatment j.

Assumption 2.2 is equivalent to restricting $q(r^o, r^t) = 0$ for any treatment response type r^t such that $r^t_j \neq j$ while $r^t_z = j$ for some $j, z \in \mathcal{Z}$ with $j \neq z$. Let R^t denote the complement of that set of treatment response types, the set of treatment response types that may occur with positive probability under Assumption 2.2:

$$R^t = \{r^t \in \mathcal{D}^{|\mathcal{Z}|} \ : \ r^t_j \neq j \text{ implies } r^t_z \neq j \ \forall \ j, z \in \mathcal{Z}\}. \tag{5}$$

Under Assumption 2.2, $Q\{(\{D_z\}_{z\in\mathcal{Z}}) \in R^t\} = 1$.

Remark 2.1. Assumption 2.2 is equivalent to the monotonicity assumption of Imbens and Angrist (1994) when $\mathcal{D} = \mathcal{Z} = \{0, 1\}$. In this sense, Assumption 2.2 is a generalization of monotonicity to multiple treatment arms. ■

3 Main Results

We first present in Section 3.1 the identified set for θ_Q under instrument exogeneity Assumption 2.1 and the generalized monotonicity Assumption 2.2 when those assumptions are consistent with P. We then show in Section 3.2 that, when those assumptions are consistent with P, that the identified set for θ_Q using those assumptions coincide with the identified set for θ_Q imposing only Assumption 2.1 without imposing Assumption 2.2.

3.1 Identified Sets

The following theorem establishes the form of the identified set for θ_Q, $I(P, \mathbf{Q})$, when $\Delta(P|\mathbf{Q})$ is nonempty with $\mathbf{Q}$ restricted to those Q satisfying Assumptions 2.1 and 2.2. The theorem is stated for the case where $\mathcal{Y} = \{0, 1\}$ and $|\mathcal{Z}| = |\mathcal{D}| = J$, though, as we discuss below, the results generalize at the cost of additional notation.

Theorem 3.1. *Suppose $\mathcal{Y} = \{0, 1\}$, and $|\mathcal{Z}| = |\mathcal{D}| = J$. Restrict $\mathbf{Q}$ to those Q that satisfy Assumptions 2.1 and 2.2. Suppose that $\Delta(P|\mathbf{Q}) \neq \emptyset$. Then, the identified set for θ_Q is given by:*

$$I(P, \mathbf{Q}) = \prod_{j=0}^{J-1} [p_{1j|j}, \ 1 - p_{0j|j}]. \tag{6}$$

Theorem 3.1 immediately implies the following result on the identified sets for the ATE of treatment j versus k:

Corollary 3.1. *Under the assumptions of Theorem 3.1, the identified set for $\mathbb{E}_Q[Y_j] - \mathbb{E}_Q[Y_k]$ is:*

$$[p_{0k|k} + p_{1j|j} - 1, 1 - p_{0j|j} - p_{1k|k}]. \tag{7}$$

Remark 3.1. Theorem 3.1 constructs the identified set for θ_Q under the stated assumptions when $I(P, \mathbf{Q})$ is nonempty, i.e., when the identified set for $\theta_Q$is nonempty. The set $I(P, \mathbf{Q})$ is nonempty when, with $\mathbf{Q}$ restricted to Q that satisfy Assumptions 2.1 and 2.2, there exists a $Q \in \mathbf{Q}$ consistent with P, i.e., a $Q \in \mathbf{Q}$ such that q solves equation (2). The set $I(P, \mathbf{Q})$ is empty when there does not exist such a Q, in which case the assumptions are inconsistent with P. In what follows we sketch an argument that allows us to explicitly characterize these conditions. Restrict $\mathbf{Q}$ to those Q that satisfy Assumptions 2.1 and 2.2. Then $\mathbf{Q}$ is given by

$$\mathbf{Q} = \{q \geq 0 : \iota' q = 1, q(r^o, r^t) = 0 \text{ for } r^t \notin R^t\} ,$$

where R^t is defined by equation (5) and ι is a $(2^J \times J^J) \times 1$ vector of ones. Therefore $\mathbf{Q}$ is a bounded polyhedron. Further note (2) defines a linear map and P is consistent with the assumptions if and only if it is the image of some $Q \in \mathbf{Q}$ under this linear map. The image of a polyhedron under a linear map is still a polyhedron, and any polyhedron can be defined by a finite number of inequalities. Therefore, P is consistent with the assumptions if and only if it satisfies a set of moment inequalities. The number of these inequalities grows rapidly with J, but it is possible to test whether they are satisfied using methods described, e.g., in Fang et al. (2023). ■

Remark 3.2. The proof of Theorem 3.1 makes it clear that the result holds under a weaker version of Assumption 2.2 that does not require $|\mathcal{Z}| = |\mathcal{D}|$. In particular, the result holds while allowing $|\mathcal{Z}| \neq |\mathcal{D}|$ while weakening Assumption 2.2 to impose only that, for any $j \in \mathcal{D}$, there exists a $z(j) \in \mathcal{Z}$ such that

$$Q\{D_{z(j)} \neq j,\ D_z = j \text{ for some } z \neq z(j)\} = 0,$$

for all $Q \in \mathbf{Q}$. In other words, the same conclusion holds while allowing possibly a different number of values of the instrument than values of the treatment if each treatment has an associated instrument value that represents the maximum encouragement to take that value of the treatment. ■

Remark 3.3. Theorem 3.1 supposes that $\mathcal{Y} = \{0, 1\}$. However, the proof of the theorem makes it clear that the result immediately extends to the case of $\mathcal{Y}$ being any finite set, with the resulting identified set for θ_Q is of the form

$$\prod_{j=0}^{J-1} \left[\theta_{j|j} + y_L(1 - \sum_y p_{yj|j})),\ \theta_{j|j} + y_U(1 - \sum_y p_{yj|j}))\right], \tag{8}$$

where $\theta_{j|j} \equiv \mathbb{E}_P[Y D_j \mid Z = j]$, $y_L = \min(\mathcal{Y}), y_U = \max(\mathcal{Y})$. The resulting implied identified set for average treatment effect of j versus k is of the form

$$\mathbb{E}_Q[Y_j] - \mathbb{E}_Q[Y_k] \in \Big[(\theta_{j|j} - \theta_{k|k}) + (y_L - y_U) + y_U \sum_y p_{yk|k} - y_L \sum_y p_{yj|j},$$
$$(\theta_{j|j} - \theta_{k|k}) + (y_U - y_L) + y_L \sum_y p_{yk|k} - y_U \sum_y p_{yj|j}\Big]. \tag{9}$$

At the cost of additional notation, the results can be generalized to allow for $\mathcal{Y}$ to take a possibly infinite number of values subject to y_L, y_U being finite. ■

Remark 3.4. Under instrument exogeneity and the monotonicity assumption of Imbens and Angrist (1994),

Balke and Pearl (1993, 1997) found the same form of the identified set for θ_Q as (6) for the case of $\mathcal{Y} = \mathcal{D} = \mathcal{Z} = \{0,1\}$. As discussed in Remark 2.1, Assumption 2.2 reduces to monotonicity in the case of $\mathcal{D} = \mathcal{Z} = \{0,1\}$. Theorem 3.1 therefore generalizes the result of Balke and Pearl (1993, 1997) to allow for more than two treatment arms and instrument values. Following the previous remark, it also generalizes Balke and Pearl (1993, 1997) to outcomes taking more than two values. ■

3.2 Identifying Power of Monotonicity for the ATE

Theorem 3.1 establishes that the identified set for θ_Q under Assumptions 2.1 and 2.2 when P is consistent with those assumptions for Z and D multi-valued is the same form as the bounds shown by Balke and Pearl (1993, 1997) for binary Z and D under the same assumptions. In the case of binary Z and D, Balke and Pearl (1993, 1997) show that, when P is consistent with both Assumptions 2.1 and 2.2, the identified set for θ_Q under Assumptions 2.1 and 2.2 corresponds to the identified set for θ_Q that only imposes Assumption 2.1. We now generalize that result to multi-valued Z and D.

First, we establish that the identified set for θ_Q from equation (6) coincides with the bounds under the weaker mean independence assumption considered by Robins (1989) and Manski (1990),

$$\mathbb{E}[Y_d \mid Z = z] = \mathbb{E}[Y_d] \quad \forall\, d \in \mathcal{D},\ z \in \mathcal{Z}, \tag{10}$$

which, specialized to the case of $\mathcal{Y} = \{0,1\}$, is the marginal independence restriction that

$$Q\{Y_d = 1 \mid Z = z\} = Q\{Y_d = 1\} \quad \forall\, d \in \mathcal{D},\ z \in \mathcal{Z}. \tag{11}$$

Note that this assumption is weaker than instrument exogeneity in Assumption 2.1, and does not imply equation (2). Following their analysis for the case of $\mathcal{Y} = \{0,1\}$, the identified set for θ_Q under the assumption of equation (11) is

$$\prod_{j=0}^{J-1} \left[\max_{z \in \mathcal{Z}} p_{1j|z},\ 1 - \max_{z \in \mathcal{Z}} p_{0j|z} \right]. \tag{12}$$

The following lemma establishes the equivalence between the identified sets in equations (6) and (12) when P is consistent with Assumptions 2.1 and 2.2.

Lemma 3.1. *Suppose $\mathcal{Y} = \{0,1\}$, and $|\mathcal{Z}| = |\mathcal{D}| = J$. Let $\mathbf{Q}$ denote the set of distributions Q that satisfy Assumptions 2.1 and 2.2. Suppose that $\Delta(P|\mathbf{Q}) \neq \emptyset$. Then, the identified sets of equations (6) and (12) coincide.*

Using Lemma 3.1, we are able to establish our main result, which asserts that imposing Assumption 2.2 in addition to Assumption 2.1 either causes the identified set for θ_Q to become empty (if P is not consistent with Assumptions 2.1 and 2.2) or leaves the identified set for θ_Q unchanged (if P is consistent with Assumptions 2.1 and 2.2). The theorem is stated for the average potential outcomes, but immediately implies the same conclusion for the ATEs.

Theorem 3.2. *Suppose* $\mathcal{Y} = \{0, 1\}$*, and* $|\mathcal{Z}| = |\mathcal{D}| = J$*. Suppose that* $\Delta(P|\mathbf{Q}) \neq \emptyset$ *when restricting* $\mathbf{Q}$ *to those* Q *that satisfy Assumptions 2.1 and 2.2. Then the identified sets on* θ_Q *under Assumptions 2.1 and 2.2 coincide with the identified set under Assumption 2.1 alone.*

Remark 3.5. Suppose that $|\mathcal{Z}| = |\mathcal{D}| = J$ and $\mathcal{Y}$ is a finite set, but with $\mathcal{Y}$ not necessarily equal to $\{0, 1\}$. Let $\theta_{j|z} = \mathbb{E}_P[YD_j \mid Z = z]$, $y_L = \min(\mathcal{Y}), y_U = \max(\mathcal{Y})$. Then, following Robins (1989) and Manski (1990), the restriction of equation (10) results in the identified set for θ_Q of the form

$$\prod_{j=0}^{J-1} \left[\max_{z \in \mathcal{Z}} \{\theta_{j|z} + y_L(1 - \sum_y p_{yj|z})\}, \ \min_{z \in \mathcal{Z}} \{\theta_{j|z} + y_U(1 - \sum_y p_{yj|z})\} \right]. \tag{13}$$

The same proof strategy as in Lemma 3.1 establishes that the identified sets of equations (8) and (13) coincide when $\Delta(P|\mathbf{Q}) \neq \emptyset$ with $\mathbf{Q}$ denoting the set of distributions Q that satisfy Assumptions 2.1 and 2.2, and then the same proof as of Theorem 3.2 establishes that the identified sets for θ_Q under Assumptions 2.1 and 2.2 coincide with the identified set for θ_Q under Assumption 2.1 alone when $\Delta(P|\mathbf{Q}) \neq \emptyset$. ■

Remark 3.6. Imposing alternative restrictions on treatment response types other than monotonicity can reduce the identified set for θ_Q compared to imposing Assumption 2.1 alone even when these assumptions are consistent with the distribution of the observable data. For example, suppose $\mathcal{Y} = \mathcal{D} = \mathcal{Z} = \{0, 1\}$ and consider the restriction on $\mathbf{Q}$ that $Q\{D_0 = D_1\} = 0$ for all $Q \in \mathbf{Q}$. In other words, in the language of Angrist et al. (1996), all individuals are either compliers or defiers. Suppose that $\Delta(P \mid \mathbf{Q})$ is nonempty. As we show in the appendix,θ_Q is identified, while it need not be identified if not imposing that restriction on treatment response types. ■

A Proofs of Main Results

A.1 Auxillary Results

Following the discussion in Section 2, we have that under exogeneity Assumption 2.1, Q is consistent with P if

$$p_{yd|z} = \sum_{(r^o,r^t):r^o_d=y,r^t_z=d} q(r^o,r^t) \qquad \forall\ y \in \mathcal{Y},\ d \in \mathcal{D},\ z \in \mathcal{Z}, \tag{14}$$

where

$$r^o = (y_0, \dots y_{|\mathcal{D}|-1}),$$
$$r^t = (d_0, \dots, d_{|\mathcal{Z}|-1}).$$

Below we derive a lemma that simplifies determining whether $q(r^o, r^t)$ satisfies equation (14) and will be used subsequently to derive our characterization of the identified set. To this end, we require some further notation.

Let

$$\mathcal{N}(r^t) = \{d \in \mathcal{D} \ :\ r^t_z \neq d \text{ for all } z \in \mathcal{Z}\},$$

$$\mathcal{N}(r^t)^c = \{d \in \mathcal{D} \ :\ r^t_z = d \text{ for some } z \in \mathcal{Z}\},$$

For a given treatment response type r^t, $\mathcal{N}(r^t)$ is the set of treatments for which that treatment response type is a never-taker, and $\mathcal{N}(r^t)^c$ is the set of treatments for which that treatment response type will take that treatment for some value of z. Using this notation, partition outcome and treatment response types (r^o, r^t) as $(r^o_n(r^t), r^o_c(r^t), r^t)$ where

$$r^o_n(r^t) = (r^o_d : d \in \mathcal{N}(r^t)),$$
$$r^o_c(r^t) = (r^o_d : d \in \mathcal{N}(r^t)^c).$$

For a given treatment response type r^t, $r^o_n(r^t)$ are those outcomes that are never observed for that response type, and $r^o_c(r^t)$ are the remaining outcomes that are observed given some potential value of Z. For example, for any $r^t \in R^t$ with R^t defined by equation (5),

$$\mathcal{N}(r^t) = \{j \in \mathcal{D} \ :\ r^t_j \neq j\},$$
$$\mathcal{N}(r^t)^c = \{j \in \mathcal{D} \ :\ r^t_j = j\},$$

and

$$r^o_n(r^t) = (r^o_j : j \in \mathcal{D},\ r^t_j \neq j),$$
$$r^o_c(r^t) = (r^o_j : j \in \mathcal{D},\ r^t_j = j).$$

For notational convenience, further define $q(r_c^o(r^t), r^t)$ as $q(r_n^o(r^t), r_c^o(r^t), r^t)$ summed over $r_n^o(r^t)$:

$$q(r_c^o(r^t), r^t) = \begin{cases} q(r^o, r^t) & \text{if } \mathcal{N}(r^t) = \emptyset \text{ so that } r_c^o(r^t) = r^o \\ \sum_{r_n^o(r^t) \in \{0,1\}^{|\mathcal{N}(r^t)|}} q(r_n^o(r^t), r_c^o(r^t), r^t) & \text{if } \mathcal{N}(r^t) \neq \emptyset \text{ so that } r_c^o(r^t) \neq r^o. \end{cases}$$

Using this notation, we have the following lemma that asserts whether $q(r^o, r^t)$ satisfies equation (14) depends only on $q(r_c^o(r^t), r^t)$.

Lemma A.1. *Suppose q satisfies (14). Then, q^* satisfies (14) if, for each $r^t \in \mathcal{D}^{|\mathcal{Z}|}$,*

$$q^*(r_c^o(r^t), r^t) = q(r_c^o(r^t), r^t) \quad \forall\ r_c^o(r^t). \tag{15}$$

Proof. We can rewrite (14) as

$$\begin{aligned} p_{yd|z} &= \sum_{r^t : r_z^t = d} \sum_{r^o : r_d^o = y} q(r^o, r^t) \\ &= \sum_{r^t : r_z^t = d} \sum_{r_c^o(r^t) : r_d^o = y} \left(\sum_{r_n^o(r^t)} q(r_n^o(r^t), r_c^o(r_t), r^t) \right) \\ &= \sum_{r^t : r_z^t = d} \sum_{r_c^o(r^t) : r_d^o = y} q(r_c^o(r_t), r^t), \end{aligned}$$

where the second equality uses that $r_c^o(r^t)$ is nonempty because $r_z^t = d$ and the third equality uses that r_d^o is an element of $r_c^o(r^t)$ for r^t such that $r_z^t = d$. The result now follows. ■

A.2 Proof of Theorem 3.1

Suppose $Q \in \Delta(P|\mathbf{Q})$. For each $j \in \mathcal{D}$,

$$\begin{aligned} Q\{Y_j = 1\} &= Q\{Y_j = 1, D_j = j\} + Q\{Y_j = 1, D_j \neq j\} \\ &= p_{1j|j} + Q\{Y_j = 1, D_j \neq j\}, \end{aligned}$$

where the second equality is using instrument exogeneity. We have $Q\{Y_j = 1, D_j \neq j\} \in [0, Q\{D_j \neq j\}]$, while instrument exogeneity implies that $Q\{D_j \neq j\} = 1 - \sum_y p_{yj|j}$. We thus have that $Q\{Y_j = 1\} \in [p_{1j|j},\ 1 - p_{0j|j}]$ for each $j \in \mathcal{D}$, and thus equation (6) provides valid bounds on θ_Q under the stated assumptions.

Let q denote latent probabilities satisfying equation (14), nonnegative, and with $q(r^o, r^t) = 0$ for $r^t \notin R^t$. The existence of such a q is implied by $\Delta(P|\mathbf{Q})$ being nonempty. We now show that if $\Delta(P|\mathbf{Q})$ is nonempty, then for any value in (6) there exists $Q \in \Delta(P|\mathbf{Q})$ for which $(\mathbb{E}_Q[Y_j] : j \in \mathcal{D})$ equals the prescribed value. To do so, we show that, for each θ in the right-hand side of (6), we can construct q^* that satisfies equation (14), is nonnegative, has $q^*(r^o, r^t) = 0$ for $r^t \notin R^t$, and with $\theta_{Q^*} = \theta$. We now construct an alternative q^* for each treatment response type r^t as follows. Fix some $\alpha = (\alpha_0, \alpha_1, ..., \alpha_{J-1})' \in [0,1]^J$. Let $q_0^*(r^o, r^t) = q(r^o, r^t)$

for all r^o. Let $K(r^t) = |\mathcal{N}(r^t)|$. Note that if $K(r^t) = 0$, then $\mathcal{N}(r^t) = \emptyset$, so $r^t_j = j$ under r^t. For such an r^t, we set $q^*(r^o, r^t) = q(r^o, r^t)$ for all $r^o \in \{0,1\}^{|\mathcal{D}|}$. If $K(r^t) \geq 1$, enumerate $\mathcal{N}(r^t)$ as $\{j[1], ..., j[K(r^t)]\}$, and for $k = 1$ to $K(r^t)$, define q^*_k iteratively as follows:

$$q^*_k((r^o_{-j[k]}, r^o_{j[k]} = 0), r^t) = (1 - \alpha_k) \sum_{r^o_{j[k]} \in \{0,1\}} q^*_{k-1}((r^o_{-j[k]}, r^o_{j[k]}), r^t) \tag{16}$$

$$q^*_k((r^o_{-j[k]}, r^o_{j[k]} = 1), r^t) = \alpha_k \sum_{r^o_{j[k]} \in \{0,1\}} q^*_{k-1}((r^o_{-j[k]}, r^o_{j[k]}), r^t) \ . \tag{17}$$

for all $r^o_{-j[k]}$, where we are partitioning $r^o = (r^o_{-j[k]}, r^o_{j[k]})$. With this construction, the marginal distribution of $Y_{j[k]}$ for treatment response type r^t is only modified in step k. In particular, for each fixed k, for $\ell \leq k-1$, for $y \in \{0,1\}$,

$$\sum_{r^o_{-j[k]}} q^*_\ell((r^o_{-j[k]}, r^o_{j[k]} = y), r^t) = \sum_{r^o_{-j[k]}} q^*_0((r^o_{-j[k]}, r^o_{j[k]} = y), r^t) = \sum_{r^o_{-j[k]}} q((r^o_{-j[k]}, r^o_{j[k]} = y), r^t) \ ,$$

which together with (17) imply

$$\sum_{r^o_{-j[k]}} q^*_k((r^o_{-j[k]}, r^o_{j[k]} = 1), r^t) = \alpha_k \sum_{r^o_{j[k]} \in \{0,1\}} \sum_{r^o_{-j[k]}} q((r^o_{-j[k]}, r^o_{j[k]}), r^t) \ .$$

On the other hand, for each fixed k, because the marginal distribution of $Y_{j[k]}$ for treatment response type r^t is not further modified after step k,

$$\begin{aligned}
\sum_{r^o_{-j[k]}} q^*_{K(r^t)}((r^o_{-j[k]}, r^o_{j[k]} = 1), r^t) &= \alpha_k \sum_{r^o_{j[k]} \in \{0,1\}} \sum_{r^o_{-j[k]}} q((r^o_{-j[k]}, r^o_{j[k]}), r^t) \\
\sum_{r^o_{-j[k]}} q^*_{K(r^t)}((r^o_{-j[k]}, r^o_{j[k]} = 0), r^t) &= (1 - \alpha_k) \sum_{r^o_{j[k]} \in \{0,1\}} \sum_{r^o_{-j[k]}} q((r^o_{-j[k]}, r^o_{j[k]}), r^t) \ .
\end{aligned} \tag{18}$$

Set

$$q^*(r^o, r^t) = q^*_{K(r^t)}(r^o, r^t) \qquad \forall \ r^o.$$

With this construction, q^* is nonnegative and equals 0 for $r^t \notin R^t$. Note that, under Assumption 2.2, r^o_j is a component of $r^o_c(r^t)$ for r^t such that $r^t_j = j$, while r^o_j is a component of $r^o_n(r^t)$ for r^t such that $r^t_j \neq j$. Thus, with this construction, for each r^t, $q^*(r^o_c(r^t), r^t) = q(r^o_c(r^t), r^t) \ \ \forall \ r^o_c(r^t)$ because of (18), and thus, by Lemma A.1, q^* satisfies equation (14). That q^* satisfies equation (14) implies that, for each $j \in \mathcal{D}$,

$$Q^*\{Y_j = 1, D_j = j\} = p_{1j|j}$$

and that

$$Q^*\{Y_j = 1, D_j \neq j\} = \sum_{r^t : r^t_j \neq j} \sum_{r^o : r^o_j = 1} q^*(r^o, r^t)$$

$$
\begin{aligned}
&= \alpha_j \sum_{r^t : r^t_j \neq j} \sum_{r^o} q(r^o, r^t) \\
&= \alpha_j \; Q\{D_j \neq j\} \\
&= \alpha_j (1 - \sum_y p_{yj|j}),
\end{aligned}
$$

where the second equality is using that (18) holds for r^t such that $r^t_j \neq j$, and the last equality is using Assumption 2.1. Thus, for each $j \in \mathcal{D}$, $Q^*\{Y_j = 1\} = p_{1j|j} + \alpha_j(1 - \sum_y p_{yj|j})$ so that θ_Q lies in the set given by equation (6). Further, we can choose $\alpha \in [0,1]^J$ such that θ_Q will equal any value in the identified set of equation (6), proving that the identified set is (6). ■

A.3 Proof of Corollary 3.1

The results follows immediately from Theorem 3.1 because $Q\{Y_j = 1\} - Q\{Y_k = 1\}$ is a function of θ_Q.

A.4 Proof of Lemma 3.1

Suppose $Q \in \Delta(P|\mathbf{Q})$. Then, for any $j, k \in \mathcal{Z}$, $y \in \{0,1\}$,

$$
\begin{aligned}
p_{yj|j} - p_{yj|k} &= Q\{Y_j = y, D_j = j\} - Q\{Y_j = y, D_k = j\} \\
&= Q\{Y_j = y, D_j = j, D_k = j\} + Q\{Y_j = y, D_j = j, D_k \neq j\} \\
&\quad - Q\{Y_j = y, D_j = j, D_k = j\} - Q\{Y_j = y, D_j \neq j, D_k = j\} \\
&= Q\{Y_j = y, D_j = j, D_k \neq j\} \\
&\geq 0,
\end{aligned}
$$

where the third equality is using that Assumption 2.2 implies that $Q\{Y_j = y, D_j \neq j, D_k = j\} = 0$. Thus, if $\Delta(P|\mathbf{Q}) \neq \emptyset$, it follows that

$$
p_{yj|j} = \max_{z \in \mathcal{Z}} p_{yj|z} \quad \text{for } y = 0, 1,
$$

and the result now follows. ■

A.5 Proof of Theorem 3.2

Let $\mathcal{B}_0$ denote the identified set for θ_Q under the mean independence assumption of equation (11), which, by results of Robins (1989) and Manski (1990), is given by equation (12). Let $\mathcal{B}_1$ denote the identified set for θ_Q under Assumptions 2.1. Let $\mathcal{B}_2$ denote the identified set for θ_Q under Assumptions 2.1 and 2.2, which by Theorem 3.1 is given by equation (6).

The restriction of equation (11) is weaker than Assumption 2.1, which is weaker than imposing both Assumptions 2.1 and 2.2. Thus,

$$
\mathcal{B}_2 \subseteq \mathcal{B}_1 \subseteq \mathcal{B}_0.
$$

By Lemma 3.1, when $\Delta(P|\mathbf{Q}) \neq \emptyset$ with $\mathbf{Q}$ denoting the set of distributions Q that satisfy Assumptions 2.1 and 2.2, we have

$$\mathcal{B}_2 = \mathcal{B}_0.$$

The result now follows. ■

B Details of Remark 3.6

If Assumption 2.1 holds and $Q\{D_1 = D_0\} = 0$, then

$$\begin{aligned}
p_{y1|1} &= Q\{Y_1 = y, D_1 = 1, D_0 = 0\} \\
p_{y0|0} &= Q\{Y_0 = y, D_1 = 1, D_0 = 0\} \\
p_{y0|1} &= Q\{Y_0 = y, D_1 = 0, D_0 = 1\} \\
p_{y1|0} &= Q\{Y_1 = y, D_1 = 0, D_0 = 1\}
\end{aligned}$$

and

$$\begin{aligned}
Q\{Y_0 = 1\} &= Q\{Y_0 = 1, D_1 = 1, D_0 = 0\} + Q\{Y_0 = 1, D_1 = 0, D_0 = 1\} \\
&= p_{10|0} + p_{10|1} \\
Q\{Y_1 = 1\} &= Q\{Y_1 = 1, D_1 = 1, D_0 = 0\} + Q\{Y_1 = 1, D_1 = 0, D_0 = 1\} \\
&= p_{11|1} + p_{11|0}.
\end{aligned}$$

Therefore, if P is consistent with Assumption 2.1 and $Q\{D_1 = D_0\} = 0$, then θ_Q is identified as

$$\theta_Q = \begin{pmatrix} \mathbb{E}_Q[Y_0] \\ \mathbb{E}_Q[Y_1] \end{pmatrix} = \begin{pmatrix} p_{10|0} + p_{10|1} \\ p_{11|0} + p_{11|1} \end{pmatrix}$$

while $Q\{Y_1 = 1\} - Q\{Y_0 = 1\}$ is identified as

$$Q\{Y_1 = 1\} - Q\{Y_0 = 1\} = p_{11|0} + p_{11|1} - p_{10|0} - p_{10|1}.$$

Under Assumption 2.1 alone, without imposing the alternative restriction on treatment response types, the identified sets for $Q\{Y_0 = 1\}$ and $Q\{Y_1 = 1\}$ are given by Balke and Pearl (1997):

$$\max \begin{Bmatrix} p_{10|1} \\ p_{10|0} \\ p_{10|0} + p_{11|0} - p_{00|1} - p_{11|1} \\ p_{01|0} + p_{10|0} - p_{00|1} - p_{01|1} \end{Bmatrix} \leq Q\{Y_0 = 1\} \leq \min \begin{Bmatrix} 1 - p_{00|1} \\ 1 - p_{00|0} \\ p_{01|0} + p_{10|0} + p_{10|1} + p_{11|1} \\ p_{10|0} + p_{11|0} + p_{01|1} + p_{10|1} \end{Bmatrix} \tag{19}$$

and

$$\max \left\{ \begin{array}{c} p_{11|0} \\ p_{11|1} \\ -p_{00|0} - p_{01|0} + p_{00|1} + p_{11|1} \\ -p_{01|0} - p_{10|0} + p_{10|1} + p_{11|1} \end{array} \right\} \leq Q\{Y_1 = 1\} \leq \min \left\{ \begin{array}{c} 1 - p_{01|1} \\ 1 - p_{01|0} \\ p_{00|0} + p_{11|0} + p_{10|1} + p_{11|1} \\ p_{10|0} + p_{11|0} + p_{00|1} + p_{11|1} \end{array} \right\}. \tag{20}$$

Algebraic manipulations show that if p is such that $p_{yd|z} > 0$ for all $y \in \mathcal{Y}, d \in \mathcal{D}, z \in \mathcal{Z}$, then the lower bounds are strictly below the upper bounds in the expressions above. Following Remark 3.1, it can be verified that P is consistent with Assumption 2.1 and the restriction on treatment response types if and only if

$$p_{01|1} - p_{00|0} - p_{10|0} \leq 0 \ .$$

Under this restriction alone, the identified sets in (19) and (20) of Balke and Pearl (1997) do not collapse to a point. Thus, in this example, imposing this alternative restriction on treatment response types in addition to Assumption 2.1 does reduce the width of the identified set (even when nonempty) for mean potential outcomes and the ATE compared to the corresponding identified sets under Assumption 2.1 alone.

References

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

Balke, A. and Pearl, J. (1993). Nonparametric bounds on causal effects from partial compliance data. Technical Report R-199, UCLA.

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

Blackwell, M. and Pashley, N. E. (2023). Noncompliance and instrumental variables for 2 k factorial experiments. *Journal of the American Statistical Association*, **118** 1102–1114.

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.

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.

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

Heckman, J. J. and Vytlacil, E. J. (2001). Instrumental variables, selection models, and tight bounds on the average treatment effect. In *Econometric Evaluation of Labour Market Policies* (M. Lechner and F. Pfeiffer, eds.). Physica-Verlag HD, 1–15.

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

Kamat, V. (2019). On the identifying content of instrument monotonicity. ArXiv:1807.01661 [econ], URL http://arxiv.org/abs/1807.01661.

Machado, C., Shaikh, A. M. and Vytlacil, E. J. (2019). Instrumental variables and the sign of the average treatment effect. *Journal of Econometrics*.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review*, **80** 319–323.

Robins, J. M. (1989). The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS* 113–159.

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