Instrumental Variables and
the Sign of the Average Treatment Effect[*]

Cecilia Machado
Graduate School of Economics
Getulio Vargas Foundation
cecilia.machado@fgv.br

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

Edward J. Vytlacil
Department of Economics
New York University
vytlacil@nyu.edu

July 24, 2013

Abstract

This paper considers identification and inference about the sign of the average effect of a binary endogenous regressor (or treatment) on a binary outcome of interest when a binary instrument is available. In this setting, the average effect of the endogenous regressor on the outcome is sometimes referred to as the average treatment effect (ATE). While maintaining instrument exogeneity, we consider three different sets of assumptions: monotonicity on the outcome equation, monotonicity on the equation for the endogenous regressor, or monotonicity on both the outcome equation and the equation for the endogenous regressor. For each of these sets of conditions, we characterize when (i) the distribution of the observed data is inconsistent with the assumptions and (ii) the distribution of the observed data is consistent with the assumptions and the sign of ATE is identified. A distinguishing feature of our results is that they are stated in terms of a reduced form parameter from the population regression of the outcome on the instrument. In particular, we find that the reduced form parameter being far enough, but not too far, from zero, implies that the distribution of the observed data is consistent with our assumptions and the sign of ATE is identified, while the reduced form parameter being too far from zero implies that the distribution of the observed data is inconsistent with our assumptions. For each set of restrictions, we also develop methods for simultaneous inference about the validity of the assumptions and the sign of the ATE. We show that our inference procedures are valid uniformly over a large class of possible distributions for the observed data that include distributions where the instrument is arbitrarily "weak." A novel feature of the methodology is that the null hypotheses involve unions of moment inequalities.

KEYWORDS: Average Treatment Effect, Endogeneity, Instrumental Variables, Simultaneous Equation Model, Union of Moment Inequalities, Bootstrap, Uniform Validity, Multiple Testing, Familywise Error Rate

JEL Codes: C12, C31, C35, C36

1 Introduction

This paper considers identification and inference about the sign of the average effect of an endogenous regressor on an outcome of interest when an instrumental variable is available. In order to obtain simple, closed-form results and for ease of exposition, we focus on the case where the outcome of interest Y, endogenous regressor D and instrumental variable Z, whose joint distribution we denote by P, are all binary. In this setting, the endogenous regressor is sometimes referred to as the treatment and the average effect of the endogenous regressor on the outcome of interest is sometimes referred to as the average treatment effect (ATE). While maintaining instrument exogeneity, we consider three different sets of assumptions: monotonicity in the outcome equation, monotonicity in the equation for the endogenous regressor, or monotonicity in both the outcome equation and the equation for the endogenous regressor. Here, monotonicity in the outcome equation requires that different individuals do not have opposite responses to the endogenous regressor, whereas monotonicity in the equation for the endogenous regressor requires that different individuals do not have opposite responses to the instrumental variable. These conditions generally only provide partial identification of ATE.

For each set of assumptions, we show that the sign of the ATE is identified to be positive if and only if the reduced form parameter

$$
\begin{aligned}
\Delta(P) &= E_P[Y|Z=1] - E_P[Y|Z=0] \\
&= P\{Y=1|Z=1\} - P\{Y=1|Z=0\} \ .
\end{aligned}
\tag{1}
$$

lies in a particular region that depends only on P and that the sign of the ATE is identified to be negative if and only if $\Delta(P)$ lies in another region that, again, depends only on P. When imposing only monotonicity in the equation for the endogenous regressor, we find it is possible to determine the sign of the ATE if $\Delta(P)$ is sufficiently far from zero. When imposing monotonicity in both the outcome equation and the equation for the endogenous regressor, we find that the sign of the ATE equals the sign of $\Delta(P)$. When imposing monotonicity in only the outcome equation, however, we find that the sign of the ATE need not equal the sign of $\Delta(P)$. In fact, it is possible for $\Delta(P)$ to be so large that one concludes the sign of the ATE is negative. For each set of restrictions, we show further that a value for $\Delta(P)$ sufficiently far from zero implies that our assumptions are false. These results may be viewed as formalizing empirical researchers' suspicions of empirical results using instrumental variables when the reduced form parameter is "too large" (or, by re-scaling appropriately, when the usual linear instrumental variables estimand is "too large" – see Remark 2.1).

For each set of restrictions, we develop methods for simultaneous inference about the validity of the assumptions and the sign of the ATE. For this purpose, we consider a multiple testing problem with three null hypotheses, where rejection of the first null hypothesis means that P is inconsistent with the assumptions, rejection of the second null hypothesis means that P is consistent with the assumptions and only a positive ATE, and rejection of the third null hypothesis means that P is consistent with the assumptions and only a negative ATE. A novel feature of the analysis is that the null hypotheses involve unions of moment inequalities. We develop a bootstrap-based testing procedure that controls the familywise error rate – the probability of any false rejection – uniformly over a large class of possible distributions for P that include distributions where the instrument is arbitrarily "weak."

In the context of monotonicity in the equation for the endogenous regressor, our analysis is most closely related to Balke and Pearl (1997), who study partial identification of the ATE and also characterize when P is consistent with their assumptions. A characterization of consistency that does not require Y to be binary can be found in Kitagawa (2008), who builds upon the work of Imbens and Rubin (1997). Kitagawa (2008) and Bhattacharya et al. (2012) also develop tests for the null hypothesis of monotonicity in the equation for the endogenous regressor. Other related literature includes the local average treatment effect literature (LATE) (Imbens and Angrist, 1994) and the local instrumental variables/non-parametric selection model literature (Heckman and Vytlacil, 2001b), both of which impose monotonicity in the equation for the endogenous regressor. In the context of monotonicity in both the outcome equation and the equation for the endogenous regressor, our analysis is most closely related to Bhattacharya et al. (2012) and Shaikh and Vytlacil (2005, 2011), who study partial identification of the ATE, but do not characterize when P is consistent with the assumptions. Related results are obtained by Chiburis (2010), though under a different instrument exogeneity assumption. See also Abrevaya et al. (2010), who focus on inference about the sign of the average treatment effect in a semi-parametric model in a related context while allowing for the treatment to be non-binary and allowing for covariates. In the context of monotonicity in the outcome equation, the most closely related results are found in Chiburis (2010), though, as mentioned previously, under a different instrument exogeneity assumption.

The remainder of the paper proceeds as follows. In Section 2, we define our notation and the assumptions that will be used in the remainder of the paper. For each set of assumptions, we characterize in terms of $\Delta(P)$ in Section 3 when (i) P is inconsistent with the assumptions, (ii) when P is consistent with the assumptions and only a positive ATE, and (iii) when P is consistent with the assumptions and only a negative ATE. We further explore when P is inconsistent with our assumptions in Section 4. Finally, in Section 5, methods for inference are developed. Proofs of all results along with a numerical exploration of some of our results in Section 4 can be found in the Appendix.

2 Notation and Assumptions

Let Y denote a binary outcome of interest, D denote a binary endogenous regressor, and Z denote a binary instrument. For example, Y might denote mortality one year after the start of the experiment, D might denote receipt of the medical treatment, and Z random assignment to the medical treatment, where the randomized experiment suffers from noncompliance so that Z differs from D with positive probability. Further denote by Y_1 the potential outcome if treated, by Y_0 the potential outcome if not treated, by D_1 the potential value of the endogenous regressor if the instrument were to be externally set to 1, and by D_0 the potential value of the endogenous regressor if the instrument were to be externally set to 0. Following Angrist et al. (1996), we will refer to realizations with $D_1 > D_0$ as "compliers", realizations with $D_1 < D_0$ as "defiers", realizations with $D_1 = 1$ and $D_0 = 1$ as "always takers," and realizations with $D_1 = 0$ and $D_0 = 0$ as "never takers." In this notation,

$$Y = DY_1 + (1 - D)Y_0 \tag{2}$$
$$D = ZD_1 + (1 - Z)D_0. \tag{3}$$

Let P be the distribution of (Y, D, Z) and Q be the distribution of (Y_0, Y_1, D_0, D_1, Z). Since

$$(Y, D, Z) = T(Y_0, Y_1, D_0, D_1, Z) \ ,$$

where T is characterized by (2) and (3), we have that

$$P = QT^{-1} \ .$$

Below we will restrict $Q \in \mathbf{Q}$, where $\mathbf{Q}$ is a set of distributions for (Y_0, Y_1, D_0, D_1, Z) satisfying certain restrictions. In particular, we will require Z to be an instrument in the sense that every $Q \in \mathbf{Q}$ satisfies the following exogeneity condition:

Assumption 2.1 *(Instrument Exogeneity): $Z \perp\!\!\!\perp (Y_0, Y_1, D_0, D_1)$ under Q.*

We will additionally require that every $Q \in \mathbf{Q}$ satisfy at least one of the following monotonicity conditions:

Assumption 2.2 *(Monotonicity of D in Z): $Q\{D_1 \geq D_0\} = 1$ or $Q\{D_1 \leq D_0\} = 1$.*

Assumption 2.3 *(Monotonicity of Y in D): $Q\{Y_1 \geq Y_0\} = 1$ or $Q\{Y_1 \leq Y_0\} = 1$.*

We do not impose instrument relevance, i.e., we allow for $P\{D = 1|Z = 1\} = P\{D = 1|Z = 0\}$. Without loss of generality, we will order Z such that $P\{D = 1|Z = 1\} \geq P\{D = 1|Z = 0\}$. Given this ordering and Assumption 2.1, we have that Assumption 2.2 is equivalent to the restriction that $Q\{D_1 \geq D_0\} = 1$.

Our object of interest is the average effect of the endogenous regressor on the outcome, defined to be

$$E_Q[Y_1 - Y_0] = Q\{Y_1 = 1\} - Q\{Y_0 = 1\} \ . \tag{4}$$

This quantity is typically referred to in the treatment effect literature as the average treatment effect (ATE). It will be useful to partition $\mathbf{Q}$ as $\mathbf{Q} = \mathbf{Q}_+ \cup \mathbf{Q}_0 \cup \mathbf{Q}_-$, where

$$
\begin{aligned}
\mathbf{Q}_+ &= \{Q \in \mathbf{Q} : Q\{Y_1 = 1\} - Q\{Y_0 = 1\} > 0\} \\
\mathbf{Q}_0 &= \{Q \in \mathbf{Q} : Q\{Y_1 = 1\} - Q\{Y_0 = 1\} = 0\} \\
\mathbf{Q}_- &= \{Q \in \mathbf{Q} : Q\{Y_1 = 1\} - Q\{Y_0 = 1\} < 0\} \ ,
\end{aligned}
$$

and define

$$
\begin{aligned}
\mathbf{Q}_{0,+} &= \mathbf{Q}_+ \cup \mathbf{Q}_0 \\
\mathbf{Q}_{0,-} &= \mathbf{Q}_- \cup \mathbf{Q}_0 \ .
\end{aligned}
$$

In other words, $\mathbf{Q}_-$ ($\mathbf{Q}_{0,-}$) is the set of distributions for (Y_0, Y_1, D_0, D_1, Z) satisfying our restrictions and having a (weakly) negative ATE, $\mathbf{Q}_0$ is the set of distributions for (Y_0, Y_1, D_0, D_1, Z) satisfying our restrictions and having a zero ATE, and $\mathbf{Q}_+$ ($\mathbf{Q}_{0,+}$) is the set of distributions for (Y_0, Y_1, D_0, D_1, Z) satisfying our restrictions and having a (weakly) positive ATE. In this notation, the ATE is identified to be positive if

$$P \in \mathbf{Q}_+ T^{-1} \cap (\mathbf{Q}_{0,-} T^{-1})^c \ , \tag{5}$$

where $\mathbf{Q}_+T^{-1} = \{QT^{-1} : Q \in \mathbf{Q}_+\}$; $\mathbf{Q}_-T^{-1}$, $\mathbf{Q}_{0,-}T^{-1}$ and $\mathbf{Q}_{0,+}T^{-1}$ are defined similarly. In other words, we identify the ATE to be positive if the distribution of (Y, D, Z) is consistent with our restrictions holding with a positive ATE but not consistent with our restrictions holding with a zero or negative ATE. Symmetrically, the ATE is identified to be negative if

$$P \in \mathbf{Q}_-T^{-1} \cap (\mathbf{Q}_{0,+}T^{-1})^c . \tag{6}$$

Analogously, the distribution of the observed data, P, is consistent with our restrictions if

$$P \in \mathbf{Q}T^{-1} . \tag{7}$$

For completeness, we note that the identified set for the ATE, as a function of P, is given by

$$\{E_Q[Y_1 - Y_0] : Q \in \mathbf{Q} \text{ and } P = QT^{-1}\} .$$

Remark 2.1 Our results below will be stated in terms of the reduced form parameter $\Delta(P)$, defined in (1). In the biostatistics literature, when Z is random assignment to treatment with possible non-compliance, $\Delta(P)$ is sometimes referred to as the "intention-to-treat" parameter. If the instrument is relevant, i.e., $P\{D = 1|Z = 1\} \neq P\{D = 1|Z = 0\}$, then, under mild regularity conditions, the usual linear instrumental variables estimand in this setting is simply $\Delta(P)$ divided by $P\{D = 1|Z = 1\} - P\{D = 1|Z = 0\}$. Under our assumptions, the sign of $\Delta(P)$ and the usual linear instrumental variables estimand are therefore the same. As a result, it will be straightforward to re-scale our results to state them in terms of this quantity. ∎

Remark 2.2 Note that Assumption 2.2 is the same monotonicity assumption found in Imbens and Angrist (1994), who also refer to it as an assumption of "no defiers." It follows from results in Vytlacil (2002) that this assumption is equivalent to the selection model of Heckman and Vytlacil (2001b, 2005). In particular, it is equivalent to assuming that there exists a representation of the model as

$$D_z = I\{\delta_0 + \delta_1 z + \eta \geq 0\} \tag{8}$$

with δ_1 being nonrandom. Similarly, Assumption 2.3 is equivalent to assuming that there exists a representation of the model as

$$Y_d = I\{\beta_0 + \beta_1 d + \epsilon \geq 0\} \tag{9}$$

with β_1 nonrandom, and Assumptions 2.2 and 2.3 is equivalent to assuming both (8) and (9) with δ_1 and β_1 nonrandom. ∎

Remark 2.3 A stronger version of Assumption 2.3 in which it is assumed further that the direction of the monotonicity is known *a priori* is referred to as the "monotone treatment response" assumption by Manski (1997) and Manski and Pepper (2000). They characterize the identified set for the ATE under this type of restriction. As discussed by Bhattacharya et al. (2008), these results do not hold if only Assumption 2.3 is assumed. In some settings, it may not be reasonable to assume that the direction of the effect is known *a priori*. Our analysis, which focuses on the sign of the ATE, is useful in such settings. ∎

3 Identifying the Sign of the Average Treatment Effect from IV

In this section, for each of our three possible restrictions on $\mathbf{Q}$, we characterize whether P satisfies (5), (6) or (7) in terms of $\Delta(P)$.

3.1 Monotonicity of D in Z

In this section, we assume that every $Q \in \mathbf{Q}$ satisfies Assumptions 2.1 and 2.2. In this case, our results essentially follow from Balke and Pearl (1997), who characterize the identified set for the ATE under these assumptions and also when P is consistent with these restrictions. See also Heckman and Vytlacil (2001a) and Kitagawa (2008), who generalize these results.

In order to state our results, we require some additional notation. Define

$$
\begin{aligned}
A_1(P) &= \max\{A_1^1(P), A_1^2(P)\}, \\
A_2(P) &= -P\{Y = 0, D = 0 | Z = 1\} - P\{Y = 1, D = 1 | Z = 0\} \\
A_3(P) &= P\{Y = 1, D = 0 | Z = 1\} + P\{Y = 0, D = 1 | Z = 0\} \\
A_4(P) &= \min\{A_4^1(P), A_4^2(P)\},
\end{aligned}
$$

where

$$
\begin{aligned}
A_1^1(P) &= P\{Y = 1, D = 0 | Z = 1\} - P\{Y = 1, D = 0 | Z = 0\}, \\
A_1^2(P) &= P\{Y = 0, D = 1 | Z = 0\} - P\{Y = 0, D = 1 | Z = 1\}, \\
A_4^1(P) &= P\{Y = 1, D = 1 | Z = 1\} - P\{Y = 1, D = 1 | Z = 0\}, \\
A_4^2(P) &= P\{Y = 0, D = 0 | Z = 0\} - P\{Y = 0, D = 0 | Z = 1\}.
\end{aligned}
$$

Note that under our assumptions $A_2(P) \leq A_4(P)$, $A_1(P) \leq A_3(P)$, and $A_2(P) \leq 0 \leq A_3(P)$.

Theorem 3.1 *If every $Q \in \mathbf{Q}$ satisfies Assumptions 2.1 and 2.2, then*

(i) $P \in \mathbf{Q}T^{-1}$ if and only if

$$\Delta(P) \in [A_1(P), A_4(P)] . \tag{10}$$

(ii) $P \in \mathbf{Q}_+ T^{-1} \cap (\mathbf{Q}_{0,-} T^{-1})^c$ if and only if

$$\Delta(P) \in (A_3(P), A_4(P)] .$$

(iii) $P \in \mathbf{Q}_- T^{-1} \cap (\mathbf{Q}_{0,+} T^{-1})^c$ if and only if

$$\Delta(P) \in [A_1(P), A_2(P)) .$$

Remark 3.1 Part (i) of Theorem 3.1 implies that P is inconsistent with our restrictions if and only if $\Delta(P) \notin [A_1(P), A_4(P)]$ Hence, P is inconsistent with our restrictions if and only if (a) $A_1(P) > A_4(P)$, (b) $A_1(P) \leq A_4(P)$ and $\Delta(P) < A_1(P)$, or (c) $A_1(P) \leq A_4(P)$ and $\Delta(P) > A_4(P)$. If $A_1(P) \leq A_4(P)$ and

$\Delta(P) < A_1(P)$, then it is possible to show that $A_1(P) \leq 0$. Similarly, if $A_1(P) \leq A_4(P)$ and $\Delta(P) > A_4(P)$, then it is possible to show that $A_4(P) \geq 0$. In this sense, part (i) of Theorem 3.1 implies that P is inconsistent with our restrictions whenever $\Delta(P)$ is "too far" from zero. ∎

Remark 3.2 Parts (ii) and (iii) of Theorem 3.1 imply that we are both unable to reject our restrictions and unable to determine the sign of the ATE whenever $\Delta(P)$ is "too close" to zero, i.e.,

$$\Delta(P) \in [A_2(P), A_3(P)] \ .$$

More generally, the width of the region of indeterminacy is bounded from above by

$$2(P\{D = 0|Z = 1\} + P\{D = 1|Z = 0\}) = 2(1 - Q\{D_1 > D_0\}) \ ,$$

which decreases with the strength of the instrument, as measured by $P\{D = 1|Z = 1\} - P\{D = 1|Z = 0\} = Q\{D_1 > D_0\}$. Using results in Imbens and Angrist (1994), we have that

$$\Delta(P) = E_Q[Y_1 - Y_0|D_1 > D_0]Q\{D_1 > D_0\}$$

under Assumptions 2.1 and 2.2. The reduced form parameter $\Delta(P)$ thus combines the strength of the instrument with the strength of the treatment on "compliers." In this way, the sign of the ATE is easier to determine when the instrument is stronger or the effect of the treatment on the "compliers" is stronger. ∎

Remark 3.3 Part (i) of Theorem 3.1 is derived from results in Balke and Pearl (1997). A more general result that does not require Y to be binary can be found in Kitagawa (2008), who builds upon the work of Imbens and Rubin (1997). Kitagawa (2008) also develops a testing procedure. For binary Y, Bhattacharya et al. (2012) develop a test of Assumptions 2.1 and 2.2 by comparing the bounds on the ATE in Manski (1990) with those in Heckman and Vytlacil (2001a). The resulting conditions are in fact equivalent to part (i) of Theorem 3.1. ∎

3.2 Monotonicity of Y in D and Monotonicity of D in Z

In this section, we assume that every $Q \in \mathbf{Q}$ satisfies Assumptions 2.1, 2.2 and 2.3. These restrictions have been previously considered in the literature by Bhattacharya et al. (2008, 2012) and Shaikh and Vytlacil (2005, 2011), who find that the sign of ATE equals the sign of $\Delta(P)$. The following theorem re-states this result and additionally characterizes when $P \in \mathbf{Q}T^{-1}$ in terms of $\Delta(P)$. We emphasize that this additional result is not found in either Bhattacharya et al. (2012) or Shaikh and Vytlacil (2005, 2011).

Theorem 3.2 *If every $Q \in \mathbf{Q}$ satisfies Assumptions 2.1, 2.2 and 2.3, then*

(i) $P \in \mathbf{Q}T^{-1}$ if and only if

$$\Delta(P) \in [A_1(P), A_4(P)] \ , \tag{11}$$

(ii) $P \in \mathbf{Q}_+T^{-1} \cap (\mathbf{Q}_{0,-}T^{-1})^c$ if and only if

$$\Delta(P) \in [A_1(P), 0) \ ,$$

(iii) $P \in \mathbf{Q}_-T^{-1} \cap (\mathbf{Q}_{0,+}T^{-1})^c$ *if and only if*

$$\Delta(P) \in (0, A_4(P)] \ .$$

Remark 3.4 Note that the conditions on $\Delta(P)$ in (11) that determine whether or not P is consistent with our assumptions are exactly the same as the ones in (10). It is therefore never possible to reject Assumption 2.2 without rejecting Assumptions 2.2 and 2.3 together. ∎

Remark 3.5 In contrast to our earlier results, the only circumstance in which we are both unable to reject our restrictions and unable to determine the sign of the ATE is if $\Delta(P) = 0$. ∎

3.3 Monotonicity of Y in D

In this section, we assume that every $Q \in \mathbf{Q}$ satisfies Assumptions 2.1 and 2.3. Note that Assumption 2.3 has not been considered without Assumption 2.2 previously in the literature. In order to state our results, we require some additional notation. Define

$$
\begin{aligned}
B_1(P) &= \max\{-P\{Y = 1, D = 1 | Z = 0\}, -P\{Y = 0, D = 0 | Z = 1\}\} \\
B_2(P) &= \max\{P\{Y = 1, D = 1 | Z = 1\}, P\{Y = 0, D = 0 | Z = 0\}\} \\
B_3(P) &= \max\{-P\{Y = 0, D = 1 | Z = 1\}, -P\{Y = 1, D = 0 | Z = 0\}\} \\
B_4(P) &= \max\{P\{Y = 0, D = 1 | Z = 0\}, P\{Y = 1, D = 0 | Z = 1\}\} \ .
\end{aligned}
$$

Note that $B_1(P) \le 0$ and $B_3(P) \le 0$, while $B_2(P) \ge 0$ and $B_4(P) \ge 0$. Using this notation, we have the following theorem:

Theorem 3.3 *If every $Q \in \mathbf{Q}$ satisfies Assumptions 2.1 and 2.3, then*

(i) $P \in \mathbf{Q}T^{-1}$ *if and only if*

$$\Delta(P) \in [\min\{B_1(P), B_3(P)\}, \max\{B_2(P), B_4(P)\}] \ , \tag{12}$$

(ii) $P \in \mathbf{Q}_+T^{-1} \cap (\mathbf{Q}_{0,-}T^{-1})^c$ *if and only if*

$$\Delta(P) \in [B_1(P), B_2(P)] \setminus [B_3(P), B_4(P)] \ ,$$

(iii) $P \in \mathbf{Q}_-T^{-1} \cap (\mathbf{Q}_{0,+}T^{-1})^c$ *if and only if*

$$\Delta(P) \in [B_3(P), B_4(P)] \setminus [B_1(P), B_2(P)] \ .$$

Remark 3.6 Analogously to our earlier results, part (i) of Theorem 3.3 implies that P is inconsistent with our assumptions if and only if $\Delta(P)$ is "too far" from zero. Here, "too far" means $\Delta(P) < \min\{B_1(P), B_3(P)\}$ or $\Delta(P) > \max\{B_2(P), B_4(P)\}$. Since $A_1(P) \ge B_3(P)$ and $A_4(P) \le B_2(P)$,

$$[A_1(P), A_4(P)] \subseteq [\min\{B_1(P), B_3(P)\}, \max\{B_2(P), B_4(P)\}] \ .$$

Furthermore, the inclusion may be strict, so it is possible to reject Assumption 2.2 without rejecting Assumption 2.3. ∎

Remark 3.7 Parts (ii) and (iii) of Theorem 3.3 imply that we are both unable to reject our restrictions and unable to determine the sign of the ATE if $\Delta(P)$ is "too close" to zero, i.e.,

$$\Delta(P) \in [\max\{B_1(P), B_3(P)\}, \min\{B_2(P), B_4(P)\}]$$

where this interval necessarily includes zero. The width of this region of indeterminacy is bounded by

$$2(P\{D = 0 | Z = 1\} + P\{D = 1 | Z = 0\}) = 2(1 - Q\{D_1 > D_0\} + Q\{D_1 < D_0\}) ,$$

which decreases as the instrument gets stronger, as measured by $P\{D = 1 | Z = 1\} - P\{D = 1 | Z = 0\} = Q\{D_1 > D_0\} - Q\{D_1 < D_0\}$. In this sense, the sign of the ATE is again easier to determine when the instrument is stronger. ∎

Remark 3.8 A possibly counterintuitive implication of Theorem 3.3 is that it is possible for $\Delta(P)$ to be so large that one determines the sign of the ATE is in fact negative or for $\Delta(P)$ to be so small that one determines the sign of the ATE is in fact positive. The first case happens when

$$\min\{B_2(P), B_4(P)\} < \Delta(P) \leq \max\{B_2(P), B_4(P)\} \text{ and } \max\{B_2(P), B_4(P)\} = B_4(P) , \qquad (13)$$

whereas the second case happens when

$$\min\{B_1(P), B_3(P)\} \leq \Delta(P) < \max\{B_1(P), B_3(P)\} \text{ and } \max\{B_1(P), B_3(P)\} = B_3(P) . \qquad (14)$$

In order to better understand this result, it is instructive to note that

$$\Delta(P) = \begin{cases} Q\{Y_1 > Y_0, D_1 > D_0\} - Q\{Y_1 > Y_0, D_1 < D_0\} & \text{if } Y_1 \geq Y_0 \\ Q\{Y_1 < Y_0, D_1 < D_0\} - Q\{Y_1 < Y_0, D_1 > D_0\} & \text{if } Y_1 \leq Y_0 \end{cases} .$$

The first case occurs when $Q\{Y_1 < Y_0, D_1 < D_0\} > Q\{Y_1 < Y_0, D_1 > D_0\}$, so we require enough "defiers" with a negative treatment effect, and the second case occurs when $Q\{Y_1 > Y_0, D_1 > D_0\} < Q\{Y_1 > Y_0, D_1 < D_0\}$, so we require enough "defiers" with a positive treatment effect. Note further that

$$\begin{aligned} B_1(P) \leq & \quad Q\{Y_1 > Y_0, D_1 > D_0\} - Q\{Y_1 > Y_0, D_1 < D_0\} \quad \leq B_2(P) \\ B_3(P) \leq & \quad Q\{Y_1 < Y_0, D_1 < D_0\} - Q\{Y_1 < Y_0, D_1 > D_0\} \quad \leq B_4(P) . \end{aligned}$$

It follows that it must be the case that $Y_1 \leq Y_0$ whenever $\Delta(P) \in (B_2(P), B_4(P)] \subseteq (0, 1]$ and that $Y_1 \geq Y_0$ whenever $\Delta(P) \in [B_1(P), B_3(P)) \subseteq [-1, 0)$. ∎

Remark 3.9 The counterintuitive result described in Remark 3.8 can only arise when P is inconsistent with Assumption 2.2 but consistent with Assumption 2.3. To see this, note that (13) or (14) together with the fact that $A_1(P) \geq B_3(P)$ and $A_4(P) \leq B_2(P)$ imply that $\min\{B_1(P), B_3(P)\} \leq \Delta(P) < A_1(P)$ or $A_4(P) < \Delta(P) \leq \max\{B_2(P), B_4(P)\}$, which implies that $\Delta(P) \notin [A_1(P), A_4(P)]$ and $\Delta(P) \in [\min\{B_1(P), B_3(P)\}, \max\{B_2(P), B_4(P)\}]$. ∎

4   Detecting Failure of the Monotonicity Restrictions

In the preceding section, we characterized when P was consistent with our restrictions in terms of the reduced form parameter $\Delta(P)$. In particular, we showed that in each case a value of $\Delta(P)$ sufficiently far from zero implied that the restrictions were violated. In this section, we maintain instrument exogeneity – Assumption 2.1 – and characterize which types of violations of Assumption(s) 2.2 (and 2.3) are detectable in the sense that they lead to $\Delta(P)$ to be sufficiently far from zero. To complement the analytical results in this section, we also provide some numerical results in Appendix A, where we explore violations of the monotonicity restrictions being detectable in the context of a parametric model for Y and D.

4.1   Monotonicity of D in Z (and Monotonicity of Y in D)

Parts (i) of Theorems 3.1 and 3.2 showed that P is consistent with Assumption(s) 2.2 (and 2.3) if and only if $\Delta(P)$ satisfies (10). The following proposition characterizes distributions Q satisfying Assumptions 2.1 for which $\Delta(P)$ fails to satisfy (10).

Proposition 4.1 *If $P = QT^{-1}$ for a distribution Q that satisfies Assumptions 2.1, then $\Delta(P) \notin [A_1(P), A_4(P)]$ if and only if*

$$Q\{Y_j = k, D_1 < D_0\} > Q\{Y_j = k, D_1 > D_0\}$$

for some $(j, k) \in \{0, 1\}^2$.

Remark 4.1 Given our normalization that $P\{D = 1 | Z = 1\} \geq P\{D = 1 | Z = 0\}$ and given Assumption 2.1, we have that the fraction of "compliers," $Q\{D_1 > D_0\}$, at least weakly exceeds the fraction of "defiers," $Q\{D_1 < D_0\}$, and does so by the magnitude of $P\{D = 1 | Z = 1\} - P\{D = 1 | Z = 0\} = Q\{D_1 > D_0\} - Q\{D_1 < D_0\}$. Thus, Proposition 4.1 implies that in order to detect a violation of Assumption 2.2 it must be the case that the fraction of "defiers" is sufficiently large (which in turn requires the instrument be sufficiently weak in that $P\{D = 1 | Z = 1\} - P\{D = 1 | Z = 0\}$ is sufficiently small) and that the distribution of potential outcomes among "defiers" and "compliers" differs, i.e.,

$$Q\{Y_j = 1 | D_1 < D_0\} \neq Q\{Y_j = 1 | D_1 > D_0\}$$

for some $j \in \{0, 1\}$. ∎

4.2   Monotonicity of Y in D

Part (i) of Theorem 3.3 showed that P is consistent with Assumption 2.3 if and only if $\Delta(P)$ satisfies (12). The following proposition characterizes distributions Q satisfying Assumption 2.1 for which $\Delta(P)$ fails to

satisfy (12). In order to state our results, we require some additional notation. Define

$$
\begin{aligned}
M_1^1 &= Q\{Y_1 > Y_0, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 1, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 1, D_1 < D_0\} \\
M_1^2 &= Q\{Y_1 > Y_0, D_1 = D_0 = 0\} + Q\{Y_1 = Y_0 = 0, D_1 = D_0 = 0\} + Q\{Y_1 = Y_0 = 0, D_1 < D_0\} \\
M_2^1 &= Q\{Y_1 > Y_0, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 1, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 1, D_1 > D_0\} \\
M_2^2 &= Q\{Y_1 > Y_0, D_1 = D_0 = 0\} + Q\{Y_0 = Y_1 = 0, D_1 = D_0 = 0\} + Q\{Y_0 = Y_1 = 0, D_1 > D_0\} \\
M_3^1 &= Q\{Y_1 < Y_0, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 0, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 0, D_1 > D_0\} \\
M_3^2 &= Q\{Y_1 < Y_0, D_1 = D_0 = 0\} + Q\{Y_0 = Y_1 = 1, D_1 = D_0 = 0\} + Q\{Y_0 = Y_1 = 1, D_1 > D_0\} \\
M_4^1 &= Q\{Y_1 < Y_0, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 0, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 0, D_1 < D_0\} \\
M_4^2 &= Q\{Y_1 < Y_0, D_1 = D_0 = 0\} + Q\{Y_1 = Y_0 = 1, D_1 = D_0 = 0\} + Q\{Y_1 = Y_0 = 1, D_1 < D_0\}
\end{aligned}
$$

and, for $1 \le j \le 4$, let

$$
M_j = \min\{M_j^1, M_j^2\} .
$$

Using this notation, we have the following result:

Proposition 4.2 *If $P = QT^{-1}$ for a distribution Q that satisfies Assumptions 2.1, then*

$$
\Delta(P) \notin [\min\{B_1(P), B_3(P)\}, \max\{B_2(P), B_4(P)\}]
$$

if and only if either

$$
\begin{aligned}
Q\{Y_1 > Y_0, D_1 > D_0\} + Q\{Y_1 < Y_0, D_1 < D_0\} \\
< \min\{Q\{Y_1 < Y_0, D_1 > D_0\} - M_1, Q\{Y_1 > Y_0, D_1 < D_0\} - M_3\}
\end{aligned} \tag{15}
$$

or

$$
\begin{aligned}
Q\{Y_1 < Y_0, D_1 > D_0\} + Q\{Y_1 > Y_0, D_1 < D_0\} \\
< \min\{Q\{Y_1 < Y_0, D_1 < D_0\} - M_2, Q\{Y_1 > Y_0, D_1 > D_0\} - M_4\} .
\end{aligned} \tag{16}
$$

Remark 4.2 Note that if there are no "defiers," then it is impossible for either (15) or (16) to hold. Hence, it is only possible to detect violations of Assumption 2.3 if Assumption 2.2 does not hold. ■

Remark 4.3 In order to satisfy (15), there must be strong negative dependence between $Y_1 - Y_0$ and $D_1 - D_0$. In addition, it seems that the probability of being an "always taker" or "never taker" must be small so that M_1 and M_3 will be small. For instance, (15) is satisfied when $Q\{Y_1 < Y_0 | D_1 > D_0\} = 1$, $Q\{Y_1 > Y_0 | D_1 < D_0\} = 1$ and $Q\{D_1 = D_0\} = 0$. Analogous comments apply to (16). In this sense, it seems that the requirements on Q in order to satisfy either (15) or (16) are rather extreme. The numerical results in Appendix A further highlight the difficulty of detecting violations of Assumption 2.3. ■

5 Inference

In this section, we let $(Y_i, D_i, Z_i), i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution $P \in \mathbf{P}$ on $\{0,1\}^3$ and, for each of the three sets of restrictions considered in the previous sections, consider the problem of simultaneously testing the restrictions and making inferences about the sign of the ATE when those restrictions are not rejected. More precisely, for each set of restrictions on $\mathbf{Q}$, we will consider the problem of testing the family of null hypotheses

$$H_j : P \in \mathbf{P}_j \text{ for } 1 \leq j \leq 3 , \tag{17}$$

where

$$
\begin{aligned}
\mathbf{P}_1 &= \{P \in \mathbf{P} : P \in \mathbf{Q}T^{-1}\} \\
\mathbf{P} \setminus \mathbf{P}_2 &= \{P \in \mathbf{P} : P \in \mathbf{Q}_+T^{-1}\} \\
\mathbf{P} \setminus \mathbf{P}_3 &= \{P \in \mathbf{P} : P \in \mathbf{Q}_-T^{-1}\} .
\end{aligned}
$$

in a way that satisfies

$$\limsup_{n \to \infty} \sup_{P \in \mathbf{P}} FWER_P \leq \alpha . \tag{18}$$

Here,

$$FWER_P = P\{\text{any false rejection}\} .$$

Note that $\mathbf{P}_1$ equals the set of distributions $P \in \mathbf{P}$ that are consistent with our restrictions, $\mathbf{P} \setminus \mathbf{P}_2$ equals the set of distributions $P \in \mathbf{P}$ that are consistent with our restrictions and the sign of the ATE being positive, and $\mathbf{P} \setminus \mathbf{P}_3$ equals the set of distributions $P \in \mathbf{P}$ that are consistent with our restrictions and the sign of the ATE being negative. Note further that at most one null hypothesis may be false. Our testing procedures below will therefore have the feature that at most one null hypothesis will be rejected. If H_1 is rejected, then we will conclude that P is inconsistent with our restrictions; if H_2 is rejected, then we will conclude that P is consistent with our restrictions and a positive ATE; if H_3 is rejected, then we will conclude that P is consistent with our restrictions and a negative ATE.

Below we will assume that that $\mathbf{P}$ is such that

$$\inf_{P \in \mathbf{P}} \inf_{(y,d,z) \in \{0,1\}^3} P\{Y = y, D = d, Z = z\} > \epsilon$$

for some $\epsilon > 0$. We will also denote by $\hat{P}_n$ the empirical distribution of $(Y_i, D_i, Z_i), i = 1, \ldots, n$.

5.1 Monotonicity of D in Z

In this section, we assume every $Q \in \mathbf{Q}$ satisfies Assumptions 2.1 and 2.2. For this choice of $\mathbf{Q}$, it follows from Theorem 3.1 that

$$
\begin{align}
\mathbf{P}_1 &= \{P \in \mathbf{P} : A_1(P) \leq \Delta(P) \leq A_4(P)\} \tag{19} \\
\mathbf{P}_2 &= \{P \in \mathbf{P} : A_3(P) \geq \Delta(P) \cup \Delta(P) > A_4(P)\} \tag{20} \\
\mathbf{P}_3 &= \{P \in \mathbf{P} : A_1(P) > \Delta(P) \cup \Delta(P) \geq A_2(P)\} . \tag{21}
\end{align}
$$

In order to describe our testing procedure, it is useful to introduce some further notation. Define

$$
\begin{align}
a_1(P) = -a_8(P) &= A_1^1(P) - \Delta(P) \\
a_2(P) = -a_9(P) &= A_1^2(P) - \Delta(P) \\
a_3(P) = -a_6(P) &= \Delta(P) - A_4^1(P) \\
a_4(P) = -a_7(P) &= \Delta(P) - A_4^2(P) \\
a_5(P) &= \Delta(P) - A_3(P) \\
a_{10}(P) &= A_2(P) - \Delta(P) .
\end{align}
$$

For $1 \leq j \leq 3$, define

$$
T_{j,n}^1 = \min_{K \in \mathcal{K}_j^1} \max_{k \in K} \frac{a_k(\hat{P}_n)}{\hat{\sigma}_{k,n}^a} ,
$$

where

$$
\begin{align}
\mathcal{K}_1^1 &= \{\{1, 2, 3, 4\}\} \\
\mathcal{K}_2^1 &= \{\{5\}, \{6\}, \{7\}\} \\
\mathcal{K}_3^1 &= \{\{8\}, \{9\}, \{10\}\} ,
\end{align}
$$

and $\hat{\sigma}_{k,n}^a$ for $1 \leq k \leq 10$ is the usual (unpooled) estimate of the standard deviation of $a_k(\hat{P}_n)$. Note that at most one of the $T_{j,n}^1$ will be strictly positive. Define

$$
\hat{c}_{1,n}(1 - \alpha) = \max_{K \in \mathcal{K}_1^*} J_{1,n}^{-1}(1 - \alpha, K, \hat{P}_n) , \tag{22}
$$

where

$$
\begin{align}
J_{1,n}(x, K, P) &= P\left\{ \max_{k \in K} \frac{a_k(\hat{P}_n) - a_k(P)}{\hat{\sigma}_{k,n}^a} \leq x \right\} \\
\mathcal{K}_1^* &= \{A \cup B \cup C : A \in \mathcal{K}_1^1, B \in \mathcal{K}_2^1, C \in \mathcal{K}_3^1\} .
\end{align}
$$

Theorem 5.1 *Consider testing (17) with $\mathbf{P}_1$, $\mathbf{P}_2$ and $\mathbf{P}_3$ given by (19), (20), and (21), respectively. The testing procedure that rejects $H_{\hat{j}}$ whenever $T_{\hat{j},n}^1 > \hat{c}_{1,n}(1 - \alpha)$, where $\hat{j} = \arg\max_{1 \leq j \leq 3} T_{j,n}^1$, satisfies (18).*

Remark 5.1 The critical value $\hat{c}_{1,n}(1 - \alpha)$ in Theorem 5.1 may be larger than necessary. By exploiting the

structure of the null hypotheses, it is in fact possible to replace $\mathcal{K}_1^*$ in Theorem 5.1 with

$$\{\{1,\ldots,5,10\}\} \cup \bigcup_{5 \leq j \leq 7, 8 \leq k \leq 10} \{\{j,k\}\} . \tag{23}$$

To see this, it suffices to examine the proof of Theorem 5.1 in the Appendix and note that (38) continues to hold if $\mathcal{K}_I^*$ is replaced with

$$\tilde{\mathcal{K}}_I^* = \left\{ \tilde{K} \in \mathcal{K}_I^* : \mathbf{P}_{\tilde{K}} \cap \{P \in \mathbf{P} : I(P) = I\} \neq \emptyset \right\} ,$$

where $I(P) = \{1 \leq j \leq 3 : P \in \mathbf{P}_j\} \subseteq \{1,2,3\}$. Using that $I(P) \in \mathcal{I}$, where

$$\mathcal{I} = \{\{1,2\},\{2,3\},\{1,3\},\{1,2,3\}\} ,$$

it follows that we may replace $\mathcal{K}_1^*$ in (22) with

$$\bigcup_{I \in \mathcal{I}} \tilde{\mathcal{K}}_I^* . \tag{24}$$

Since $J_{1,n}^{-1}(1-\alpha, \tilde{K}, \hat{P}_n)$ is monotone in $\tilde{K}$ with respect to set inclusion, we may further restrict attention to only those sets $\tilde{K}$ in (24) such that there is no strictly larger set with respect to set inclusion. The sets that remain after doing so are given by (23). $\blacksquare$

Remark 5.2 The critical value $\hat{c}_{1,n}(1-\alpha)$ in (22) may be viewed as a "least favorable" critical value in the same way that critical values based on assuming that all moments are binding in the moment inequality literature are "least favorable." To see this, it is useful to note that $\hat{c}_{1,n}(1-\alpha)$ is the same critical value that would be used to test the null hypothesis that

$$P \in \bigcap_{1 \leq j \leq 3} \mathbf{P}_j = \bigcup_{K \in \mathcal{K}_1^*} \{P \in \mathbf{P} : a_k(P) \leq 0\}$$

at level α using the test statistic $\max_{1 \leq j \leq 3} T_{j,n}^1$. In contrast to the moment inequality literature, where the null hypotheses only involve a single set of inequalities, the null hypothesis involves a union of different sets of inequalities. As a result, there is no longer a single "least favorable" critical value, but rather one for each set of inequalities in the union. It is for this reason that the maximum appears in (22). It is possible to construct critical values that are not "least favorable" by modifying other approaches in the moment inequality literature, such as the recent approach by Romano et al. (2012). $\blacksquare$

5.2 Monotonicity of Y in D and D in Z

In this section, we assume every $Q \in \mathbf{Q}$ satisfies Assumptions 2.1, 2.2 and 2.3. For this choice of $\mathbf{Q}$, it follows from Theorem 3.2 that

$$
\begin{align}
\mathbf{P}_1 &= \{P \in \mathbf{P} : A_1(P) \leq \Delta(P) \leq A_4(P)\} \tag{25} \\
\mathbf{P}_2 &= \{P \in \mathbf{P} : 0 \geq \Delta(P) \cup \Delta(P) > A_4(P)\} \tag{26} \\
\mathbf{P}_3 &= \{P \in \mathbf{P} : A_1(P) > \Delta(P) \cup \Delta(P) \geq 0\} . \tag{27}
\end{align}
$$

14

Recall the definitions of $a_k(P)$ and $\hat{\sigma}_{k,n}^a$ for $1 \leq k \leq 10$ in Section 5.1 and define

$$a_{11}(P) = -a_{12}(P) \quad = \quad \Delta(P) \ .$$

For $1 \leq j \leq 3$, define

$$T_{j,n}^2 = \min_{K \in \mathcal{K}_j^2} \max_{k \in K} \frac{a_k(\hat{P}_n)}{\hat{\sigma}_{k,n}^a} \ ,$$

where

$$
\begin{aligned}
\mathcal{K}_1^2 &= \{\{1,2,3,4\}\} \\
\mathcal{K}_2^2 &= \{\{6\},\{7\},\{11\}\} \\
\mathcal{K}_3^2 &= \{\{8\},\{9\},\{12\}\} \ ,
\end{aligned}
$$

and $\hat{\sigma}_{k,n}^a$ for $11 \leq k \leq 12$ is the usual (unpooled) estimate of the standard deviation of $a_k(\hat{P}_n)$. Note that at most one of the $T_{j,n}^3$ will be strictly positive. Define

$$\hat{c}_{2,n}(1-\alpha) = \max_{K \in \mathcal{K}_3^*} J_{2,n}^{-1}(1-\alpha, K, \hat{P}_n) \ ,$$

where

$$
\begin{aligned}
J_{2,n}(x, K, P) &= P\left\{\max_{k \in K} \frac{a_k(\hat{P}_n) - a_j(P)}{\hat{\sigma}_{k,n}^a} \leq x\right\} \\
\mathcal{K}_2^* &= \{A \cup B \cup C : A \in \mathcal{K}_1^2, B \in \mathcal{K}_2^2, C \in \mathcal{K}_3^2\} \ .
\end{aligned}
$$

Theorem 5.2 *Consider testing* (17) *with* $\mathbf{P}_1$, $\mathbf{P}_2$ *and* $\mathbf{P}_3$ *given by* (25), (26), *and* (27), *respectively. The testing procedure that rejects* $H_{\hat{j}}$ *whenever* $T_{\hat{j},n}^2 > \hat{c}_{2,n}(1-\alpha)$, *where* $\hat{j} = \arg\max_{1 \leq j \leq 3} T_{j,n}^2$, *satisfies* (18).

Remark 5.3 By arguing as in Remark 5.1, it is in fact possible to show that $\mathcal{K}_2^*$ may be replaced by

$$\{\{1,\dots,4,11,12\}\} \cup \bigcup_{j \in \{6,7,11\}, k \in \{8,9,12\}} \{\{j,k\}\} \ . \blacksquare$$

5.3 Monotonicity of Y in D

In this section, we assume every $Q \in \mathbf{Q}$ satisfies Assumptions 2.1 and 2.3. For this choice of $\mathbf{Q}$, it follows from Theorem 3.3 that

$$
\begin{aligned}
\mathbf{P}_1 &= \{P \in \mathbf{P} : \min\{B_1(P), B_3(P)\} \leq \Delta(P) \leq \max\{B_2(P), B_4(P)\}\} & (28) \\
\mathbf{P}_2 &= \{P \in \mathbf{P} : B_2(P) \leq B_4(P) \cup \Delta(P) \leq B_4(P) \cup \Delta(P) > B_2(P), & (29) \\
&\qquad B_3(P) \leq B_1(P) \cup \Delta(P) < B_1(P) \cup \Delta(P) \geq B_3(P)\} & (30) \\
\mathbf{P}_3 &= \{P \in \mathbf{P} : B_2(P) \geq B_4(P) \cup \Delta(P) > B_4(P) \cup \Delta(P) \leq B_2(P), & (31) \\
&\qquad B_3(P) \geq B_1(P) \cup \Delta(P) \geq B_1(P) \cup \Delta(P) < B_3(P)\} \ . & (32)
\end{aligned}
$$

In order to describe our testing procedure, it is useful to introduce some further notation. Define

$$
\begin{aligned}
b_1(P) = -b_{19}(P) &= B_1^1(P) - \Delta(P) \\
b_2(P) = -b_{20}(P) &= B_1^2(P) - \Delta(P) \\
b_3(P) = -b_{31}(P) &= B_3^1(P) - \Delta(P) \\
b_4(P) = -b_{32}(P) &= B_3^2(P) - \Delta(P) \\
b_5(P) = -b_{13}(P) &= \Delta(P) - B_2^1(P) \\
b_6(P) = -b_{14}(P) &= \Delta(P) - B_2^2(P) \\
b_7(P) = -b_{25}(P) &= \Delta(P) - B_4^1(P) \\
b_8(P) = -b_{26}(P) &= \Delta(P) - B_4^2(P) \\
b_9(P) = -b_{21}(P) &= B_2^1(P) - B_4^1(P) \\
b_{10}(P) = -b_{23}(P) &= B_2^1(P) - B_4^2(P) \\
b_{11}(P) = -b_{22}(P) &= B_2^2(P) - B_4^1(P) \\
b_{12}(P) = -b_{24}(P) &= B_2^2(P) - B_4^2(P) \\
b_{15}(P) = -b_{27}(P) &= B_3^1(P) - B_1^1(P) \\
b_{16}(P) = -b_{29}(P) &= B_3^2(P) - B_1^1(P) \\
b_{17}(P) = -b_{28}(P) &= B_3^1(P) - B_1^2(P) \\
b_{18}(P) = -b_{30}(P) &= B_3^2(P) - B_1^2(P) \ .
\end{aligned}
$$

For $1 \leq j \leq 3$, define

$$
T_{j,n}^3 = \min_{K \in \mathcal{K}_j^3} \max_{k \in K} \frac{b_k(\hat{P}_n)}{\hat{\sigma}_{k,n}^b} \ ,
$$

where

$$
\begin{aligned}
\mathcal{K}_1^3 &= \{A \cup B : A \in \{\{1,2\},\{3,4\}\}, B \in \{\{5,6\},\{7,8\}\} \\
\mathcal{K}_2^3 &= \{A \cup B : A \in \{\{13\},\{14\},\{7,8\},\{9,10\},\{11,12\}\}, B \in \{\{19\},\{20\},\{3,4\},\{15,16\},\{17,18\}\}\} \\
\mathcal{K}_3^3 &= \{A \cup B : A \in \{\{25\},\{26\},\{5,6\},\{21,22\},\{23,24\}\}, B \in \{\{31\},\{32\},\{1,2\},\{27,28\},\{29,30\}\}\} \ ,
\end{aligned}
$$

and $\hat{\sigma}_{k,n}^b$ for $1 \leq k \leq 32$ is the usual (unpooled) estimate of the standard deviation of $b_k(\hat{P}_n)$. Define

$$
\hat{c}_{3,n}(1-\alpha) = \max_{K \in \mathcal{K}_3^*} J_{3,n}^{-1}(1-\alpha, K, \hat{P}_n) \ ,
$$

where

$$
\begin{aligned}
J_{3,n}(x, K, P) &= P \left\{ \max_{k \in K} \frac{b_k(\hat{P}_n) - b_k(P)}{\hat{\sigma}_{k,n}^b} \leq x \right\} \\
\mathcal{K}_3^* &= \{A \cup B \cup C : A \in \mathcal{K}_1^3, B \in \mathcal{K}_2^3, C \in \mathcal{K}_3^3\}
\end{aligned}
$$

Theorem 5.3 *Consider testing* (17) *with* $\mathbf{P}_1$, $\mathbf{P}_2$ *and* $\mathbf{P}_3$ *given by* (28), (29), *and* (31), *respectively. The testing procedure that rejects* $H_{\hat{j}}$ *whenever* $T_{\hat{j},n}^3 > \hat{c}_{3,n}(1-\alpha)$, *where* $\hat{j} = \arg\max_{1 \leq j \leq 3} T_{j,n}^3$, *satisfies* (18).

Remark 5.4 By arguing as in Remark 5.1, it is in fact possible to show that $\mathcal{K}_3^*$ may be replaced by

$$\{\{1,\ldots,8\}\} \cup \{A \cup B : A \in \mathcal{K}_2^3, B \in \mathcal{K}_3^3, C \nsubseteq A \cup B \text{ for all } C \in \tilde{\mathcal{K}}\}\ ,$$

where

$$\tilde{\mathcal{K}} = \{\{1,2,19\},\{3,4,31\},\{3,4,32\},\{7,8,25\},\{7,8,26\},\{5,6,13\},\{5,6,14\}\}\ . \blacksquare$$

A Detecting Failure of the Monotonicity Restrictions: Numerical Results

Below we provide numerical results to complement the analytical results from Section 4. We consider a parametric, latent variable model for Y and D that satisfies instrument exogeneity – Assumption 2.1 – and examine which parameterizations of the latent variable model result in detectable violations of the monotonicity restrictions. The numerical results highlight that the ability to detect violations of Assumption 2.2 requires strong (positive or negative) dependence between treatment response to the instrument and outcome response to the treatment, i.e., strong dependence between $D_1 - D_0$ and $Y_1 - Y_0$. The results also highlight the difficulty of detecting violations of Assumption 2.3 in that, in addition to extremely strong dependence between $D_1 - D_0$ and $Y_1 - Y_0$, the fraction of "always takers" and "never takers" must be close to zero in order to be able to detect violations of Assumption 2.3.

A.1 Monotonicity of D in Z

Consider the following parametric model for Y and D:

$$\begin{aligned} Y &= I\{\beta D + \epsilon \geq 0\} \\ D &= I\{\delta Z + \eta \geq 0\} \end{aligned} \tag{33}$$

with $Z \perp\!\!\!\perp (\epsilon, \eta, \beta, \delta)$, $(\epsilon, \eta, \beta, \delta) \sim N(\mu, \Sigma)$, and $E[\delta] > 0$. Note that this model satisfies Assumption 2.1 and that Assumption 2.2 is violated whenever $\mathrm{Var}[\delta] > 0$. $\mathrm{Corr}[\beta, \delta]$ measures the dependence between treatment response to the instrument and outcome response to the treatment. $\mathrm{Var}[\delta]$ and $E[\delta]$ measure the strength of the instrument, which is decreasing in $\mathrm{Var}[\delta]$ and increasing in $E[\delta]$. From Proposition 4.1, we have that the ability to detect violations of Assumption 2.2 is increasing in the size of the violation and decreasing in the strength of the instrument. In addition, the ability to detect violations requires sufficient difference in latent outcome distributions between "compliers" and "defiers." The difference between these distributions is increasing in $|\mathrm{Corr}[\beta, \delta]|$. We therefore examine below how the ability to detect violations of Assumption 2.2 varies with $\mathrm{Var}[\delta]$, $E[\delta]$, and $\mathrm{Corr}[\beta, \delta]$. In particular, we consider parameterizations of (33) with

$$\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \mu_\delta \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \sigma_{\beta,\delta} & \sigma_\delta^2 \end{pmatrix}$$

and vary μ_δ from 0.1 to 1, σ_δ^2 from 0.2 to 50, and $\sigma_{\beta,\delta}$ so that $\mathrm{Corr}[\beta, \delta]$ varies from -1 to 1.

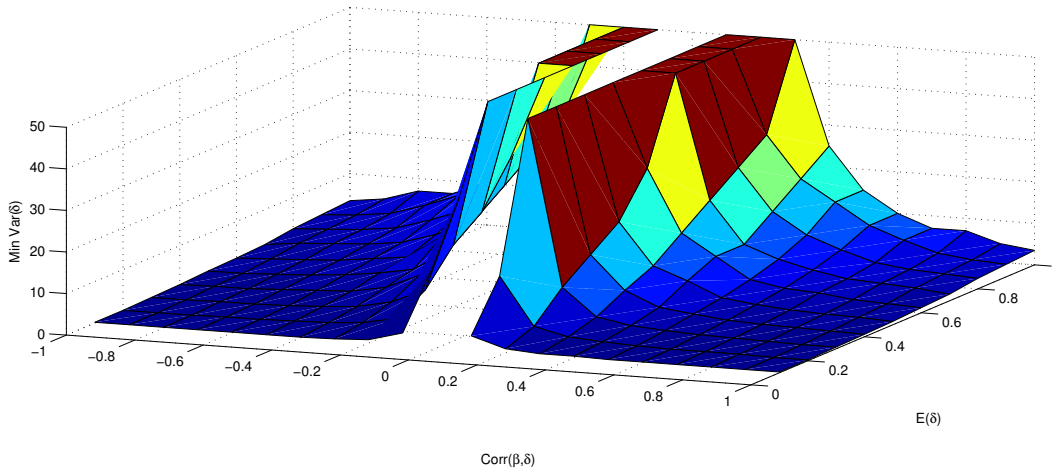Figure 1: Detecting Violations of D Monotonic in Z: Minimum Var[δ]

Figure 1 displays the minimum value of Var[δ] for which it is possible to detect violations of Assumption 2.2 for different values of E[δ], Corr[β, δ]. For presentation purposes, we have truncated the graph at 50 for the minimum value of Var(δ). The minimum value of Var[δ] for which it is possible to detect violations of Assumption 2.2 is increasing in E[δ], though not dramatically so. In contrast, the minimum value of Var[δ] for which it is possible to detect violations of Assumption 2.2 asymptotes to infinity as Corr[β, δ] approaches zero.

Figure 2 displays the maximum strength of the instrument, as indexed by $E[\delta]$, for which it is possible to detect violations of Assumption 2.2 for different values of Var[δ] and Corr[β, δ]. The maximum value of E[δ] for which it is possible to detect violations is increasing in Var[δ]: if the violation is more severe, then the instrument can be stronger with the violation still being detectable. As Corr[β, δ] approaches zero, the maximum value of E[δ] for which it is possible to detect violations approaches 0. For any Corr[β, δ] $\neq 0$, there is a strength of instrument sufficiently weak such that the violation of Assumption 2.2 can still be detected. On the other hand, if Corr[β, δ] $= 0$, then it is not possible to detect violation of Assumption 2.2 for any value of E[δ] and Var[δ].

The lefthand-side of Figure 3 displays the maximum value of Corr[β, δ] < 0 for which we can detect violations of Assumption 2.2 for different values of $E[\delta]$ and Var[δ]; the righthand-side of Figure 3 displays the minimum value of Corr[β, δ] > 0 for which we can detect violations of Assumption 2.2 for different values of $E[\delta]$ and Var[δ]. The figure is plotted from an "overhead" view, with warmer colors indicating higher values for the maximum/minumum value of Corr[β, δ] for which the violation is detectable and white space for values of $E[\delta]$ and Var[δ] for which there is no value of Corr[β, δ] for which the violation is detectable. The ability to detect the violation of Assumption 2.2 is increasing in $|\text{Corr}(\beta, \delta)|$, but, for a fairly large range of values of $E[\delta]$ and Var[δ], there exists no value of Corr[β, δ] for which the violation is detectable.

Figure 2: Detecting Violations of D Monotonic in Z: Maximum E[δ]
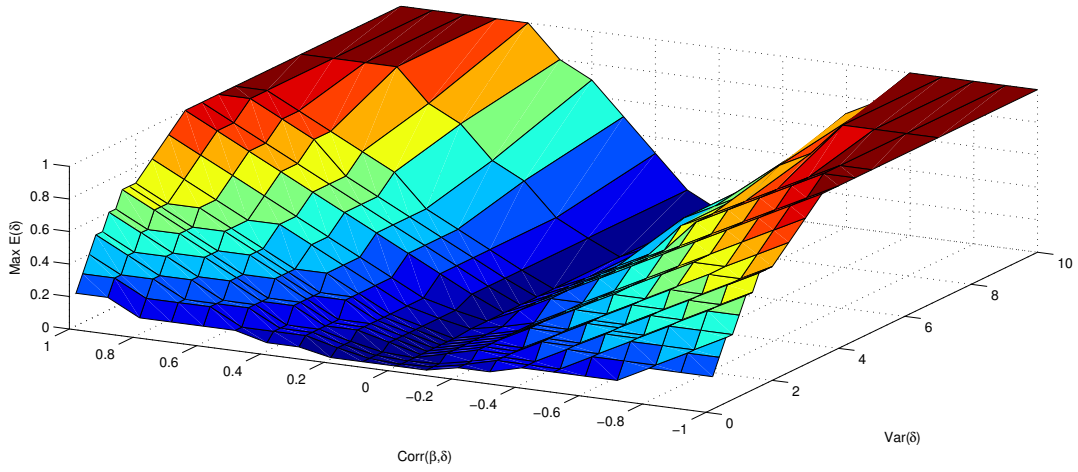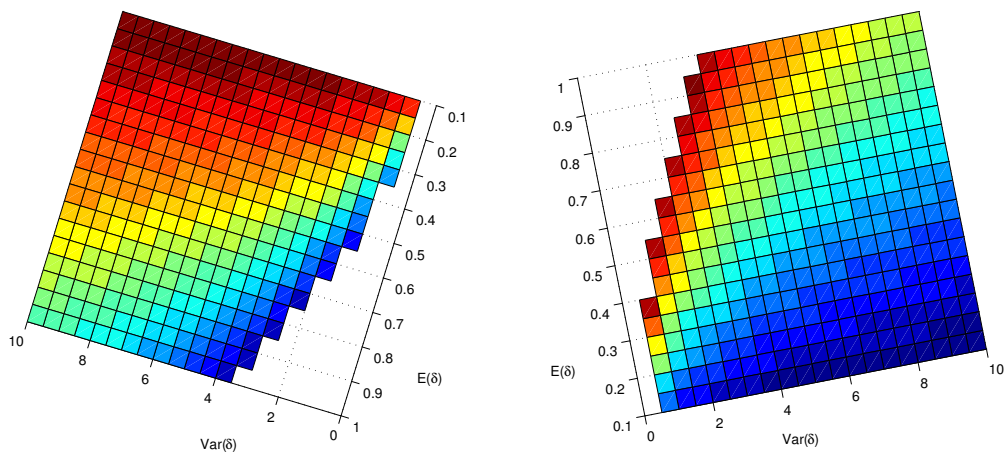

Figure 3: Detecting Violations of D Monotonic in Z: Minimum/Maximum Corr[β, δ]

A.2 Monotonicity of Y in D

Extensive experimentation revealed that it is difficult to find parameterizations of (33) for which it is possible to detect violations of Assumption 2.3. For example, with $\text{Corr}[\eta, \beta] = 0$, $\text{Corr}[\eta, \delta] = 0$ and $\text{Corr}[\beta, \delta] \approx \pm 1$, we were unable to find any parameterizations for which it is possible to detect violations Assumption 2.3. The only parameterizations we found for which it is possible to detect violations of Assumption 2.3 involved $\text{Corr}[\beta, \delta] \approx 1$, $\text{Corr}[\eta, \beta] \approx -1$, $\text{Corr}[\eta, \delta] \approx -1$, and both $\text{Var}[\beta]$ and $\text{Var}[\delta]$ large. This remained true even for extreme violations of Assumption 2.3, such as $\text{Var}[\beta] = 10,000$. The results suggest that in a model of the form of (33), it is difficult to find parameterizations such that the fractions of "always takers" and "never takers" are small enough so that it is possible to detect violations of Assumption 2.3.

The extreme difficulty in finding parameterizations of (33) for which it is possible to detect violations of Assumption 2.3 led us to consider the following parametric, latent variable model for Y and D:

$$
\begin{aligned}
Y &= I\{\beta D + \epsilon \geq 0\} \\
D &= I\{\alpha_t(\delta) + \delta_t(\delta)Z + \eta \geq 0\}
\end{aligned}
\tag{34}
$$

with $Z \perp\!\!\!\perp (\epsilon, \eta, \beta, \delta)$, $(\epsilon, \eta, \beta, \delta) \sim N(\mu, \Sigma)$, and

$$
\alpha_t(\delta) = \begin{cases} -t & \text{if } \delta > 0 \\ t & \text{if } \delta \leq 0 \end{cases}
$$

$$
\delta_t(\delta) = \begin{cases} \delta + 2t & \text{if } \delta > 0 \\ \delta - 2t & \text{if } \delta \leq 0 \end{cases}.
$$

Here, the parameter $t > 0$ is used as an index to control the fractions of "always takers" and "never takers." In particular, these fractions are decreasing in t. As in the analysis of the previous subsection, we consider

$$
\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \\ .1 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \sigma_\beta^2 & 0 \\ 0 & 0 & \sigma_{\beta,\delta} & 10 \end{pmatrix},
$$

Extensive experimentation again revealed that it is difficult to find parameterizations of (34) for which it is possible to detect violations of Assumption 2.3 for small values of t, though less difficult as t gets larger (and thus the probability of being an "always taker" or "never taker" approaches zero). For example, when $t \geq 4$, $Q\{D_1 = D_0\} \approx 0$, and, for such t, it is possible to detect violations of Assumption 2.3 if $\text{Corr}[\beta, \delta]$ and $\text{Var}[\beta^*]$ are sufficiently large, such as $\text{Corr}[\beta, \delta] = .8$ and $\text{Var}[\delta] \geq 3$.

B Proofs for Section 3

PROOF OF THEOREM 3.1: First consider assertion (i). For $1 \leq j \leq 2$, $\Delta(P) = A_1^j(P) + A_4^j(P)$, so that $\Delta(P) \geq A_1^j(P) \implies A_4^j(P) \geq 0$ and $\Delta(P) \leq A_4^j(P) \implies A_1^j(P) \leq 0$. Thus, $\Delta(P) \in [A_1(P), A_4(P)]$ if and only if $A_1(P) \leq 0$ and $A_4(P) \geq 0$. The result then follows from Balke and Pearl (1997). Now

consider assertions (ii) and (iii). From Balke and Pearl (1997), the identified set for $E_Q[Y_1 - Y_0]$ is given by $[\Delta(P) - A_3(P), \Delta(P) - A_2(P)]$. Combining this result with (i) gives the stated results. ∎

PROOF OF THEOREM 3.2: Following the same linear programming strategy as in Balke and Pearl (1997) except with the additional constraints on **Q** that

$$Q\{Y_1 > Y_0, D_1 = j, D_0 = k\} = 0$$

for all $(j, k) \in \{0, 1\}^2$ and $Q \in \mathbf{Q}$ results in the restriction that $\Delta(P) \in [A_1(P), 0]$. Instead imposing the additional constraints on **Q** that

$$Q\{Y_1 < Y_0, D_1 = j, D_0 = k\} = 0$$

for all $(j, k) \in \{0, 1\}^2$ and $Q \in \mathbf{Q}$ results in the restriction $\Delta(P) \in [0, A_4(P)]$. The result now follows. ∎

PROOF OF THEOREM 3.3: Following the same linear programming strategy as in Balke and Pearl (1997) except replacing the constraint that

$$Q\{Y_1 = j, Y_0 = k, D_1 < D_0\} = 0 \tag{35}$$

for all $(j, k) \in \{0, 1\}^2$ and $Q \in \mathbf{Q}$ with the constraint that

$$Q\{Y_1 > Y_0, D_1 = j, D_0 = k\} = 0$$

for all $(j, k) \in \{0, 1\}^2$ and $Q \in \mathbf{Q}$ results in the restriction that $\Delta(P) \in [B_3(P), B_4(P)]$. Instead replacing the constraint that (35) holds for all $(j, k) \in \{0, 1\}^2$ and $Q \in \mathbf{Q}$ with the constraint that

$$Q\{Y_1 < Y_0, D_1 = j, D_0 = k\} = 0$$

for all $(j, k) \in \{0, 1\}^2$ and $Q \in \mathbf{Q}$ results in the restriction that $\Delta(P) \in [B_1(P), B_2(P)]$. The result now follows. ∎

C Proofs for Section 4

PROOF OF PROPOSITION 4.1: Using (2) - (3) and Assumption 2.1, $\Delta(P)$ may be expressed as

$$\left\{ Q\{Y_1 > Y_0, D_1 > D_0\} - Q\{Y_1 < Y_0, D_1 > D_0\} \right\}$$
$$- \left\{ Q\{Y_1 > Y_0, D_1 < D_0\} - Q\{Y_1 < Y_0, D_1 < D_0\} \right\}. \tag{36}$$

Furthermore,

$$
\begin{aligned}
A_1^1(P) &= Q\{Y_0 = 1, D_1 < D_0\} - Q\{Y_0 = 1, D_1 > D_0\} \\
A_1^2(P) &= Q\{Y_1 = 0, D_1 < D_0\} - Q\{Y_1 = 0, D_1 > D_0\} \\
A_4^1(P) &= Q\{Y_1 = 1, D_1 > D_0\} - Q\{Y_1 = 1, D_1 < D_0\} \\
A_4^2(P) &= Q\{Y_0 = 0, D_1 > D_0\} - Q\{Y_0 = 0, D_1 < D_0\}.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\Delta(P) - A_1^1(P) &= Q\{Y_1 = 1, D_1 > D_0\} - Q\{Y_1 = 1, D_1 < D_0\} \\
\Delta(P) - A_1^2(P) &= Q\{Y_0 = 0, D_1 > D_0\} - Q\{Y_0 = 0, D_1 < D_0\} \\
\Delta(P) - A_4^1(P) &= -Q(Y_0 = 1, D_1 > D_0\} + Q\{Y_0 = 1, D_1 < D_0\} \\
\Delta(P) - A_4^2(P) &= -Q\{Y_1 = 0, D_1 > D_0\} + Q\{Y_1 = 0, D_1 < D_0\} .
\end{aligned}
$$

The desired result now follows immediately. ∎

PROOF OF PROPOSITION 4.2: Using Assumption 2.1, we have that

$$
B_1(P) = -\min\left\{ \begin{array}{l} Q\{Y_1 > Y_0, D_1 < D_0\} + Q\{Y_1 > Y_0, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 1, D_0 = 1\}, \\ Q\{Y_1 > Y_0, D_1 < D_0\} + Q\{Y_1 > Y_0, D_1 = D_0 = 0\} + Q\{Y_1 = Y_0 = 0, D_1 = 0\} \end{array} \right\},
$$

$$
B_2(P) = \min\left\{ \begin{array}{l} Q\{Y_1 > Y_0, D_1 > D_0\} + Q\{Y_1 > Y_0, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 1, D_1 = 1\}, \\ Q\{Y_1 > Y_0, D_1 > D_0\} + Q\{Y_1 > Y_0, D_1 = D_0 = 0\} + Q\{Y_0 = Y_1 = 0, D_0 = 0\} \end{array} \right\},
$$

$$
B_3(P) = -\min\left\{ \begin{array}{l} Q\{Y_1 < Y_0, D_1 > D_0\} + Q\{Y_1 < Y_0, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 0, D_1 = 1\} \\ Q\{Y_1 < Y_0, D_1 > D_0\} + Q\{Y_1 < Y_0, D_1 = D_0 = 0\} + Q\{Y_0 = Y_1 = 1, D_0 = 0\} \end{array} \right\},
$$

$$
B_4(P) = \min\left\{ \begin{array}{l} Q\{Y_1 < Y_0, D_1 < D_0\} + Q\{Y_1 < Y_0, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 0, D_0 = 1\} \\ Q\{Y_1 < Y_0, D_1 < D_0\} + Q\{Y_1 < Y_0, D_1 = D_0 = 0\} + Q\{Y_1 = Y_0 = 1, D_1 = 0\} \end{array} \right\}.
$$

so that

$$
\begin{aligned}
B_1(P) &= -Q\{Y_1 > Y_0, D_1 < D_0\} - M_1 \\
B_2(P) &= Q\{Y_1 > Y_0, D_1 > D_0\} + M_2 \\
B_3(P) &= -Q\{Y_1 < Y_0, D_1 > D_0\} - M_3 \\
B_4(P) &= Q\{Y_1 < Y_0, D_1 < D_0\} + M_4
\end{aligned}
$$

The desired result now follows immediately. ∎

D   Proofs for Section 6

PROOF OF THEOREM 5.1: Suppose by way of contradiction that (18) fails. Then there exists a subsequence $\{P_{n_m} \in \mathbf{P} : m \geq 1\}$ and $\alpha' > \alpha$ such that

$$
FWER_{P_{n_m}} \to \alpha' . \tag{37}
$$

Let

$$
I(P) = \{1 \leq j \leq 3 : P \in \mathbf{P}_j\} \subseteq \{1, 2, 3\} .
$$

Since there are only finitely many possible values for $I(P)$ and $FWER_P = 0$ when $I(P) = \emptyset$, we may assume further that $I(P_{n_m}) = I \neq \emptyset$. Let

$$\mathcal{K}_I^* = \left\{ \bigcup_{j \in I} A_j : A_j \in \mathcal{K}_j^1 \text{ for } j \in I \right\} .$$

Using this notation, we have that

$$\{ P \in \mathbf{P} : I(P) = I \} \subseteq \bigcup_{\tilde{K} \in \mathcal{K}_I^*} \mathbf{P}_{\tilde{K}} , \tag{38}$$

where

$$\mathbf{P}_{\tilde{K}} = \bigcap_{k \in \tilde{K}} \{ P \in \mathbf{P} : a_k(P) \leq 0 \} .$$

Note that in (38) we have an inclusion instead of an equality due to the fact that $\mathbf{P}_{\tilde{K}}$ is always defined with weak inequalities, whereas the null hypotheses are sometimes defined with strict inequalities. Since there are only finitely many $\tilde{K}$ in $\mathcal{K}_I^*$, we may assume further that there is $K \in \mathcal{K}_I^*$ such that

$$\{ P_{n_m} \in \mathbf{P} : m \geq 1 \} \subseteq \mathbf{P}_K . \tag{39}$$

Next, note that

$$
\begin{aligned}
\max_{j \in I} T_{j,n}^1 &= \max_{j \in I} \min_{\tilde{K} \in \mathcal{K}_j^1} \max_{k \in \tilde{K}} \frac{a_k(\hat{P}_n)}{\hat{\sigma}_{k,n}^a} \\
&\leq \max_{k \in K} \frac{a_k(\hat{P}_n)}{\hat{\sigma}_{j,n}^a} .
\end{aligned}
\tag{40}
$$

To establish (40), let $j \in I$ and $K^* \in \mathcal{K}_I^*$ be given. It follows that there exists $\tilde{K} \in \mathcal{K}_j^1$ such that $\tilde{K} \subseteq K^*$. Hence,

$$\max_{k \in \tilde{K}} \frac{a_k(\hat{P}_n)}{\hat{\sigma}_{k,n}^a} \leq \max_{k \in K^*} \frac{a_k(\hat{P}_n)}{\hat{\sigma}_{k,n}^a} ,$$

which in turn implies that

$$\min_{\tilde{K} \in \mathcal{K}_j^1} \max_{k \in \tilde{K}} \frac{a_k(\hat{P}_n)}{\hat{\sigma}_{k,n}^a} \leq \max_{k \in K^*} \frac{a_k(\hat{P}_n)}{\hat{\sigma}_{k,n}^a} .$$

Since the inequality is valid for any $j \in I$ and $K^* \in \mathcal{K}_I^*$,

$$\max_{j \in I} \min_{\tilde{K} \in \mathcal{K}_j^1} \max_{k \in \tilde{K}} \frac{a_k(\hat{P}_n)}{\hat{\sigma}_{k,n}^a} \leq \min_{K^* \in \mathcal{K}_I^*} \max_{k \in K^*} \frac{a_k(\hat{P}_n)}{\hat{\sigma}_{k,n}^a} ,$$

from which (40) follows because $K \in \mathcal{K}_I^*$. Similarly,

$$
\begin{aligned}
\hat{c}_{1,n}(1-\alpha) &= \max_{\tilde{K} \in \mathcal{K}_1^*} J_{1,n}^{-1}(1-\alpha, \tilde{K}, \hat{P}_n) \\
&\geq J_{1,n}^{-1}(1-\alpha, K, \hat{P}_n) .
\end{aligned}
\tag{41}
$$

To establish (41), note that there exists $K^* \in \mathcal{K}_1^*$ such that $K \subseteq K^*$, so

$$J_{1,n}^{-1}(1 - \alpha, K^*, \hat{P}_n) \geq J_{1,n}^{-1}(1 - \alpha, K, \hat{P}_n) \ .$$

The inequalities (40) and (41) therefore imply that

$$
\begin{aligned}
FWER_{P_{n_m}} &= P_{n_m} \left\{ \max_{j \in I} T_{j,n_m}^1 > \hat{c}_{1,n_m}(1 - \alpha) \right\} \\
&\leq P_{n_m} \left\{ \max_{k \in K} \frac{a_k(\hat{P}_{n_m})}{\hat{\sigma}_{j,n_m}^a} > J_{1,n_m}^{-1}(1 - \alpha, K, \hat{P}_{n_m}) \right\} \ .
\end{aligned}
\tag{42}
$$

Using (39) and Theorem E.1, we see that the righthand-side of (42) tends to α, contradicting (37), and thereby establishing the desired result. ■

PROOF OF THEOREM 5.2: The proof is essentially identical to the one presented for Theorem 5.1 and is therefore omitted. ■

PROOF OF THEOREM 5.3: The proof is essentially identical to the one presented for Theorem 5.1 and is therefore omitted. ■

E Auxiliary Results

In this appendix, we establish the following result:

Theorem E.1 *Let* $(X_i, Y_i, Z_i), i = 1, \ldots, n$ *be an i.i.d. sequence of random variables with distribution* $P \in \mathbf{P}$ *on* $\mathbf{R}^k \times \mathbf{R}^k \times \{0,1\}$. *Suppose* $\mathbf{P}$ *is such that*

$$\epsilon < \inf_{P \in \mathbf{P}} P\{Z = 1\} \leq \sup_{P \in \mathbf{P}} P\{Z = 1\} < 1 - \epsilon \tag{43}$$

for some $\epsilon > 0$, *and for each* $1 \leq j \leq k$ *that*

$$\limsup_{\lambda \to \infty} \sup_{P \in \mathbf{P}} E_P \left[\left(\frac{X_j - \mu_{X_j|Z=1}(P)}{\sigma_{X_j|Z=1}(P)} \right)^2 I\left\{ \left| \frac{X_j - \mu_{X_j|Z=1}(P)}{\sigma_{X_j|Z=1}(P)} \right| > \lambda \right\} \bigg| Z = 1 \right] = 0 \tag{44}$$

and

$$\limsup_{\lambda \to \infty} \sup_{P \in \mathbf{P}} E_P \left[\left(\frac{Y_j - \mu_{Y_j|Z=1}(P)}{\sigma_{Y_j|Z=1}(P)} \right)^2 I\left\{ \left| \frac{Y_j - \mu_{Y_j|Z=1}(P)}{\sigma_{Y_j|Z=1}(P)} \right| > \lambda \right\} \bigg| Z = 0 \right] = 0 \ . \tag{45}$$

Let

$$J_n(x, P) = P\left\{ \max_{1 \leq j \leq k} T_{n,j}(P) \leq x \right\} \ , \tag{46}$$

where

$$T_{n,j}(P) = \frac{\frac{1}{n_1} \sum_{1 \leq i \leq n: Z_i=1} X_{j,i} - \mu_{X_j|Z=1}(P) - \frac{1}{n_0} \sum_{1 \leq i \leq n: Z_i=0} Y_{j,i} - \mu_{Y_j|Z=0}(P)}{\sqrt{\frac{\sigma_{X_j|Z=1}^2(\hat{P}_n)}{n_1} + \frac{\sigma_{Y_j|Z=0}^2(\hat{P}_n)}{n_0}}} \ .$$

Then

$$\limsup_{n \to \infty} \sup_{P \in \mathbf{P}} P\left\{ \max_{1 \leq j \leq k} T_{n,j}(P) > J_n^{-1}(1 - \alpha, \hat{P}_n) \right\} \leq \alpha \ .$$

Before presenting the proof of Theorem E.1, we present a series of useful lemmata.

Lemma E.1 *Let $(X_i, Z_i), i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution $P \in \mathbf{P}$ on $\mathbf{R} \times \{0, 1\}$. Suppose $\mathbf{P}$ is such that*

$$\inf_{P \in \mathbf{P}} P\{Z = 1\} > \epsilon$$

for some $\epsilon > 0$ and that

$$\limsup_{\lambda \to \infty} \sup_{P \in \mathbf{P}} E_P \left[\left| X - \mu_{X|Z=1}(P) \right| I \left\{ \left| X - \mu_{X|Z=1}(P) \right| > \lambda \right\} | Z = 1 \right] = 0 \ . \tag{47}$$

Then, for any $\{P_n \in \mathbf{P} : n \geq 1\}$,

$$\frac{1}{n_1} \sum_{1 \leq i \leq n : Z_i = 1} X_i - \mu_{X|Z=1}(P_n) \overset{P_n}{\to} 0 \ ,$$

where $n_1 = \sum_{1 \leq i \leq n} Z_i$.

PROOF: First assume w.l.o.g. that $\mu_{X|Z=1}(P_n) = 0$. Thus, $E_{P_n}[ZX] = 0$. Next, note that (47) implies that

$$\limsup_{\lambda \to \infty} \frac{1}{P_n\{Z = 1\}} E_{P_n} \left[|ZX| I \{|ZX| > \lambda\} | Z = 1 \right] = 0 \ .$$

Since $P_n\{Z = 1\} > \epsilon$, it follows that

$$\limsup_{\lambda \to \infty} E_{P_n} \left[|ZX| I \{|ZX| > \lambda\} | Z = 1 \right] = 0 \ .$$

By Lemma 11.4.2 of Romano and Shaikh (2012), we therefore have that

$$\frac{1}{n} \sum_{1 \leq i \leq n} X_i Z_i \overset{P_n}{\to} 0 \ .$$

Since $|Z - \mu_Z(P_n)| \leq 1$, we also have that

$$\limsup_{\lambda \to \infty} E_{P_n} \left[|Z - \mu_Z(P_n)| I \{|Z - \mu_Z(P_n)| > \lambda\} | Z = 1 \right] = 0 \ .$$

Thus,

$$\frac{1}{n} \sum_{1 \leq i \leq n} Z_i = P_n\{Z = 1\} + o_{P_n}(1) \ .$$

To complete the argument, note that

$$\frac{1}{n_1} \sum_{1 \leq i \leq n : Z_i = 1} X_i = \left(\frac{1}{n} \sum_{1 \leq i \leq n} X_i Z_i \right) \Big/ \left(\frac{1}{n} \sum_{1 \leq i \leq n} Z_i \right) \ .$$

The desired result now follows since $P_n\{Z = 1\} > \epsilon$. ∎

Lemma E.2 *Let $(X_i, Y_i, Z_i), i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution $P \in \mathbf{P}$ on $\mathbf{R}^k \times \mathbf{R}^k \times \{0, 1\}$. Suppose (43) holds for some $\epsilon > 0$ and for all $1 \leq j \leq k$ that (44) and (45) hold.*

Then, for any $\{P_n \in \mathbf{P} : n \geq 1\}$,

$$||\Omega_{X|Z=1}(\hat{P}_n) - \Omega_{X|Z=1}(P_n)|| \overset{P_n}{\to} 0 \tag{48}$$

$$||\Omega_{Y|Z=0}(\hat{P}_n) - \Omega_{Y|Z=0}(P_n)|| \overset{P_n}{\to} 0 , \tag{49}$$

where $||\cdot||$ denotes the component-wise maximum of the absolute value of all elements.

PROOF: We provide only the proof for (48), as the same argument establishes (49). To establish (48), first note that we may assume w.l.o.g. for all $1 \leq j \leq k$ that $\mu_{X_j|Z=1}(P_n) = 0$ and $\sigma_{X_j|Z=1}(P_n) = 1$. The (j, ℓ) element of $\Omega_{X|Z=1}(P_n)$ is thus given by

$$E_{P_n}[X_j X_\ell | Z = 1]$$

and the (j, ℓ) element of $\Omega_{X|Z=1}(\hat{P}_n)$ is given by

$$\frac{\frac{1}{n_1} \sum_{1 \leq i \leq n : Z_i = 1} X_{i,j} X_{i,\ell} - \left(\frac{1}{n_1} \sum_{1 \leq i \leq n : Z_i = 1} X_{i,j} \right) \left(\frac{1}{n_1} \sum_{1 \leq i \leq n : Z_i = 1} X_{i,\ell} \right)}{\sigma_{X_j|Z=1}(\hat{P}_n) \sigma_{X_\ell|Z=1}(\hat{P}_n)} ,$$

where $n_1 = \sum_{1 \leq i \leq n} Z_i$. From Lemma B.3 in Bhattacharya et al. (2012), we see that

$$\sigma_{X_j|Z=1}(\hat{P}_n) \overset{P_n}{\to} 1$$

$$\sigma_{X_\ell|Z=1}(\hat{P}_n) \overset{P_n}{\to} 1 .$$

From Lemma E.1, we see that

$$\frac{1}{n_1} \sum_{1 \leq i \leq n : Z_i = 1} X_{i,j} \overset{P_n}{\to} 0$$

$$\frac{1}{n_1} \sum_{1 \leq i \leq n : Z_i = 1} X_{i,\ell} \overset{P_n}{\to} 0 .$$

Using the inequality

$$|a||b|I\{|a||b| > \lambda\} \leq a^2 I\{|a| > \sqrt{\lambda}\} + b^2 I\{|b| > \sqrt{\lambda}\} ,$$

we see that

$$\limsup_{\lambda \to \infty} E_{P_n}\left[|X_j X_\ell| I\{|X_j X_\ell| > \lambda\} | Z = 1 \right] = 0 .$$

Since $|E_{P_n}[X_j X_\ell | Z = 1]| \leq 1$ by the Cauchy-Schwartz inequality, we have further that

$$\limsup_{\lambda \to \infty} E_{P_n}\left[|X_j X_\ell - E_{P_n}[X_j X_\ell | Z = 1]| I\{|X_j X_\ell - E_{P_n}[X_j X_\ell | Z = 1]| > \lambda\} | Z = 1 \right] = 0 .$$

Thus, Lemma E.1 implies that

$$\frac{1}{n_1} \sum_{1 \leq i \leq n : Z_i = 1} X_{i,j} X_{i,\ell} = E_{P_n}[X_j X_\ell | Z = 1] + o_{P_n}(1) .$$

The desired result now follows immediately. ∎

Lemma E.3 *Let $(X_i, Y_i, Z_i), i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution $P \in \mathbf{P}$*

on $\mathbf{R}^k \times \mathbf{R}^k \times \{0,1\}$. *Suppose* (43) *holds for some* $\epsilon > 0$ *and for all* $1 \le j \le k$ *that* (44) *and* (45) *hold.*
Define

$$D(P) = diag\left(\frac{\frac{\sigma^2_{X_1|Z=1}(P)}{P\{Z=1\}}}{\frac{\sigma^2_{X_1|Z=1}(P)}{P\{Z=1\}} + \frac{\sigma^2_{Y_1|Z=0}(P)}{P\{Z=0\}}}, \ldots, \frac{\frac{\sigma^2_{X_k|Z=1}(P)}{P\{Z=1\}}}{\frac{\sigma^2_{X_k|Z=1}(P)}{P\{Z=1\}} + \frac{\sigma^2_{Y_k|Z=0}(P)}{P\{Z=0\}}}\right) .$$

Then,

$$||D(\hat{P}_n)\Omega_{X|Z=1}(\hat{P}_n) - D(P_n)\Omega_{X|Z=1}(P_n)|| \overset{P_n}{\to} 0 \tag{50}$$

$$||(I - D(\hat{P}_n))\Omega_{Y|Z=0}(\hat{P}_n) - (I - D(P_n))\Omega_{Y|Z=0}(P_n)|| \overset{P_n}{\to} 0 , \tag{51}$$

where I is the k-dimensional identity matrix and $||\cdot||$ denotes the component-wise maximum of the absolute value of all elements. Hence,

$$||V(\hat{P}_n) - V(P_n)|| \overset{P_n}{\to} 0 , \tag{52}$$

where

$$V(P) = D(P)\Omega_{X|Z=1}(P) + (I - D(P))\Omega_{Y|Z=0}(P) . \tag{53}$$

PROOF: We provide only the proof for (50); the same argument establishes (51) and (52) then follows immediately from the triangle inequality. To establish (50), first note that $D(P_n)$ is invertible and that from Lemma B.4 of Bhattacharya et al. (2012)

$$||D(P_n)^{-1}D(\hat{P}_n) - I|| \overset{P_n}{\to} 0 .$$

Next, note for a universal constant C that

$$||D(\hat{P}_n)\Omega_{X|Z=1}(\hat{P}_n) - D(P_n)\Omega_{X|Z=1}(P_n)||$$
$$\le C||D(P_n)||\,||D(P_n)^{-1}D(\hat{P}_n)\Omega_{X|Z=1}(\hat{P}_n) - \Omega_{X|Z=1}(P_n)||$$
$$\le C^2||D(P_n)||\,||\Omega_{X|Z=1}(\hat{P}_n)||\left(||D(P_n)^{-1}D(\hat{P}_n) - I|| + ||\Omega_{X|Z=1}(\hat{P}_n) - \Omega_{X|Z=1}(P_n)||\right)$$

Since the elements of $D(P_n)$ and $\Omega_{X|Z=1}(\hat{P}_n)$ are all bounded, the norm of these matrices are also bounded. It therefore suffices to show that

$$||\Omega_{X|Z=1}(\hat{P}_n) - \Omega_{X|Z=1}(P_n)|| \overset{P_n}{\to} 0 ,$$

which follows from Lemma E.2. ∎

Lemma E.4 *Let $(X_i, Y_i, Z_i), i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution $P \in \mathbf{P}$ on $\mathbf{R}^k \times \mathbf{R}^k \times \{0,1\}$. Suppose* (43) *holds for some $\epsilon > 0$ and for all $1 \le j \le k$ that* (44) *and* (45) *hold. Then, for any $\{P_n \in \mathbf{P} : n \ge 1\}$,*

$$\max_{1 \le j \le k} \left\{\int_0^\infty |r_j(\lambda, \hat{P}_n) - r_j(\lambda, P)|d\lambda\right\} \overset{P_n}{\to} 0 \tag{54}$$

$$\max_{1 \le j \le k} \left\{\int_0^\infty |s_j(\lambda, \hat{P}_n) - s_j(\lambda, P)|d\lambda\right\} \overset{P_n}{\to} 0 , \tag{55}$$

where

$$r_j(\lambda, P) = E_P\left[\left(\frac{X_j - \mu_{X_j|Z=1}(P)}{\sigma_{X_j|Z=1}(P)}\right)^2 I\left\{\frac{X_j - \mu_{X_j|Z=1}(P)}{\sigma_{X_j|Z=1}(P)} > \lambda\right\}\bigg|Z=1\right] \tag{56}$$

$$s_j(\lambda, P) = E_P\left[\left(\frac{Y_j - \mu_{Y_j|Z=0}(P)}{\sigma_{Y_j|Z=0}(P)}\right)^2 I\left\{\frac{Y_j - \mu_{Y_j|Z=0}(P)}{\sigma_{Y_j|Z=0}(P)} > \lambda\right\}\bigg|Z=0\right] . \tag{57}$$

PROOF: We provide only the proof for (54); the same argument establishes (55). To establish (54), consider any $1 \leq j \leq k$. First note that we may assume w.l.o.g. that $\mu_{X_j|Z=1}(P_n) = 0$ and $\sigma_{X_j|Z=1}(P_n) = 1$. Next, note for any $1 \leq j \leq k$ that $r_j(\lambda, \hat{P}_n) = A_n - 2B_n + C_n$, where

$$A_n = \frac{1}{\sigma_{X_j|Z=1}(\hat{P}_n)}\frac{1}{n_1}\sum_{1 \leq i \leq n: Z_i=1} X_{i,j}^2 I\{|X_{i,j} - \mu_{X_j|Z=1}(\hat{P}_n)| > \lambda\sigma_{X_j|Z=1}(\hat{P}_n)\}$$

$$B_n = \frac{\mu_{X_j|Z=1}(\hat{P}_n)}{\sigma_{X_j|Z=1}(\hat{P}_n)}\frac{1}{n_1}\sum_{1 \leq i \leq n: Z_i=1} X_{i,j} I\{|X_{i,j} - \mu_{X_j|Z=1}(\hat{P}_n)| > \lambda\sigma_{X_j|Z=1}(\hat{P}_n)\}$$

$$C_n = \frac{\mu_{X_j|Z=1}(\hat{P}_n)^2}{\sigma_{X_j|Z=1}(\hat{P}_n)}\frac{1}{n_1}\sum_{1 \leq i \leq n: Z_i=1} I\{|X_{i,j} - \mu_{X_j|Z=1}(\hat{P}_n)| > \lambda\sigma_{X_j|Z=1}(\hat{P}_n)\} .$$

From Lemma E.1, we see that $\mu_{X_j|Z=1}(\hat{P}_n) \xrightarrow{P_n} 0$. From Lemma B.3 in Bhattacharya et al. (2012), we see that $\sigma_{X_j|Z=1}(\hat{P}_n) \xrightarrow{P_n} 1$. From Lemma E.1, we also see that

$$\frac{1}{n_1}\sum_{1 \leq i \leq n: Z_i=1} |X_{i,j}| = E_{P_n}[|X_j|] + o_{P_n}(1) .$$

Since $E_{P_n}[|X_j|] \leq 1$ by the Cauchy-Schwartz inequality, it follows that $B_n = o_{P_n}(1)$ uniformly in λ. A similar argument establishes that $C_n = o_{P_n}(1)$ uniformly in λ. In summary,

$$r_j(\lambda, P_n) = \frac{1}{\sigma_{X_j|Z=1}(\hat{P}_n)}\frac{1}{n_1}\sum_{1 \leq i \leq n: Z_i=1} X_{i,j}^2 I\{|X_{i,j} - \mu_{X_j|Z=1}(\hat{P}_n)| > \lambda\sigma_{X_j|Z=1}(\hat{P}_n)\} + \Delta_n$$

uniformly in λ, where $\Delta_n = o_{P_n}(1)$.

For $\delta > 0$, define the events

$$E_n(\delta) = \{|\mu_{X_j|Z=1}(\hat{P}_n)| < \delta \cap 1 - \delta < \sigma_{X_j|Z=1}(\hat{P}_n) < 1 + \delta\}$$

$$E_n'(\delta) = \left\{\sup_{t \in \mathbf{R}}\left|\frac{1}{n_1}\sum_{1 \leq i \leq n: Z_i=1} X_{i,j}^2 I\{|X_{i,j}| > t\} - E_{P_n}[X_j^2 I\{|X_j| > t\}|Z=1]\right| < \delta\right\}$$

$$E_n''(\delta) = \{|\Delta_n| < \delta\} .$$

We now argue that $P_n\{E_n(\delta) \cap E_n'(\delta) \cap E_n''(\delta)\} \to 1$. Since $\mu_{X_j|Z=1}(\hat{P}_n) \xrightarrow{P_n} 0$, $\sigma_{X_j|Z=1}(\hat{P}_n) \xrightarrow{P_n} 1$, and $\Delta_n = o_{P_n}(1)$, it suffices to argue that

$$P_n\{E_n'(\delta)\} \to 1 . \tag{58}$$

To see this, note that

$$\frac{1}{n_1} \sum_{1 \le i \le n: Z_i=1} X_{i,j}^2 I\{|X_{i,j}| > t\} - E_{P_n}[X_j^2 I\{|X_j| > t\}|Z = 1]$$

$$= \frac{\frac{1}{P_n\{Z=1\}} \frac{1}{n} \sum_{1 \le i \le n} Z_i X_{i,j}^2 I\{|X_{i,j}| > t\}}{\frac{\bar{Z}_n}{P_n\{Z=1\}}} - \frac{1}{P_n\{Z=1\}} E_{P_n}[ZX_j^2 I\{|X_j| > t\}]$$

$$= \left(1 - \frac{P_n\{Z=1\}}{\bar{Z}_n}\right) \frac{1}{P_n\{Z=1\}} \frac{1}{n} \sum_{1 \le i \le n} Z_i X_{i,j}^2 I\{|X_{i,j}| > t\}$$

$$\frac{1}{P_n\{Z=1\}} \left(\frac{1}{n} \sum_{1 \le i \le n} Z_i X_{i,j}^2 I\{|X_{i,j}| > t\} - E_{P_n}[ZX_j^2 I\{|X_j| > t\}]\right)$$

$$= \left(1 - \frac{P_n\{Z=1\}}{\bar{Z}_n}\right) \frac{\bar{Z}_n}{P_n\{Z=1\}} \frac{1}{n_1} \sum_{1 \le i \le n: Z_i=1} X_{i,j}^2 I\{|X_{i,j}| > t\} \tag{59}$$

$$\frac{1}{P_n\{Z=1\}} \left(\frac{1}{n} \sum_{1 \le i \le n} Z_i X_{i,j}^2 I\{|X_{i,j}| > t\} - E_{P_n}[ZX_j^2 I\{|X_j| > t\}]\right) . \tag{60}$$

From Lemma E.1, we see that

$$\frac{\bar{Z}_n}{P_n\{Z=1\}} \xrightarrow{P_n} 1$$

and

$$\frac{1}{n_1} \sum_{1 \le i \le n: Z_i=1} X_{i,j}^2 I\{|X_{i,j}| > t\} = E_{P_n}[X_j^2 I\{|X_j| > t\}|Z = 1] + o_{P_n}(1) .$$

The Cauchy-Schwartz inequality implies that $E_{P_n}[X_j^2 I\{|X_j| > t\}|Z = 1] \le 1$. Hence, (59) is $o_{P_n}(1)$ uniformly in t. Note further that the class of functions

$$\{zx^2 I\{|x| > t\} : t \in \mathbf{R}\} \tag{61}$$

is a VC class of functions. Therefore, by Theorem 2.6.7 and Theorem 2.8.1 of van der Vaart and Wellner (1996), we see that the class of functions (61) is Glivenko-Cantelli uniformly over $\mathbf{P}$. Since $P_n\{Z = 1\} > \epsilon$, it follows that the supremum over $t \in \mathbf{R}$ of (60) tends in probability to zero under P_n. The desired conclusion (58) follows. $\blacksquare$

To complete the argument, it now suffices to argue as in the proof of Lemma S.12.2 in Romano and Shaikh (2012). $\blacksquare$

Lemma E.5 *Let* $(X_i, Y_i, Z_i), i = 1, \ldots, n$ *be an i.i.d. sequence of random variables with distribution* $P \in \mathbf{P}$ *on* $\mathbf{R}^k \times \mathbf{R}^k \times \{0, 1\}$. *Suppose (43) holds for some* $\epsilon > 0$ *and for all* $1 \le j \le k$ *that (44) and (45) hold. Then, for any* $\{P_n \in \mathbf{P} : n \ge 1\}$,

$$\rho(\hat{P}_n, P_n) \xrightarrow{P_n} 0 ,$$

where

$$\rho(Q, P) = \max \Bigg\{ ||V(Q) - V(P)||, |Q\{Z = 1\} - P\{Z = 1\}|,$$

$$\max_{1 \le j \le k} \left\{ \int_0^\infty |r_j(\lambda, Q) - r_j(\lambda, P)| d\lambda \right\}, \max_{1 \le j \le k} \left\{ \int_0^\infty |s_j(\lambda, Q) - s_j(\lambda, P)| d\lambda \right\} \Bigg\} . \quad (62)$$

Here, $V(P)$, $r_j(\lambda, P)$, and $s_j(\lambda, P)$ are defined as in (53), (56), and (57), respectively, and $|| \cdot ||$ denotes the component-wise maximum of the absolute value of all elements.

PROOF: By arguing as in the proof of Lemma E.2, we have that

$$\hat{P}_n\{Z = 1\} - P_n\{Z = 1\} = \bar{Z}_n - P_n\{Z = 1\} \overset{P_n}{\to} 0 .$$

The desired result now follows from Lemmas E.3 and E.4. ∎

Lemma E.6 *Let $\mathbf{P}$ be a set of distributions on $\mathbf{R}^k \times \mathbf{R}^k \times \{0, 1\}$ such that (43) holds for some $\epsilon > 0$ and for all $1 \le j \le k$ that (44) and (45) hold. Let $\mathbf{P}'$ be the set of all distributions on $\mathbf{R}^k \times \mathbf{R}^k \times \{0, 1\}$. Define $\rho(Q, P)$ as in (62) and $J_n(x, P)$ as in (46). Then, for any $\{Q_n \in \mathbf{P}' : n \ge 1\}$ and $\{P_n \in \mathbf{P} : n \ge 1\}$ satisfying $\rho(Q_n, P_n) \to 0$,*

$$\limsup_{n \to \infty} \sup_{P \in \mathbf{P}} |J_n(x, Q_n) - J_n(x, P_n)| \to 0 . \quad (63)$$

PROOF: Consider sequences $\{Q_n \in \mathbf{P}' : n \ge 1\}$ and $\{P_n \in \mathbf{P} : n \ge 1\}$ satisfying $\rho(Q_n, P_n) \to 0$. By arguing as in the proof of Lemma S.12.1 in Romano and Shaikh (2012), we see that

$$\lim_{\lambda \to \infty} \limsup_{n \to \infty} r_j(\lambda, P_n) = 0$$
$$\lim_{\lambda \to \infty} \limsup_{n \to \infty} r_j(\lambda, Q_n) = 0$$
$$\lim_{\lambda \to \infty} \limsup_{n \to \infty} s_j(\lambda, P_n) = 0$$
$$\lim_{\lambda \to \infty} \limsup_{n \to \infty} s_j(\lambda, Q_n) = 0 .$$

We now establish (63). Suppose by way of contradiction that (63) fails. It follows that there exists a subsequence such that n_m such that $V(P_{n_m}) \to V^*$, $V(Q_{n_m}) \to V^*$, and either

$$\sup_{x \in \mathbf{R}} |J_{n_m}(x, P_{n_m}) - \Phi_{V^*}(x)| \not\to 0 \quad (64)$$

or

$$\sup_{x \in \mathbf{R}} |J_{n_m}(x, Q_{n_m}) - \Phi_{V^*}(x)| \not\to 0 . \quad (65)$$

Let $W_n(P_n)$ be the vector whose jth element for $1 \le j \le k$ is given by

$$\frac{\frac{1}{n_1} \sum_{1 \le i \le n : Z_i = 1} X_{i,j} - \mu_{X_j | Z = 1}(P) - \frac{1}{n_0} \sum_{1 \le i \le n : Z_i = 0} Y_{i.j} - \mu_{Y_j | Z = 0}(P)}{\sqrt{\frac{\sigma^2_{X_j | Z = 1}(\hat{P}_n)}{n_1} + \frac{\sigma^2_{Y_j | Z = 0}(\hat{P}_n)}{n_0}}} .$$

From Lemmas B.4 and B.5 in Bhattacharya et al. (2012) and Slutsky's Lemma, we see that

$$W_{n_m}(P_{n_m}) \xrightarrow{d} \Phi_{V^*}(x)$$

under P_{n_m}. It therefore follows from Polya's Theorem that (64) can not hold. Similarly, we see that (65) can not hold. The desired conclusion thus follows. ■

PROOF OF THEOREM E.1: The desired result follows immediately from Lemmas E.5 and E.6 and Theorem 2.4 in Romano and Shaikh (2012). ■

References

ABREVAYA, J., HAUSMAN, J. and KHAN, S. (2010). Testing for causal effects in a generalized model with endogenous regressors. *Econometrica*, **6** 2043–2061.

ANGRIST, J. D., IMBENS, G. D. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, **91** 444–454.

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

BHATTACHARYA, J., SHAIKH, A. and VYTLACIL, E. (2008). Treatment effect bounds under monotonicity conditions: An application to Swan-Ganz catherization. *American Economic Review, Papers and Proceedings*, **98** 351–356.

BHATTACHARYA, J., SHAIKH, A. M. and VYTLACIL, E. (2012). Treatment effect bounds: An application to Swan-Ganz catheterization. *Journal of Econometrics*, **168** 223–243.

CHIBURIS, R. (2010). Semiparametric bounds on treatment effects. *Journal of Econometrics*, **159** 267–275.

HECKMAN, J. and VYTLACIL, E. (2001a). Instrumental variables, selection models, and tight bounds on the average treatment effect. In *Econometric Evaluations of Active Labor Market Policies in Europe* (M. Lechner and F. Pfeiffer, eds.). Heidelberg; New York: Physica-Verlag.

HECKMAN, J. J. and VYTLACIL, E. J. (2001b). Local instrumental variables. In *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya* (C. Hsiao, K. Morimune and J. Powell, eds.). Cambridge: Cambridge University Press.

HECKMAN, J. J. and VYTLACIL, E. J. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, **73** 669–738.

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

IMBENS, G. W. and RUBIN, D. B. (1997). Estimating outcome distributions for compliers in instrumental variable models. *Review of Economic Studies*, **64** 555–574.

KITAGAWA, T. (2008). A bootstrap test for instrument validity in the heterogeneous treatment effect model. Mimeo, University College London.

MANSKI, C. (1990). Nonparametric bounds on treatment effects. *The American Economic Review Papers and Proceedings*, **80** 319–323.

MANSKI, C. and PEPPER, J. (2000). Monotone instrumental variables: With an application to the returns to schooling. *Econometrica*, **68** 997–1010.

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

ROMANO, J., SHAIKH, A. and WOLF, M. (2012). A simple two-step method for testing moment inequalities with an application to inference in partially identified models. Tech. rep., University of Chicago.

ROMANO, J. P. and SHAIKH, A. M. (2012). On the uniform asymptotic validity of subsampling and the bootstrap. *The Annals of Statistics*, **40** 2798–2822.

SHAIKH, A. M. and VYTLACIL, E. J. (2005). Threshold crossing models and bounds on treatment effects: A nonparametric analysis. Technical Working Paper 0307, National Bureau of Economic Research.

SHAIKH, A. M. and VYTLACIL, E. J. (2011). Partial identification in triangular simultaneous equations models with binary dependent variables. *Econometrica*, **79** 949–955.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer, New York.

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