

NBER WORKING PAPER SERIES

THE CAUSAL INTERPRETATION OF TWO-STAGE LEAST SQUARES WITH
MULTIPLE INSTRUMENTAL VARIABLES

Magne Mogstad
Alexander Torgovitsky
Christopher R. Walters

Working Paper 25691
<http://www.nber.org/papers/w25691>

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2019, Revised July 2020

We thank Stéphane Bonhomme, Vishal Kamat, Azeem Shaikh, and Ed Vytlačil for helpful comments, and Christine Blandhol, John Bonney, and Conroy Lau for outstanding research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at <http://www.nber.org/papers/w25691.ack>

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2019 by Magne Mogstad, Alexander Torgovitsky, and Christopher R. Walters. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables
Magne Mogstad, Alexander Torgovitsky, and Christopher R. Walters
NBER Working Paper No. 25691
March 2019, Revised July 2020
JEL No. C01,C1,C26

ABSTRACT

Empirical researchers often combine multiple instrumental variables (IVs) for a single treatment using two-stage least squares (2SLS). When treatment effects are heterogeneous, a common justification for including multiple IVs is that the 2SLS estimand can be given a causal interpretation as a positively-weighted average of local average treatment effects (LATEs). This justification requires the well-known monotonicity condition. However, we show that with more than one instrument, this condition can only be satisfied if choice behavior is effectively homogenous. Based on this finding, we consider the use of multiple IVs under a weaker, partial monotonicity condition. We characterize empirically verifiable sufficient and necessary conditions for the 2SLS estimand to be a positively-weighted average of LATEs under partial monotonicity. We apply these results to an empirical analysis of the returns to college with multiple instruments. We show that the standard monotonicity condition is at odds with the data. Nevertheless, our empirical checks show that the 2SLS estimate retains a causal interpretation as a positively-weighted average of the effects of college attendance among complier groups.

Magne Mogstad
Department of Economics
University of Chicago
1126 East 59th Street
Chicago, IL 60637
and NBER
magne.mogstad@gmail.com

Christopher R. Walters
Department of Economics
University of California, Berkeley
530 Evans Hall #3880
Berkeley, CA 94720-3880
and NBER
crwalters@econ.berkeley.edu

Alexander Torgovitsky
University of Chicago
1126 E. 59th Street
Chicago, IL, 60637
atorgovitsky@gmail.com

1 Introduction

Instrumental variables (IVs) are widely used to estimate causal relationships. In practice, researchers often combine multiple IVs using two-stage least squares (2SLS). In Section 2, we report a survey of empirical papers using IV that were published in leading journals since 2000. More than half of these papers report results from a specification with multiple IVs for a single treatment, typically combined using 2SLS.

The textbook motivation for combining multiple IVs is statistical efficiency. However, this requires an assumption of constant treatment effects. In contrast, allowing for heterogeneous treatment effects is a key motivation in the modern program evaluation literature, and one which is supported by a large body of empirical work.¹ In an influential paper, Imbens and Angrist (1994, “IA” hereafter) provided an alternative justification for using 2SLS with multiple IVs, one which allows for heterogeneous treatment effects. They showed that the 2SLS estimand can be interpreted as a positively-weighted average of local average treatment effects (LATEs) for subpopulations whose treatment status is affected by the instruments. This result holds for any number of instruments, as long as IA’s “monotonicity” condition is satisfied.²

The fact that widespread empirical practice rests on the monotonicity condition raises a number of questions. What requirements does this condition place on choice behavior when there are multiple IVs? Can we expect these requirements to be satisfied? If not, is it still possible to retain a causal interpretation of the 2SLS estimand while allowing for unobserved heterogeneity in both treatment effects and choice behavior? The contribution of our paper is to answer these questions and, by doing so, offer theoretical and empirical guidance for researchers who wish to use 2SLS with multiple IVs.

In Section 3, we begin our theoretical analysis by showing that the monotonicity condition cannot be satisfied with more than one instrument without restricting choice behavior to be effectively homogenous.³ For example, if the treatment is college at-

¹ Heckman (2001) compiled a list of empirical evidence on heterogeneous treatment effects prior to 2001. More recent papers that find evidence of unobserved heterogeneity in treatment effects include Bitler, Gelbach, and Hoynes (2006), Doyle Jr. (2007), Moffitt (2008), Carneiro and Lee (2009), Firpo, Fortin, and Lemieux (2009), Carneiro, Heckman, and Vytlacil (2011), Maestas, Mullen, and Strand (2013), Bitler, Hoynes, and Domina (2014), Walters (2014), Felfe and Lalive (2014), French and Song (2014), Havnes and Mogstad (2015), Kirkeboen, Leuven, and Mogstad (2016), Kline and Walters (2016), Carneiro, Lokshin, and Umapathi (2016), Cornelissen, Dustmann, Raute, and Schönberg (forthcoming), Nybom (2017), and Brinch, Mogstad, and Wiswall (2017).

² In discussing the use of multiple IVs instead of a single binary instrument, Angrist and Pischke (2009, p. 173) write that “The econometric tool remains 2SLS and the interpretation remains fundamentally similar to the basic LATE result, with a few bells and whistles.”

³Our analysis here builds upon points made by Heckman and Vytlacil (2005, Section 6) and Heckman,

tendance and the instruments are tuition and proximity, the monotonicity condition requires all individuals to respond more to tuition than to proximity, or vice versa. This is a concerning implication; it shows that appealing to IA monotonicity to justify combining multiple IVs comes at the cost of assuming homogeneity in choice behavior.

Motivated by this observation, we then consider a weaker, *partial* version of the monotonicity condition. The partial monotonicity condition is that the IA monotonicity condition is satisfied for each instrument separately, holding all of the other instruments fixed. We show that partial monotonicity is satisfied if each instrument makes every individual weakly more likely to choose treatment. For example, a sufficient condition for partial monotonicity is that all individuals are at least as likely to attend college if they live closer to a college or face lower tuitions. However, unlike the IA monotonicity condition, partial monotonicity does not restrict heterogeneity in the relative impacts of different instruments; it allows for some individuals to respond more to tuition than to proximity, and for others to respond more to proximity than to tuition.

In Section 4, we show that even though partial monotonicity permits heterogeneous choice behavior, it can still be sufficient to give the 2SLS estimand a causal interpretation as a positively-weighted average of LATEs. Moreover, we characterize sufficient and necessary conditions for this interpretation, and show that the conditions can be checked empirically. In a simple case with two binary instruments this amounts to verifying that the *unconditional* correlations between the treatment and each instrument have the expected sign. More generally, our results provide a set of tests that researchers can report alongside 2SLS estimates formed from multiple IVs. These tests are implemented in our companion Stata module, `mivcausal`.⁴

In Section 5, we apply our results to estimating the returns to college attendance, using the same data as Carneiro et al. (2011). We focus on two instruments used by those authors: local labor market conditions and distance to college. We show that these instruments generate different estimated marginal treatment effect (MTE) schedules, a finding which is inconsistent with IA monotonicity. Nevertheless, our empirical checks show that under partial monotonicity the 2SLS estimand retains an interpretation as a positively-weighted average of the causal effects of college attendance among complier groups. This finding illustrates how assumptions weaker than IA monotonicity can be sufficient for the modest empirical goal of using 2SLS to recover a weighted average of LATEs.

Urzua, and Vytlačil (2006, Section III.D).

⁴ The module is available for download from the Boston College Statistical Software Components (SSC) archive.

2 Survey on the Use of Multiple Instruments

Empirical researchers often combine multiple IVs using 2SLS. To document this practice, we searched the Web of Science Database for articles published between January 2000 and October 2018 containing the words “instrument” or “instrumental variable” in the abstract, title, or topic words. We restricted the search to the following five journals: *Journal of Political Economy*, *American Economic Review*, *Quarterly Journal of Economics*, *Review of Economic Studies*, and *Econometrica*. In total, 266 articles matched our search criteria.

We restrict our attention to the 122 of these papers that use at least one IV in an empirical application. The other 144 papers not included were either methodological papers without an empirical application, or were papers that used the word “instrument” in a different context, such as to describe a policy or financial instrument. Column (1) of Table 1 tabulates the papers used in our survey by the journal in which they were published. The other columns of Table 1 categorize these papers based on the number of IVs used, the number of IVs compared to the number of endogenous variables, and the choice of IV estimator.

Column (2) includes papers containing at least one specification in the main body of the paper that included more than one IV. The number of IVs is defined as the number of variables excluded from the outcome equation. If a paper has multiple IVs, but each reported specification used no more than one instrument, then we do not count it in column (2). The bottom row of column (2) reveals that more than half of the papers in our sample used more than one instrument.

In column (3) of Table 1, we count the subset of papers from column (2) that have at least one specification in the main body of the paper with more IVs than endogenous variables. This is typically referred to as an “over-identified” specification in a constant-effects model. Comparing column (3) to column (2), we see that most papers that used more than one IV had a specification with fewer endogenous variables than IVs. Nearly all of these specifications have just a single endogenous variable, consistent with the framework that we will use in our analysis.

A few of the papers in column (3) used more IVs than endogenous variables in a way that was either nonstandard or unclear to us. In column (4), we remove these papers and focus on only those that combined multiple IVs using 2SLS or optimally-weighted generalized method of moments (GMM). This leaves 43% of papers across the five journals. This confirms our claim that combining more instruments than endogenous variables through 2SLS (or occasionally GMM) is widespread empirical practice. This practice is the focus of this paper.

Table 1: IV papers by journal and type

	(1)	(2)	(3)	(4)
	All papers	Papers that use multiple IVs	Papers with more IVs than endogenous variables	Papers with more IVs than endogenous variables using 2SLS/GMM
<i>American Economic Review</i>	100% 44	39% 17	36% 16	34% 15
<i>Quarterly Journal of Economics</i>	100% 28	54% 15	46% 13	43% 12
<i>Journal of Political Economy</i>	100% 23	65% 15	48% 11	43% 10
<i>Econometrica</i>	100% 15	67% 10	60% 9	53% 8
<i>Review of Economic Studies</i>	100% 12	67% 8	67% 8	58% 7
All	100% 122	53% 65	47% 57	43% 52

In Table 2, we categorize the papers in column (4) of Table 1 into three types by the relationship between their multiple IVs. Case **A** is when at least two of the IVs are economically distinct quantities. For example, Carneiro et al. (2011) instrument for college attendance with measures of local labor market conditions, distance to college, and local tuition in public four-year colleges. As shown in column (2), the vast majority of papers fall into case **A**. Given its empirical importance, our analysis will be centered around this case.

The other two cases are less common. Case **B** are papers in which all IVs in the multiple IV specification are formed via interactions between pre-determined covariates and a single exogenous instrument. For example, Angrist and Krueger (1991) use a full set of indicators interacted between state and quarter of birth as instruments for years of schooling. As we see in column (3), case **B** accounts for only 13% of the papers, making it the least common of the three types of papers with multiple IVs. Case **C** is when the multiple IV specification uses functions of a single exogenous variable. For example, using distance and its square as instruments for college attendance would fall into this case. As shown in column (4), Case **C** is also relatively rare, accounting for only 19% of the papers.

Table 2: Multiple IV papers by journal and relationship between their IVs

	(1)	(2)	(3)	(4)
	Papers with more IVs than endogenous variables using 2SLS/GMM	Case A	Case B	Case C
<i>American Economic Review</i>	100% 15	60% 9	13% 2	27% 4
<i>Quarterly Journal of Economics</i>	100% 12	75% 9	8% 1	17% 2
<i>Journal of Political Economy</i>	100% 10	70% 7	0% 0	30% 3
<i>Econometrica</i>	100% 8	63% 5	38% 3	0% 0
<i>Review of Economic Studies</i>	100% 7	71% 5	14% 1	14% 1
All	100% 52	67% 35	13% 7	19% 10

This table classifies the papers from column (4) of Table 1 into three cases of multiple instruments. Case **A** are studies that use multiple economically distinct instruments. Case **B** are studies that use covariate interactions with a single instrument. Case **C** are studies that use multiple functions of a single instrument. These cases are mutually exclusive and exhaustive; some proportions do not sum to 100% due to rounding.

3 Monotonicity Conditions with Multiple IVs

This section contains our results on the interpretation of the IA and partial monotonicity conditions when there are multiple IVs. We first develop an equivalent characterization of the IA condition, which facilitates a graphical interpretation of its content. Among other things, the graphical interpretation reveals that the IA condition has nothing to do with “monotonicity” in the usual sense of the phrase. Then, we argue that IA monotonicity severely restricts choice heterogeneity. This motivates interest in the weaker, partial monotonicity assumption. Throughout the paper, we consider the binary treatment case, since allowing for multiple treatments already complicates the IA monotonicity condition even with a single instrument (Angrist and Imbens, 1995; Heckman et al., 2006). For our theoretical analysis, we condition on covariates

nonparametrically and keep this conditioning implicit in the notation.

3.1 Definition of the IA Monotonicity Condition

Consider a population of individuals $i \in \mathcal{I}$. Denote individual i 's potential treatment status if some instrument vector Z_i were set to z by $D_i(z) \in \{0, 1\}$, where z takes values in some subset \mathcal{Z} of \mathbb{R}^L . We assume that the support of Z_i is contained in \mathcal{Z} , possibly as a proper subset. When $L > 1$, we view each component of the vector Z_i as comprising an economically distinct quantity. That is, if $L = 2$ then $Z_{i,1}$ and $Z_{i,2}$ will denote the two distinct instruments, each of which can take two or more values.

Imbens and Angrist (1994, "IA") introduced the following assumption on the potential treatment states, which they described as "monotonicity."

Assumption IAM. (IA Monotonicity) For all $z, z' \in \mathcal{Z}$ either $D_i(z) \geq D_i(z')$ for all $i \in \mathcal{I}$, or $D_i(z) \leq D_i(z')$ for all $i \in \mathcal{I}$.

Heckman and Vytlacil (2005, pp. 715-716) observed that Assumption IAM requires uniformity across individuals, not monotonicity in the instrument. The results ahead provide further justification of their observation. Nevertheless, to conform with the existing literature, we still refer to Assumption IAM as "IA monotonicity." For clarity, we refer to the usual definition of monotonicity as "actual monotonicity."

Assumption AM. (Actual Monotonicity) If $z' \geq z$ in the vector sense (component-wise), then $D_i(z') \geq D_i(z)$ for all $i \in \mathcal{I}$.

We show ahead that IA monotonicity (Assumption IAM) neither implies nor is implied by actual monotonicity (Assumption AM).

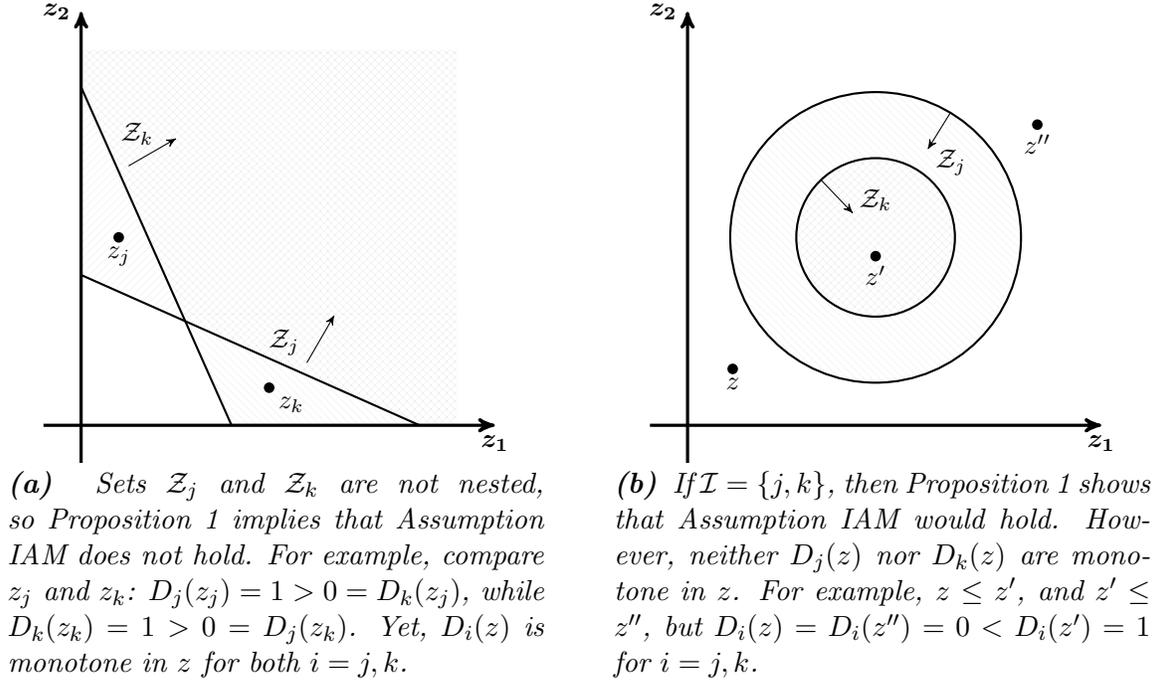
3.2 A Graphical Characterization of IA Monotonicity

Assumption IAM is a comparison across all individuals for any two values of Z_i . To interpret this condition when Z_i is a vector, it is useful to rephrase it as a comparison across all values of Z_i for any two individuals. The equivalent condition is that for any two individuals j and k , either j must take treatment under all instrument values that k does, or the opposite. This is the content of the following proposition.⁵

Proposition 1. For any $i \in \mathcal{I}$, define $\mathcal{Z}_i \equiv \{z \in \mathcal{Z} : D_i(z) = 1\}$ as the set of all instrument values for which individual i would take treatment. Assumption IAM holds if and only if for all $j, k \in \mathcal{I}$, either $\mathcal{Z}_j \subseteq \mathcal{Z}_k$, or $\mathcal{Z}_k \subseteq \mathcal{Z}_j$.

⁵ Proofs for all propositions are contained in Appendix A.

Figure 1: Assumption IAM neither implies nor is implied by monotonicity of $D_i(z)$ in z



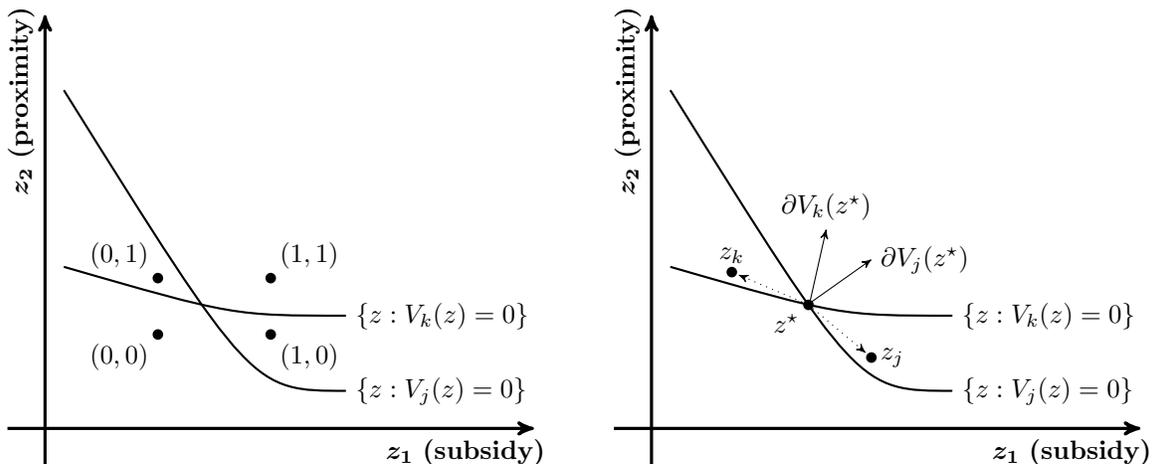
Proposition 1 shows that Assumption IAM can be interpreted as a “nesting condition” among the sets of instrument values that induce different agents to take treatment.⁶ This means that with two instruments one can gain intuition about the content of Assumption IAM by drawing sets in \mathbb{R}^2 .

For example, in Figure 1a, we have drawn two sets \mathcal{Z}_j and \mathcal{Z}_k that are not nested. Proposition 1 says that Assumption IAM fails, which can be verified by comparing the choices individuals j and k would make at the points marked z_j and z_k . Yet, for both individuals j and k , the instrument has a monotonic effect in the sense that $D_i(z)$ is increasing in z . That is, if $z' \geq z$ as a vector (component-wise), then $D_i(z') \geq D_i(z)$. This shows that Assumption AM does not imply Assumption IAM.

Figure 1b depicts the opposite case, in which $\mathcal{Z}_k \subseteq \mathcal{Z}_j$. If \mathcal{I} only consists of individuals like j and k , then Proposition 1 implies that Assumption IAM is satisfied. However, the instrument does not have a monotonic effect on treatment choice. For example, moving from z to $z' \geq z$ moves both individuals’ choices from 0 to 1, but moving from z' to $z'' \geq z'$ moves their choices back to 0. This shows that Assumption IAM does not imply monotonicity in the usual sense of Assumption AM.

⁶ This nesting condition is different than the “equivalent monotonicity condition” used by Vytlacil (2002, p. 335), although it shares a superficial resemblance. Vytlacil (2002, p. 336) used the sets \mathcal{Z}_i for his proof of the existence of an equivalent threshold-crossing model.

Figure 2: Assumption IAM requires homogenous choice behavior



(a) Each instrument takes two binary values. Individual j has $\mathcal{Z}_j = \{(1,0), (1,1)\}$ and individual k has $\mathcal{Z}_k = \{(0,1), (1,1)\}$. Points $(1,0)$ and $(0,1)$ violate the nesting condition in Proposition 1.

(b) An illustration of Proposition 2. When (3) fails to hold for two individuals $j, k \in \mathcal{I}$, one can find points z_j and z_k which violate the nesting condition in Proposition 1 by taking small steps in the directions of the dotted arrows.

3.3 Implications for Heterogeneity in Choice Behavior

In this section, we examine the restrictions that Assumption IAM places on choice behavior. To do this, we use a random utility model. Assume that individual i 's indirect utility from choosing d when the instrument is z is given by $V_i(d, z)$. The individual chooses $D_i(z) = 1$ if and only if $V_i(1, z) \geq V_i(0, z)$:

$$D_i(z) = \arg \max_{d \in \{0,1\}} V_i(d, z) \equiv \begin{cases} 1, & \text{if } V_i(z) \geq 0 \\ 0, & \text{if } V_i(z) < 0 \end{cases}, \quad (1)$$

where $V_i(z) \equiv V_i(1, z) - V_i(0, z)$ and ties are resolved in favor of treatment.

For concreteness, consider the familiar setting of the returns to schooling in which $D_i(z)$ represents a binary decision to attend college. Suppose that $z = (z_1, z_2)$, where z_1 is a tuition subsidy, and z_2 is proximity to a college (e.g. Kane and Rouse, 1993; Card, 1995). Larger values of either instrument encourage college attendance, so that $D_i(z)$ is a monotonic function of z and Assumption AM is satisfied. As observed in the previous section, this neither implies nor is implied by Assumption IAM. In Figure 2, we draw two possible indifference curves along which individuals j and k would be on the margin between attending and not attending college.

Suppose that z only takes the values $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$ shown in Figure 2a. Then individual j would attend college if and only if they received a tuition subsidy, regardless of whether they lived in close proximity. Individual k would attend college if and only if they lived in close proximity, regardless of whether they received a tuition subsidy. That is, $(1, 0)$ is in \mathcal{Z}_j but not \mathcal{Z}_k , and $(0, 1)$ is in \mathcal{Z}_k but not \mathcal{Z}_j , so that Assumption IAM fails. Thus, Assumption IAM does not permit individuals to differ in their responses to different incentives to attend college: all individuals must find either a tuition subsidy or distance to be more compelling. This suggests a strong form of preference homogeneity.

We can sharpen this statement when the instrument is continuous and the net indirect utility function is differentiable. This is shown in the next proposition, which is illustrated in Figure 2b.

Proposition 2. *Suppose that $D_i(z)$ is determined by (1). Let z^* be a point in the interior of \mathcal{Z} , and let $\mathcal{I}(z^*) \equiv \{i \in \mathcal{I} : V_i(z^*) = 0\}$ denote the set of individuals who are indifferent between treatment and non-treatment when faced with z^* . Suppose further that $V_i(z)$ is a continuously differentiable function of z in a neighborhood of z^* for all $i \in \mathcal{I}$. Then Assumption IAM implies that*

$$\partial_1 V_j(z^*) \partial_2 V_k(z^*) = \partial_1 V_k(z^*) \partial_2 V_j(z^*) \quad (2)$$

for all $j, k \in \mathcal{I}(z^*)$, where $\partial_\ell V_i(z) \equiv \frac{\partial}{\partial z_\ell} V_i(z)$ for $\ell = 1, 2$.

Proposition 2 says that if Assumption IAM holds, then any two individuals who are indifferent between treatment and non-treatment at z^* must have the same marginal rate of substitution between the two components of the instrument. That is, assuming that the second component has an impact at z^* (so that $\partial_2 V_i(z^*) \neq 0$), Assumption IAM implies

$$\frac{\partial_1 V_j(z^*)}{\partial_2 V_j(z^*)} = \frac{\partial_1 V_k(z^*)}{\partial_2 V_k(z^*)} \quad (3)$$

for all individuals j and k who are indifferent at z^* . This is a strong statement about preference homogeneity.

For example, suppose that individual i 's net utility from attending college is given by the random coefficients specification

$$V_i(z) = B_{i,0} + B_{i,1}z_1 + z_2 \quad \text{so that} \quad D_i(z) = \mathbb{1}[B_{i,0} + B_{i,1}z_1 + z_2 \geq 0], \quad (4)$$

where $B_{i,1} \geq 0$ controls variation in the taste for subsidies relative to proximity. Propo-

sition 2 shows that Assumption IAM does not hold if $B_{i,1}$ varies with i .⁷ Thus, the college attendance decision of every individual is either affected more by tuition subsidies (if $b_1 \equiv B_{i,1} \geq 1$), or by proximity (if $b_1 < 1$), and all individuals trade off these incentives at the same rate. Assumption IAM does not permit heterogeneity in these behavioral responses.

3.4 Partial Monotonicity

Assumption IAM creates unattractive implications for choice behavior because it requires cross-instrument comparisons, such as the comparison between $(0, 1)$ and $(1, 0)$ in Figure 2a. We can eliminate these comparisons by considering a condition that compares only values of a single component of the instrument, holding all other components fixed. To state such a condition, we divide vectors $z \in \mathcal{Z}$ into their ℓ th component, z_ℓ , and all other $(L - 1)$ components, $z_{-\ell}$. We write $z = (z_\ell, z_{-\ell})$ to emphasize the separation of the ℓ th component.

Consider the following assumption:⁸

Assumption PM. (*Partial Monotonicity*) Take any $\ell = 1, \dots, L$, and let $(z_\ell, z_{-\ell})$ and $(z'_\ell, z_{-\ell})$ be two points in \mathcal{Z} . Then either $D_i(z_\ell, z_{-\ell}) \geq D_i(z'_\ell, z_{-\ell})$ for all $i \in \mathcal{I}$, or $D_i(z_\ell, z_{-\ell}) \leq D_i(z'_\ell, z_{-\ell})$ for all $i \in \mathcal{I}$.

Assumption IAM clearly implies Assumption PM. When $L = 1$, Assumption PM is equivalent to Assumption IAM; when $L > 1$, it is strictly weaker. To see this, recall Figure 2a, where we determined that Assumption IAM fails. Holding $z_2 = 0$ fixed, both individuals j and k are weakly induced to treatment by switching z_1 from 0 to 1. The same is true when the roles of z_1 and z_2 are swapped. If \mathcal{I} consisted of only individuals like j and k , then Assumption PM would be satisfied.

Figure 2 suggests that a simple sufficient condition for Assumption PM is monotonicity in the usual sense of Assumption AM. This is the content of the following proposition.

Proposition 3. *Assumption AM implies Assumption PM.*

Unlike Assumption IAM, Assumption AM can be easily satisfied in random utility models with heterogeneous preferences. For example, if $V_i(z)$ follows the random coefficients specification (4), then it will be satisfied if $B_{i,1}$ is positive for all i . This is

⁷ Heckman et al. (2006, p. 399) note that Assumption IAM *can* fail in random coefficient specifications like (4). Our analysis shows that it *must* fail when the instruments are continuous.

⁸ Mountjoy (2019) uses a similar assumption in a setting with multiple unordered treatments.

easy to interpret and justify: all individuals are more likely to attend college if tuition is lower, even if they differ in their preferences for tuition relative to proximity. More generally, Proposition 3 shows that Assumption PM is satisfied whenever $V_i(z)$ is weakly increasing in z .⁹

Our running example of the returns to college was chosen because it is familiar, not because it favors either Assumption IAM or PM. In Appendix B, we review three empirical papers from our survey in Section 2. These papers are on diverse topics, but each one combines multiple economically distinct IVs using 2SLS, and so falls into column (2) (“Case **A**”) of Table 2. For each paper, we discuss the substantive assumptions represented by both Assumptions IAM and PM. Our conclusion from that exercise is consistent with our theoretical results: Assumption IAM is an extremely strong condition with multiple instruments, whereas Assumption PM—while not innocuous—is much less controversial.

It is important to observe that Assumption AM is sufficient but not necessary for Assumption PM. For example, consider the influential study by Angrist and Evans (1998), who used the sex composition of a family’s first two children to instrument for family size. Let D_i indicate whether the family had exactly two or more than two children. Let $Z_{i,1}$ and $Z_{i,2}$ indicate the sexes of their first two children. Angrist and Evans (1998) present empirical evidence that families for which these two sexes are the same ($Z_i = (0, 0)$ or $Z_i = (1, 1)$) are more likely to have a third child. Suppose that this is due to a uniform preference ordering in the population, so that for all $i \in \mathcal{I}$

$$\begin{aligned} D_i(0, 0) &\geq D_i(0, 1), & D_i(0, 0) &\geq D_i(1, 0), \\ \text{and } D_i(1, 1) &\geq D_i(0, 1), & D_i(1, 1) &\geq D_i(1, 0). \end{aligned} \tag{5}$$

This is consistent with Assumption PM, but it violates Assumption AM.

Another way to see the difference between Assumptions AM and PM is to consider the interacted random coefficient specification

$$V_i(z) = B_{i,0} + B_{i,1}z_1 + z_2 + B_{i,2}z_1z_2, \tag{6}$$

where we suppose for simplicity that $\mathcal{Z} = \{0, 1\}^2$. If all individuals $i \in \mathcal{I}$ have $B_{i,1} \geq 0$, $B_{i,2} \leq -1$, but $B_{i,1} \leq -B_{i,2}$, then Assumption AM fails while Assumption PM is satisfied. The reason is due to the strong negative interaction effect (submodularity)

⁹ More generally still, Theorem 4 of Milgrom and Shannon (1994) implies that Assumption PM will be satisfied if $V_i(d, z)$ has the single-crossing property in $(d; z)$. The single-crossing property is ordinal, and is strictly weaker than monotonicity in the usual sense.

between z_1 and z_2 on indirect net utility, which is controlled here by $B_{i,2}$. This implies that $D_i(z_1, z_2)$ is increasing as a function of z_1 when $z_2 = 0$, but decreasing when $z_2 = 1$, and similarly when the roles of z_1 and z_2 are reversed. This violates Assumption AM, even though Assumption PM is satisfied.¹⁰

4 Interpreting 2SLS Under Partial Monotonicity

Imbens and Angrist (1994, Theorem 2) showed that under standard instrument exogeneity and relevance conditions, Assumption IAM ensures that the 2SLS estimand can be written as a weighted average of causal effects for complier subpopulations. The weights are convex in that they are non-negative and sum to one. Their result holds regardless of the number of instruments, as long as the first stage for the 2SLS estimand is fully saturated. However, it crucially depends on Assumption IAM, which we have shown requires a strong form of preference homogeneity with two or more distinct instruments. In this section, we show that the Imbens and Angrist (1994) result can be partially salvaged under Assumption PM.

4.1 Potential Outcomes and Exogeneity Condition

To define the 2SLS estimand, we need to introduce an outcome variable, Y_i . We write potential outcomes for Y_i as $Y_i(1)$ and $Y_i(0)$ to correspond to setting D_i to treatment ($D_i = 1$) and non-treatment states ($D_i = 0$). The notation incorporates the usual exclusion restriction that Z_i has no direct causal effect on Y_i . The observed outcome is $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0) = Y_i(D_i)$. The observed treatment state is related to the potential treatment states analyzed in the previous section as

$$D_i = \sum_{z \in \mathcal{Z}} \mathbb{1}[Z_i = z] D_i(z) = D_i(Z_i).$$

Throughout the paper, we maintain the following exogeneity condition:

Assumption E. (*Exogeneity*) $(Y_i(0), Y_i(1), \{D_i(z)\}_{z \in \mathcal{Z}}) \perp\!\!\!\perp Z_i$.

Assumption E is common in nonparametric IV models, and identical to Condition 1 of Imbens and Angrist (1994). For simplicity, we state the condition in terms of full independence, although our analysis of the 2SLS estimand only involves mean outcomes, so only requires a weaker form of Assumption E.¹¹

¹⁰ To see that the above configuration satisfies Assumption PM, note that $B_{i,1} \geq 0$ implies that $D_i(0, 0) \leq D_i(1, 0)$, $B_{i,1} \leq -B_{i,2}$ implies that $D_i(0, 1) \geq D_i(1, 1)$, $D_i(0, 0) \leq D_i(0, 1)$ by virtue of the normalized coefficient on z_2 , and $B_{i,2} \leq -1$ implies that $D_i(1, 0) \geq D_i(1, 1)$.

¹¹ The weaker form still requires that $\{D_i(z)\}_{z \in \mathcal{Z}} \perp\!\!\!\perp Z_i$, but relaxes full joint independence to the restriction

Table 3: Groups under Assumption PM with (7)

	z^1	z^2	z^3	z^4		
G_i (group g)	$D_i(0, 0)$	$D_i(0, 1)$	$D_i(1, 0)$	$D_i(1, 1)$	\mathcal{C}_g	\mathcal{D}_g
Always-taker (at)	1	1	1	1	\emptyset	\emptyset
Eager complier (ec)	0	1	1	1	$\{2\}$	\emptyset
Reluctant complier (rc)	0	0	0	1	$\{4\}$	\emptyset
Never-taker (nt)	0	0	0	0	\emptyset	\emptyset
Z_1 complier (1c)	0	0	1	1	$\{3\}$	\emptyset
Z_2 complier (2c)	0	1	0	1	$\{2, 4\}$	$\{3\}$

The sets \mathcal{C}_g and \mathcal{D}_g are integers k for which group type g acts as a complier or defier when considering the contrast z^{k-1} to z^k . This notation is explained in more detail in Section 4.4.

4.2 Two Binary Instruments

We first consider the case in which $Z_i = (Z_{i,1}, Z_{i,2})$ consists of two binary instruments $Z_{i,1} \in \{0, 1\}$ and $Z_{i,2} \in \{0, 1\}$, with support $\mathcal{Z} = \{0, 1\}^2$. We also assume that the direction of Assumption PM is known a priori to be

$$D_i(0, 0) \leq D_i(0, 1) \leq D_i(1, 1),$$

and $D_i(0, 0) \leq D_i(1, 0) \leq D_i(1, 1).$ (7)

This particular ordering is not required, but it helps with the notation and exposition to not have to consider multiple different orderings simultaneously; our more general results in Section 4.4 relax this condition.¹² The following proposition establishes that under these conditions the population \mathcal{I} can be partitioned into six mutually exclusive and exhaustive groups.

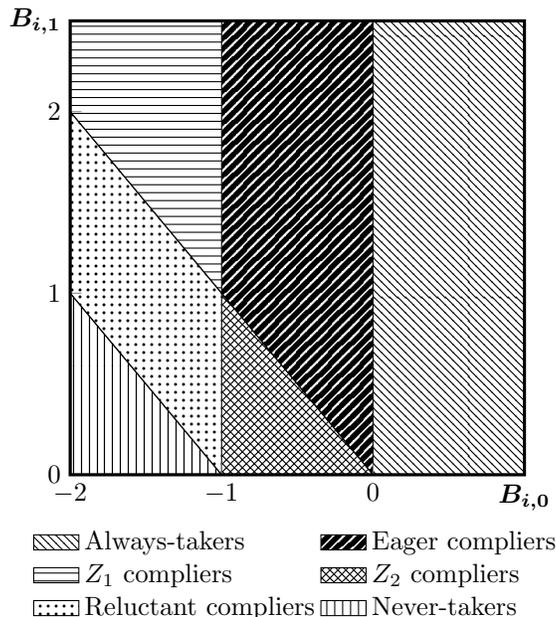
Proposition 4. *If $\mathcal{Z} = \{0, 1\}^2$ and Assumption PM is satisfied with the order shown in (7), then each $i \in \mathcal{I}$ belongs to exactly one of the six groups in Table 3.*

The terminology in Table 3 modifies that of Angrist, Imbens, and Rubin (1996). Always- and never- takers exhibit the same behavior regardless of either instrument. Z_1 compliers take treatment if and only if $Z_{i,1}$ is switched on, while Z_2 compliers take treatment if and only if $Z_{i,2}$ is switched on. Eager compliers participate in treatment if either instrument is on, and reluctant compliers only participate if both instruments

that $\mathbb{E}[Y_i(d) | \{D_i(z)\}_{z \in \mathcal{Z}}, Z_i = z']$ not depend on z' for both $d = 0$ and $d = 1$.

¹² Also, note that the direction of the comparisons are still identified under Assumption PM, just as they are under Assumption IAM, since Assumption E implies that $\mathbb{P}[D_i(z) = 1] = \mathbb{P}[D_i | Z_i = z]$.

Figure 3: Correspondence between random coefficients and behavioral types



These regions show how random coefficient realizations relate to behavioral types if treatment were determined by the binary choice model (4) with $B_{i,1} \geq 0$. For example, a Z_2 complier has $D_i(0,0) = 0$, $D_i(1,0) = 0$, $D_i(0,1) = 1$, and $D_i(1,1) = 1$. The first three choices imply that $B_{i,0} < 0$, $B_{i,0} + B_{i,1} < 0$, and $B_{i,0} + 1 > 0$, with the fourth choice implied by the third. The region of such $(B_{i,0}, B_{i,1})$ realizations is shown in cross-hatches.

are on. For any of the six groups, an increase in either instrument weakly increases treatment, which implies that Assumption AM also holds in this case. Assumption IAM will be violated if the population includes both Z_1 and Z_2 compliers, since in this case a change in Z_i from $(0, 1)$ to $(1, 0)$ would induce Z_1 compliers to enter treatment and Z_2 compliers to exit treatment. Figure 3 shows how realizations of random coefficients would map into these six group types if potential treatment states were generated by (4).

As in Imbens and Angrist (1994, Theorem 2), we consider the 2SLS estimand with a saturated first stage, and a second stage with D_i and a constant. That is, the 2SLS estimand is formed by using 1 (a constant), $Z_{i,1}$, $Z_{i,2}$, and $Z_{i,1}Z_{i,2}$ as instruments for 1 and D_i . Since the first stage is saturated, this 2SLS procedure generates the same coefficient estimate on D_i as the IV estimator that uses the propensity score $p(Z_i) \equiv \mathbb{P}[D_i = 1|Z_i]$ as the sole instrument for D_i . Let $\beta_{2sls} = \text{Cov}(Y_i, p(Z_i)) / \text{Cov}(D_i, p(Z_i))$ denote this coefficient. Let $G_i \in \{\text{at}, \text{nt}, \text{1c}, \text{2c}, \text{ec}, \text{rc}\}$ denote individual i 's group among the six shown in Table 3, and define $\pi_g \equiv \mathbb{P}[G_i = g]$ and $\Delta_g \equiv \mathbb{E}[Y_i(1) - Y_i(0)|G_i = g]$

to be the population shares and average treatment effects for each group. The following proposition establishes the relationship between these quantities and β_{2sls} .

Proposition 5. *Suppose Z_i has support $\mathcal{Z} \equiv \{0, 1\}^2$ and that Assumption PM is satisfied with the ordering in (7). Suppose in addition that Assumption E is satisfied, and that β_{2sls} exists. Then*

$$\beta_{2sls} = \sum_{g \in \{1c, 2c, ec, rc\}} \omega_g \Delta_g,$$

where the ω_g are weights that sum to 1. Both ω_{ec} and ω_{rc} are always non-negative. If $\pi_{1c} \geq \pi_{2c}$, then ω_{1c} is also non-negative, while the sign of ω_{2c} is given by

$$\text{sgn}(\omega_{2c}) = \mathbb{1}[\pi_{2c} > 0] \times \text{sgn} \left(\mathbb{P}[D_i = 1 | Z_{i,2} = 1] - \mathbb{P}[D_i = 1 | Z_{i,2} = 0] \right).$$

If $\pi_{2c} \geq \pi_{1c}$, then ω_{2c} is non-negative, and the sign of ω_{1c} is given by

$$\text{sgn}(\omega_{1c}) = \mathbb{1}[\pi_{1c} > 0] \times \text{sgn} \left(\mathbb{P}[D_i = 1 | Z_{i,1} = 1] - \mathbb{P}[D_i = 1 | Z_{i,1} = 0] \right).$$

Proposition 5 shows that the 2SLS estimand is a linear combination of average treatment effects for the four groups that change treatment status in response to one or both of the instruments. The groups are non-overlapping, and the weights on the groups (ω_g) sum to unity.¹³ This implies that $\beta_{2sls} = \Delta \equiv \Delta_g$ in the unlikely situation that there are constant treatment effects across behavioral groups. In general, however, the 2SLS estimand might not be a positively-weighted average of treatment effects for the four groups, because some of the weights might be negative. The weights for reluctant compliers and eager compliers are always non-negative. If $\pi_{1c} \geq \pi_{2c}$, the weight given to the Z_1 compliers is also non-negative, but the weight given to the Z_2 compliers can be either positive or negative.

The intuition is that a shift of Z_i from $(0, 1)$ to $(1, 0)$ induces Z_1 compliers to enter treatment and Z_2 compliers to exit treatment. If $\pi_{1c} \geq \pi_{2c}$, then the net effect of this shift is still more participation in treatment. However, the Z_2 compliers act as “defiers,” and therefore receive negative weight for this binary contrast. Whether the overall weight given to the Z_2 compliers is positive or negative depends on whether this negative weight is outweighed by the positive weight given to the Z_2 compliers in the two other instrument contrasts for which they enter treatment: $Z_i = (0, 0)$ to $Z_i = (0, 1)$, and $Z_i = (1, 0)$ to $Z_i = (1, 1)$. If instead $\pi_{2c} \geq \pi_{1c}$, then the roles of the Z_1 and Z_2 compliers are reversed.

¹³ Formulas for ω_g are given in the proof.

4.3 When is the 2SLS Estimand a Positively-Weighted Average of Compliers?

Proposition 5 shows that one can check whether the 2SLS estimand is a positively-weighted average of causal effects by examining observable relationships between treatment and instruments. Specifically, if both Z_1 and Z_2 compliers are present in the population and $\pi_{1c} \geq \pi_{2c}$, then ω_{2c} will be negative if and only if the coefficient on $Z_{i,2}$ in a regression of D_i on $Z_{i,2}$ (and a constant) is negative. Likewise, if $\pi_{2c} \geq \pi_{1c}$, then ω_{1c} will be negative if and only if the coefficient on $Z_{i,1}$ in a regression of D_i on $Z_{i,1}$ (and a constant) is negative. One can either check both cases, or check only the relevant case, which is identified by the sign of $p(1, 0) - p(0, 1) = \pi_{1c} - \pi_{2c}$.

It is important to note that except for π_{at} and π_{nt} , the group shares are not themselves separately identified. This is because there are five linearly independent group shares (after accounting for the fact that they sum to unity), but only four values of the propensity score. As mentioned, the sign of $p(1, 0) - p(0, 1)$ indicates whether Z_1 or Z_2 compliers are more common. However, it is not possible to determine whether the less common group is entirely absent. If it were the case that $\pi_{1c} = 0$ or $\pi_{2c} = 0$, then Assumption IAM would hold and the 2SLS estimand would necessarily be a positively-weighted average of the remaining three complier groups.

Finding negative weights based on the formulas in Proposition 5 represents a situation in which the unconditional relationship between an instrument and the treatment has a different sign than the *ceteris paribus* impact of the instrument. This may be rare in practice, since researchers often have prior beliefs regarding the impacts of the instruments (e.g. if each instrument is an encouragement to take treatment), and a researcher may be unlikely to use an instrument if the raw correlation with the treatment contradicts the expected sign. A necessary condition for such a contradiction is that the instruments are negatively correlated.

Proposition 6. *Suppose Z_i consists of two binary instruments that satisfy Assumption PM with the ordering in (7), that Assumption E is satisfied, and that β_{2sls} exists. If $\text{Cov}(Z_{i,1}, Z_{i,2}) \geq 0$, then both ω_{1c} and ω_{2c} are non-negative.*

An important special case of Proposition 6 is when the instruments are independent, so that $\text{Cov}(Z_{i,1}, Z_{i,2}) = 0$ and the 2SLS weights are guaranteed to be positive. The leading scenario in which the instruments would be negatively correlated is when $Z_{i,j} = 1$ tends to imply $Z_{i,k} = 0$ for $j \neq k$. For example, $Z_{i,1}$ and $Z_{i,2}$ may indicate two different arms in an experiment corresponding to different types of encouragement to take the treatment. In this setting, $\text{Cov}(Z_{i,1}, Z_{i,2}) < 0$ and negative weights are possible.

In Appendix C, we use the full sufficient and necessary characterization in Proposition 5 to develop formal statistical tests about the signs of the weights. We consider four tests of the null hypothesis that the weights are positive. In a small simulation study, we find that the test based on the procedure of Romano, Shaikh, and Wolf (2014) performs the best, with size only slightly below its nominal level. We also consider a test of the null hypothesis that at least one weight is negative. All five tests can be implemented using the `mivcausal` Stata package.

4.4 Multivalued Instruments

Suppose that Z_i consists of two or more distinct, discrete instruments, and that its support has K elements total. Label these elements as $\text{supp}(Z_i) \equiv \{z^1, \dots, z^K\}$ in increasing order of the propensity score, so that $p(z^k) \geq p(z^{k-1})$ for all $k \geq 2$. In the case considered in Section 4.2, $K = 4$ and the instrument values would be ordered as $z^1 = (0, 0)$, $z^2 = (0, 1)$, $z^3 = (1, 0)$, and $z^4 = (1, 1)$ if $p(1, 0) \geq p(0, 1)$, with the roles of z^2 and z^3 switched in the opposite case.

Let \mathcal{G} represent the set of all realizations of $\{D_i(z)\}_{z \in \mathcal{Z}}$ that are consistent with Assumption PM and the observed values of $p(z)$. In Section 4.2, where there were two binary instruments with the ordering (7), \mathcal{G} was composed of the six groups in Table 3. Table 4 displays another example consistent with the ordering (5) in which there are seven groups. As before, let G_i denote individual i 's group membership, let π_g denote the proportion of the population in each group g , and let Δ_g denote group g 's average treatment effect.

For each $g \in \mathcal{G}$, define the set:

$$\mathcal{C}_g = \{k \in \{2, \dots, K\} : D_i(z^k) = 1 \text{ and } D_i(z^{k-1}) = 0 \text{ for all } i \text{ with } G_i = g\}.$$

This is the set of instrument values k at which individuals in group g are compliers in the sense that they would not take treatment if $Z_i = z^{k-1}$, but would take treatment if $Z_i = z^k$. Similarly, define

$$\mathcal{D}_g = \{k \in \{2, \dots, K\} : D_i(z^k) = 0 \text{ and } D_i(z^{k-1}) = 1 \text{ for all } i \text{ with } G_i = g\}$$

as the set of instrument values at which individuals in group g act as defiers. For example, in Section 4.2, we had $\mathcal{C}_{1c} = \{3\}$, $\mathcal{D}_{1c} = \emptyset$, $\mathcal{C}_{2c} = \{2, 4\}$, and $\mathcal{D}_{2c} = \{3\}$. We also had $\mathcal{C}_{ec} = \{1\}$, $\mathcal{C}_{rc} = \{4\}$, $\mathcal{C}_{at} = \mathcal{C}_{nt} = \emptyset$, and $\mathcal{D}_g = \emptyset$ for each $g \in \{at, nt, ec, rc\}$; recall Table 3. Table 4 reports the sets \mathcal{C}_g and \mathcal{D}_g , for an alternative case consistent with (5).

Table 4: Groups under Assumption PM with (5)

	z^1	z^2	z^3	z^4		
G_i (group g)	$D_i(1, 0)$	$D_i(0, 1)$	$D_i(1, 1)$	$D_i(0, 0)$	\mathcal{C}_g	\mathcal{D}_g
Always-taker (at)	1	1	1	1	\emptyset	\emptyset
Group 2	0	1	1	1	$\{2\}$	\emptyset
Group 3	1	0	1	1	$\{3\}$	$\{2\}$
Group 4	0	0	1	1	$\{3\}$	\emptyset
Group 5	0	0	0	1	$\{4\}$	\emptyset
Group 6	0	0	1	0	$\{3\}$	$\{4\}$
Never-taker (nt)	0	0	0	0	\emptyset	\emptyset

As before, consider the same 2SLS specification used by Imbens and Angrist (1994, Theorem 2) with a saturated first stage, and a second stage that contains D_i and a constant. Let β_{2sls} denote the 2SLS estimand corresponding to the coefficient on D_i . The following proposition provides an interpretation of the 2SLS estimand.

Proposition 7. *Suppose Z_i takes K values $\{z^1, \dots, z^K\}$ labeled so that the propensity score is increasing and suppose that the support of Z_i is rectangular, that is $\text{supp}(Z_i) = \text{supp}(Z_{i,1}) \times \dots \times \text{supp}(Z_{i,L})$. If Assumptions PM and E are satisfied, and if β_{2sls} exists, then*

$$\beta_{2sls} = \sum_{g \in \mathcal{G}: \mathcal{C}_g \neq \emptyset} \omega_g \Delta_g,$$

where ω_g are weights such that $\sum_{g \in \mathcal{G}: \mathcal{C}_g \neq \emptyset} \omega_g = 1$, and

$$\text{sgn}(\omega_g) = \mathbb{1}[\pi_g > 0] \times \text{sgn} \left(\sum_{k=2}^K (\mathbb{1}[k \in \mathcal{C}_g] - \mathbb{1}[k \in \mathcal{D}_g]) \text{Cov} \left(D_i, \mathbb{1}[p(Z_i) \geq p(z^k)] \right) \right).$$

Proposition 7 shows that under Assumption PM, the 2SLS estimator produces a weighted average of treatment effects for groups that comply with some instrument change. The weights on each group could be positive or negative, but this can be checked empirically. To do this, one must generate the sets \mathcal{C}_g and \mathcal{D}_g by applying Assumption PM to the set of K instrument values and the observed ordering of the propensity score, as we did in the examples in Tables 3 and 4. The sign of the weight for group g is then determined by the overall sum of $\text{Cov}(D_i, \mathbb{1}[p(Z_i) \geq p(z^k)])$ for instrument values k at which they comply less the sum of these terms at values k for which they defy.

The additional rectangular support condition in Proposition 7 is necessary to ensure that the contrasts picked up in $\beta_{2\text{sls}}$ are restricted by Assumption PM. For example, if in the special case in Section 4.2 the support of Z_i were only $\{(0, 1), (1, 0)\}$, then either the Z_1 compliers or Z_2 compliers would always have negative weight. This is intuitive since Assumption PM does not place any direct restrictions on behavior in the contrast between $(0, 1)$ and $(1, 0)$.

4.5 Using Each Instrument Separately

A natural alternative to combining two or more instruments through 2SLS is to use each instrument separately. This may be unattractive due to a loss of statistical precision. However, it is justified under Assumption PM with the important caveat that all of the other instruments generally must be conditioned on as control variables in both the first and the second stage of the 2SLS estimator.¹⁴ This is necessary because Assumption E is stronger than the exogeneity condition in the traditional constant effects linear IV model. Whereas the traditional condition only requires Z_i to be exogenous with respect to the outcome process, Assumption E also requires Z_i to be independent of $\{D_i(z)\}_{z \in \mathcal{Z}}$, and thus exogenous with respect to potential treatments as well.

To see why conditioning is necessary, consider again the case with two binary instruments. In the heterogeneous effects framework, using the first instrument by itself requires considering marginal potential treatment states with respect to $Z_{i,1}$. These are related to the joint potential treatment states by

$$D_{i,1}(z_1) \equiv (1 - Z_{i,2})D_i(z_1, 0) + Z_{i,2}D_i(z_1, 1). \quad (8)$$

To apply the Imbens and Angrist (1994) LATE interpretation using only $Z_{i,1}$ as the instrument would require that $Z_{i,1}$ is independent of $(D_{i,1}(0), D_{i,1}(1))$. We can see from (8) that this is unlikely to hold outside of the case where $Z_{i,1}$ and $Z_{i,2}$ are themselves independent. However, Assumption E does imply that $Z_{i,1}$ and $(D_{i,1}(0), D_{i,1}(1))$ will be independent *conditional* on $Z_{i,2}$. Similarly, Assumption PM implies that Assumption IAM holds for the marginal potential treatments *conditional* on $Z_{i,2}$. See Mogstad, Torgovitsky, and Walters (2020) for a complete development of these ideas.

In Table 5, we report on the ways that researchers use multiple instruments separately using the same sample of papers from Section 2. Of the 65 papers in column (2) of Table 1 that used multiple IVs for the same treatment variable, we found 20 that fit a separate 2SLS model using a single IV for at least one specification reported in the main body of the paper. Column (2) of Table 5 restricts these papers to those that do

¹⁴ This has been noted elsewhere, e.g. Carneiro et al. (2011, footnote 6).

Table 5: Multiple IV papers that use IVs separately

	(1)	(2)	(3)
	Papers that ever use IVs separately	Papers that do not control for omitted IVs	Papers that always use IVs separately and never control for omitted IVs
<i>American Economic Review</i>	7	6	2
<i>Quarterly Journal of Economics</i>	5	5	4
<i>Journal of Political Economy</i>	6	5	2
<i>Econometrica</i>	1	1	0
<i>Review of Economic Studies</i>	1	1	1
All	20	18	9

This table includes the subset of the papers from column (2), Table 1 that fit a separate 2SLS model using a single IV in at least one specification in the main body of the paper.

not condition on the other omitted IVs when fitting such a model, while column (3) further restricts column (2) to papers that never used all IVs together in any specification. As shown in column (2), 90% of the papers that use multiple IVs separately do not control for the omitted instruments. Unless the multiple IVs are independent, these papers are reporting an estimated quantity that likely does not have a causal interpretation.

5 Estimating the Returns to College

We illustrate the results of the previous sections by estimating the returns to college attendance.

5.1 Sample and Instruments

Carneiro, Heckman, and Vytlacil (2011, “CHV”) used multiple IVs to estimate the returns to college in a sample of white men from the National Longitudinal Survey of Youth 1979 Cohort. We use their sample and focus on the two instruments that are the strongest predictors of college attendance: average log earnings in a youth’s county of residence at age 17, and the presence of a four-year college in a youth’s

Table 6: Summary statistics

	Mean (1)	Standard deviation (2)
Average log wage, 1989-1993	2.378	0.499
Attended college	0.495	0.500
Corrected AFQT score	0.449	0.952
Mother's years of schooling	12.102	2.335
Number of siblings	2.927	1.909
Urban residence at age 14	0.744	0.436
Year of birth	1959.759	2.340
Permanent local log earnings	10.283	0.188
Local log earnings in 1991	10.293	0.165
Low local earnings at age 17	0.250	0.433
Nearby four-year college at age 14	0.525	0.500
Sample size	1747	

Notes: This table reports means and standard deviations of key variables for white men in the National Longitudinal Survey of Youth 1979 sample. Column (1) displays mean characteristics, and column (2) shows standard deviations. AFQT score is corrected for differences in years of schooling at the time individuals took the test. Permanent local log earnings equals average log earnings from 1973 to 2000 for an individual's county of residence at age 17. Low local earnings at age 17 is an indicator equal to one if the residual from a regression of average log earnings for an individual's county at age 17 on the other covariates falls below the 25th percentile in the sample.

county of residence at age 14.¹⁵ For these two instruments, Assumption PM holds if all individuals view college as more attractive when local labor market opportunities are weaker and when a college is nearby. This seems plausible, since higher opportunity costs and longer distances are likely undesirable for everyone. In contrast, Assumption IAM requires all individuals to place the same relative weight on these economically distinct types of incentives.

¹⁵ These two instruments were also used in several earlier studies of the returns to education e.g. Card (1995) Cameron and Heckman (1998) Kling (2001), and Cameron and Taber (2004).

Table 7: OLS and 2SLS estimates of the returns to college

	OLS		2SLS				
	No controls (1)	Controls (2)	First stage		Second stage excluding:		
			No interaction (3)	Interaction (4)	Local earnings (5)	Nearby college (6)	Both (7)
College attendance (return per year)	0.085 (0.006)	0.042 (0.006)			0.198 (0.065)	-0.024 (0.100)	0.137 (0.049)
Low local earnings at 17		0.103 (0.033)	0.104 (0.022)	0.124 (0.033)		0.092 (0.051)	
Nearby four-year college at 14		0.000 (0.027)	0.061 (0.023)	0.069 (0.025)	-0.054 (0.033)		
Low local earnings*Nearby college		-0.067 (0.046)		-0.036 (0.045)			
Corrected AFQT score		0.096 (0.014)	0.232 (0.011)	0.232 (0.011)	-0.049 (0.062)	0.157 (0.094)	0.008 (0.047)
Mother's years of schooling		0.013 (0.005)	0.030 (0.004)	0.030 (0.004)	-0.006 (0.010)	0.021 (0.014)	0.001 (0.008)
Number of siblings		-0.005 (0.005)	-0.018 (0.005)	-0.018 (0.005)	0.006 (0.008)	-0.010 (0.009)	0.002 (0.007)
Urban residence at 14		0.020 (0.024)	0.031 (0.024)	0.032 (0.024)	-0.001 (0.029)	0.027 (0.028)	0.002 (0.027)
Permanent local log earnings		0.073 (0.073)	-0.081 (0.067)	-0.083 (0.067)	0.127 (0.088)	0.056 (0.068)	0.060 (0.070)
Local log earnings in 1991		0.727 (0.078)	0.224 (0.069)	0.228 (0.069)	0.581 (0.104)	0.780 (0.115)	0.641 (0.090)
First stage <i>F</i> -statistic for excluded instruments					21.970	7.182	15.451
<i>P</i> -value: overidentification test							0.080
Sample size	1747	1747	1747	1747	1747	1747	1747

Notes: This table reports ordinary least squares (OLS) and two-stage least squares (2SLS) estimates of the effects of college attendance on log wages for white men in the NLSY 1979. Columns (1) and (2) show coefficients from OLS regressions of log hourly wages on college attendance. Columns (3)-(7) display coefficients from two-stage least squares models instrumenting college attendance with low earnings in an individual's county at 17 and the presence of a four-year public college in an individual's county at 14. Columns (3) and (4) show first stage specifications with and without the interaction of the two instruments. Column (5) displays the estimated second stage using the low local earnings variable as the excluded instrument, column (6) shows the second stage using the nearby college variable as the excluded instrument, and column (7) displays the second stage excluding both instruments. The second stage estimates are based on the additively separable first stage in column (3). Columns (2)-(7) also control for cohort indicators. Returns to college are expressed annually by dividing the college coefficient by 4. Robust standard errors in parentheses.

Table 6 reports descriptive statistics for key variables in the CHV sample. The outcome variable Y_i is the log of individual i 's average hourly wage from 1989 to 1993, and the treatment D_i equals one for individuals that attended college. Throughout our analysis we control for a vector X_i of covariates that includes AFQT scores, mother's years of education, number of siblings, urban residence at age 14, year of birth indicators, permanent local earnings (defined as average log earnings in the age 17 county of residence from 1973 to 2000), and average log earnings in the county of residence in 1991.¹⁶ The latter two variables account for general differences in earnings across counties, so that conditional on X_i the local earnings instrument can be interpreted as capturing idiosyncratic fluctuations in local labor market opportunities at the time of the college enrollment decision.

The bottom rows of Table 6 display descriptive statistics for the two instruments. To parallel the discussion of Section 4.2 we create a binary version of the local earnings instrument, labeled $Z_{i,1}$, which equals one if the residual from a regression of local earnings on X_i is in the bottom quartile of the sample.¹⁷ By construction, this definition means that 25% of youths experienced low local earnings at age 17. Fifty-three percent of individuals lived near a four-year college at age 14, which we label with a binary indicator, $Z_{i,2}$.

5.2 2SLS Estimates

We begin by reporting ordinary least squares (OLS) and 2SLS estimates of the returns to college. Following CHV, we divide all college coefficients by four to express returns on an annual basis. Column (1) of Table 7 displays the coefficient from a bivariate OLS regression of log wages on college attendance, which gives an annual return of 9%. Column 2 shows that controlling for the covariates and instruments reduces the college premium to 4% per year.

An extensive literature argues that OLS estimates of the returns to schooling may be biased by unobserved differences in ability (e.g., Card 2001). Motivated by this

¹⁶ We follow CHV and use a version of AFQT that corrects for differences in years of schooling at the time of the test. CHV also include labor market experience as an additional control variable in the second stage of their estimation procedure. Because experience may partially mediate the effects of college enrollment we instead control for age (captured by the year of birth indicators) and interpret the effects of college attendance inclusive of any impacts on experience.

¹⁷ In results not reported here, we found that other thresholds for discretizing the local earnings instrument yielded similar results but were less predictive of college attendance.

concern, columns 3-7 of Table 7 report 2SLS estimates of the system:

$$D_i = \psi + \Pi_1 Z_{i,1} + \Pi_2 Z_{i,2} + X_i' \delta + \eta_i, \quad (9)$$

$$Y_i = \alpha + \beta D_i + X_i' \gamma + \epsilon_i. \quad (10)$$

We compare 2SLS estimates from models excluding the instruments from the second stage one at a time, adding the other to the set of controls X_i in equation (10) as in Section 4.5, as well as a model that excludes both instruments from the second stage simultaneously.

The first stage estimates show that both instruments boost college attendance. Column (3) of Table 7 reports OLS estimates of equation (9), and column (4) shows estimates from a saturated specification that adds an interaction between the two instruments. The interaction term is small and statistically insignificant, and its value implies that the partial effect of each instrument is always positive, indicating that the ordering of the propensity score is consistent with (7) from our discussion of two binary instruments in Section 4.2. In view of the insignificant interaction term, we maintain the more parsimonious additively separable model from column (3) going forward.

This specification shows that youths experiencing low local earnings at age 17 are 10 percentage points more likely to attend college conditional on the controls, and this estimate is highly statistically significant. In the framework of Section 4.2, this implies that 10% of the population complies with $Z_{i,1}$, including eager compliers, reluctant compliers, and Z_1 compliers. Youths that live near a four-year college at age 14 are 6 percentage points more likely to attend college, an estimate that is also statistically significant. The larger first stage for the local earnings instrument implies that the share of Z_1 compliers exceeds the share of Z_2 compliers ($\pi_{1c} > \pi_{2c}$). However, the levels of these shares are not identified, so we have no empirical basis for assuming that $\pi_{2c} = 0$.

Columns (5) and (6) of Table 7 reveal that the two instruments generate very different estimates of the returns to college. Treating the low local earnings variable as the excluded instrument yields a positive and significant 2SLS estimate of 19.8% per year, while excluding the nearby college variable produces an insignificant estimated return of negative 2.4%. These two estimates are outside the typical range of such estimates (usually around 10%; see Card 1999), but both estimates are also statistically imprecise.

As can be seen in column (7), excluding both instruments from the second stage yields a more reasonable and precisely estimated return of 13.7% per year. At the same time, the standard rationale for this procedure relies on the strong restrictions on

choice behavior underlying Assumption IAM. Moreover, the overidentification test for the combined model is rejected at marginal significance levels ($p = 0.08$), suggesting heterogeneity in returns across subpopulations affected by each instrument. This raises the possibility that violations of Assumption IAM could contaminate the combined 2SLS estimates.

5.3 Assessing IA Monotonicity

We explore violations of Assumption IAM by comparing estimates of marginal treatment effects (MTEs) for the two instruments (Heckman and Vytlacil, 1999, 2005, 2007). The MTEs are based on the model

$$D_i(z) = \mathbb{1}[p(z, X_i) \geq U_i], \quad (11)$$

$$(Y_i(1), Y_i(0), U_i) \perp\!\!\!\perp Z_i | X_i, \quad (12)$$

where U_i is distributed uniformly, conditional on X_i and Z_i . Vytlacil (2002) showed that the model defined by (11) and (12) is equivalent to the framework of Imbens and Angrist (1994). In particular, the separable threshold-crossing representation of potential treatment choices in equation (11) is equivalent to Assumption IAM, with the homogeneous ordering of preferences over the instruments represented by the one-dimensional unobservable U_i .

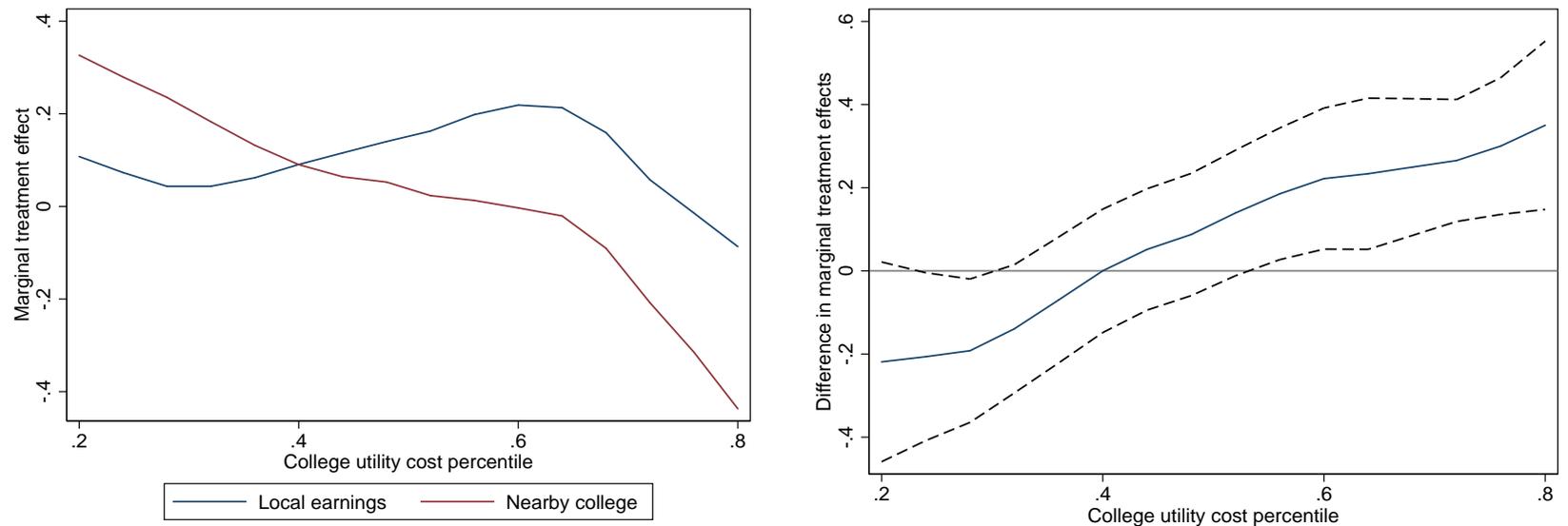
Heckman and Vytlacil (2005) define the MTE as the average treatment effect at a percentile of the unobserved cost U_i and value of the covariates X_i , that is,

$$\text{MTE}(u, x) \equiv \mathbb{E}[Y_i(1) - Y_i(0) | U_i = u, X_i = x].$$

They show that if there is continuous variation in the probability of treatment, then the MTEs are identified by derivatives of the average outcomes with respect to the propensity score: $\text{MTE}(u, x) = \partial \mathbb{E}[Y_i | p(Z_i, X_i) = u, X_i = x] / \partial u$. A core implication of Assumption IAM (or the equivalent (11)) is that the MTE schedule is invariant to which instrument is used to calculate these derivatives. In contrast, Assumption PM allows for the possibility that each instrument is associated with a different unobserved cost of treatment and its own MTE function (Mogstad et al., 2020). Comparing MTEs generated by different instruments therefore provides a test of Assumption IAM.¹⁸

¹⁸ This observation is based on Mogstad et al. (2020), who develop MTE methods for multiple IVs that enable testing Assumption IAM and aggregating information across instruments under Assumption PM.

Figure 4: Marginal treatment effects of college attendance based on local earnings and nearby college instruments



A. Marginal treatment effects for each instrument

B. Difference in marginal treatment effects

Notes: This figure plots estimates of marginal treatment effects (MTEs) of a year of college attendance on log wages using low average log earnings in an individual's county at age 17 (local earnings) and the presence of a college in an individual's county at age 14 (nearby college) as instruments. Panel A plots separate MTE estimates based on each instrument, and panel B plots the difference in estimated MTEs with 90 percent confidence intervals based on 1,000 bootstrap replications. MTE estimates are estimated derivatives of the conditional mean log wage with respect to the propensity score evaluated at the mean of the controls. The predicted probability of college attendance comes from a logit regression of college attendance on the two instruments and controls. The conditional mean of log wages is estimated based on local regressions of log wages on a quadratic function of the estimated propensity score and controls. Local regressions are evaluated at increments of 0.04 from 0.2 to 0.8 using a triangle kernel and bandwidth of 0.32. The controls in both steps include quadratic functions of AFQT, mother's education, number of siblings, permanent local earnings, and local earnings in 1991, along with urban status and cohort indicators. AFQT and mother's education are interacted with the instruments in the logit model and with the linear propensity score term in the outcome regressions. The MTE estimates using a given instrument include the other instrument and its interactions with the linear propensity score term, AFQT, and mother's education as additional controls.

Since MTEs are not nonparametrically identified when the instruments are discrete, we use parametric models for $p(Z_i, X_i)$ and $\mathbb{E}[Y_i|p(Z_i, X_i), X_i]$ to estimate MTE curves separately for each instrument. This amounts to extrapolating from the LATEs identified by each instrument to compare treatment effects at a common value of the propensity score, which must be the same if Assumption IAM holds.¹⁹ We estimate MTEs by first fitting a logit model for $p(Z_i, X_i)$, then running local regressions of Y_i on X_i and a quadratic function of the estimated propensity score, weighted with a triangle kernel function maximized at the target value of u . We allow for covariate heterogeneity by interacting the linear propensity score term with elements of X_i . The MTE estimates are derivatives of the fitted conditional mean function evaluated at the sample average of the covariates. Paralleling the 2SLS models in columns (4) and (5) of Table 7, we exclude each instrument separately while including the other as a control variable, thereby producing separate MTE estimates based on the separate variation in each instrument. We focus on MTEs for values of u from 0.2 to 0.8 because estimates at extreme values of u were very imprecise. Inference is conducted based on 1,000 bootstrap replications of the entire procedure.

The results reveal that the two instruments generate different MTE schedules. Panel A of Figure 4 shows that the local earnings instrument yields a relatively flat MTE curve, suggesting limited variation in treatment effects as a function of unobserved responsiveness to labor market opportunity costs. In contrast, the MTE curve for the nearby college instrument is sharply downward sloping, implying substantial treatment effect heterogeneity along the distance dimension. As a result, MTEs based on the nearby college instrument tend to be larger at low values of u and smaller at high values of u compared to those based on the local earnings instrument. The bootstrap confidence intervals displayed in Panel B show that the differences in MTE schedules are statistically significant at low and high values of u , and a joint test rejects equality of the two sets of MTEs at conventional levels ($p = 0.046$). Subject to the modeling restrictions required to estimate MTEs with discrete instruments, these results indicate that Assumption IAM does not hold jointly for the local earnings and nearby college instruments.

5.4 Implications for Combined 2SLS Estimates

Our results so far suggest that Assumption IAM may be violated for the combination of instruments considered here. At the same time, estimates based on the individual

¹⁹ See Brinch et al. (2017), Mogstad, Santos, and Torgovitsky (2018), Mogstad and Torgovitsky (2018), and Kline and Walters (2019) for related discussion of extrapolation based on discrete instruments.

Table 8: *Testing for positive 2SLS weights*

	Dependent variable:	
	College attendance (1)	Nearby college (2)
Low local earnings at 17	0.119 (0.024)	0.102 (0.024)
Nearby college at 14	0.071 (0.023)	1.000 -
<i>P</i> -value: Positive weights		1.000
<i>P</i> -value: Negative weights		0.001

Notes: This table displays regressions of the variable listed in each column on the variable listed in each row in the NLSY sample. All models control for covariates. Robust standard errors are in parentheses. The first p -value comes from a test of the null hypothesis that the 2SLS weights are all positive, and the second comes from a test of the null hypothesis that at least one weight is negative.

instruments are imprecise, so it may still be useful to combine these instruments via 2SLS. How do violations of Assumption IAM affect the interpretation of the combined 2SLS results in Table 7?

We answer this question by applying Propositions 5 and 6 to the CHV sample. Table 8 reports coefficients from regressions of D_i on each instrument separately along with the coefficient from a regression of $Z_{i,2}$ on $Z_{i,1}$. These models also control for X_i . Column (1) demonstrates that controlling for the covariates (but not the other instrument), the correlation between each instrument and the treatment is positive and statistically significant. The formulas in Proposition 5 then imply that the weights for each complier group must be positive under Assumption AM even if both Z_1 and Z_2 compliers are present. Similarly, column (2) shows that the partial correlation between the two instruments is also positive. By Proposition 6, this implies that the 2SLS weights are positive even if Assumption IAM is violated. The joint distribution of the two instruments therefore turns out to be sufficient to yield positive weights in this case.

The final rows of Table 8 use these results in formal statistical tests for positive and negative 2SLS weights. We consider tests of both the null hypothesis that the weights are all positive, and the complementary hypothesis that at least one weight is negative.²⁰ As expected given the strong positive correlations in Table 8, the null hy-

²⁰ The tests, which were implemented using our companion Stata package `mivcausal`, are described in more detail in Appendix C. The p -value reported for the null of positive weights is from the test based on Romano et al. (2014).

pothesis of negative weights is rejected at conventional levels ($p = 0.001$), while a test of the null of positive weights does not reject and generates a high p -value of 1.0. These findings show that in the CHV application, the combined 2SLS estimand retains an interpretation as a positively-weighted average of treatment effects when Assumption IAM is replaced by the weaker Assumption AM. This has been emphasized elsewhere as an attractive property of estimands in settings with treatment effect heterogeneity (e.g. Angrist and Pischke, 2009). Our results demonstrate that if estimating a positive weighted average of treatment effects is the researcher’s goal, then they may be justified in combining instruments via 2SLS without relying on the onerous homogeneity restrictions implied by IA monotonicity.

6 Conclusion

The IA monotonicity condition is a cornerstone of modern IV analysis. It is appealed to often, but rarely justified explicitly. As we have shown, it will not hold when there are multiple IVs without severe restrictions on choice heterogeneity. This creates a dilemma for using IV methods to aggregate findings into a larger body of knowledge: Each instrument is associated with a different set of complier groups, but combining multiple IVs together using IA monotonicity requires assuming that these groups are effectively identical in terms of their choice behavior.

In this paper, we have clarified the implications of IA monotonicity and considered combining instruments under a weaker condition called partial monotonicity. Partial monotonicity is not a strong assumption about choice behavior. It is satisfied under the usual mathematical interpretation of monotonicity that each instrument encourages all individuals either towards or away from treatment. We have shown that it still preserves the interpretation of the 2SLS estimand as a positively-weighted average of causal effects for complier groups, except in rare cases. These rare cases can and should be checked for when reporting 2SLS estimates with multiple IVs. The conditions for positive weights turn out to hold in our application to estimating the returns to college attendance, even while the IA monotonicity condition seems to be clearly violated.

While our results provide theoretical and empirical guidance for researchers who wish to use 2SLS to combine multiple IVs, it is important to recognize that positively-weighted averages of LATEs do not necessarily answer interesting scientific or policy counterfactuals. This point motivates our companion paper (Mogstad et al., 2020), in which we develop a framework for aggregating multiple IVs to conduct inference about specific target parameters that do answer well-posed counterfactual questions. The framework generalizes the approach of Mogstad et al. (2018) to replace IA monotonic-

ity with partial monotonicity. The key idea is that even under partial monotonicity, each instrument still carries identifying content about unobserved quantities that are instrument-invariant, such as the average treatment effect. Harmonizing this content across instruments allows for information to flow between instruments and to aggregate into greater whole. This provides a way to combine exogenous variation from multiple sources for policy evaluation, while still maintaining plausible conditions on choice behavior.

A Proofs

Proof of Proposition 1. (\Rightarrow) Suppose that the nesting statement is not true. Then there exist $j, k \in \mathcal{I}$ such that $\mathcal{Z}_j \not\subseteq \mathcal{Z}_k$ and $\mathcal{Z}_k \not\subseteq \mathcal{Z}_j$. Since the empty set \emptyset is a subset of every set (including itself), this implies that both \mathcal{Z}_j and \mathcal{Z}_k are not empty. Thus, there exists a $z_j \in \mathcal{Z}_j$ such that $z_j \notin \mathcal{Z}_k$, and there exists a $z_k \in \mathcal{Z}_k$ such that $z_k \notin \mathcal{Z}_j$. By the definition of these sets, this means that

$$\begin{aligned} D_j(z_j) &= 1 > 0 = D_j(z_k) \\ \text{and } D_k(z_j) &= 0 < 1 = D_k(z_k). \end{aligned} \tag{13}$$

Thus, $D_i(z_j) \geq D_i(z_k)$ for some $i = j$, but $D_i(z_j) < D_i(z_k)$ for some other $i = k$. This violates Assumption IAM. By contraposition, it follows that Assumption IAM implies the nesting statement.

(\Leftarrow) Conversely, if Assumption IAM is not true, then there exist $j, k \in \mathcal{I}$ and $z_j, z_k \in \mathcal{Z}$ such that (13) holds. By definition, (13) implies that $z_j \in \mathcal{Z}_j$, but $z_j \notin \mathcal{Z}_k$, and that $z_k \in \mathcal{Z}_k$, but $z_k \notin \mathcal{Z}_j$. That is, $\mathcal{Z}_k \not\subseteq \mathcal{Z}_j$, and $\mathcal{Z}_j \not\subseteq \mathcal{Z}_k$. It follows that the nesting statement also implies Assumption IAM. *Q.E.D.*

Proof of Proposition 2. Suppose to the contrary that there exist $j, k \in \mathcal{I}(z^*)$ for which (2) does not hold. Then the matrix

$$\partial V_{jk}(z^*) \equiv \begin{bmatrix} \partial_1 V_j(z^*) & \partial_2 V_j(z^*) \\ \partial_1 V_k(z^*) & \partial_2 V_k(z^*) \end{bmatrix} \equiv \begin{bmatrix} \partial V_j(z^*) \\ \partial V_k(z^*) \end{bmatrix}$$

is invertible. Thus, the span of the rows of $\partial V_{jk}(z^*)$ is \mathbb{R}^2 , so there exists a unit vector $v^* \in \mathbb{R}^2$ such that $\partial V_j(z^*)v^* > 0$, while $\partial V_k(z^*)v^* < 0$. Taking a Taylor series expansion at $z^* + \epsilon v^*$ for sufficiently small $\epsilon > 0$, we have that

$$\begin{aligned} V_j(z^* + \epsilon v^*) &\approx V_j(z^*) + \epsilon [\partial V_j(z^*)v^*] > 0, \\ \text{while } V_k(z^* + \epsilon v^*) &\approx V_k(z^*) + \epsilon [\partial V_k(z^*)v^*] < 0 \end{aligned}$$

since $V_j(z^*) = V_k(z^*) = 0$. On the other hand, an ϵ step in the direction $-v^*$ yields

$$V_j(z^* - \epsilon v^*) < 0 \quad \text{while} \quad V_k(z^* - \epsilon v^*) > 0.$$

Using the random utility model, (1), we have that

$$D_j(z^* + \epsilon v^*) = 1 > D_j(z^* - \epsilon v^*) = 0$$

and $D_k(z^* + \epsilon v^*) = 0 < D_k(z^* - \epsilon v^*) = 1,$

which shows that Assumption IAM is violated. This establishes the result through contraposition. *Q.E.D.*

Proof of Proposition 3. Take any $(z_\ell, z_{-\ell})$ and $(z'_\ell, z_{-\ell})$ in \mathcal{Z} . Since $z_\ell, z'_\ell \in \mathbb{R}$, either $z_\ell \geq z'_\ell$ or $z'_\ell \geq z_\ell$. Suppose that the first case holds. Then $(z_\ell, z_{-\ell}) \geq (z'_\ell, z_{-\ell})$, so $D_i(z_\ell, z_{-\ell}) \geq D_i(z'_\ell, z_{-\ell})$ for all i , as required by Assumption PM. *Q.E.D.*

Proof of Proposition 4. That an individual i cannot belong to more than one of the six groups can be verified by inspection. To see that i must belong to at least one of these groups, note that under (7), either

$$D_i(0, 0) \leq D_i(0, 1) \leq D_i(1, 0) \leq D_i(1, 1) \tag{14}$$

or $D_i(0, 0) \leq D_i(1, 0) \leq D_i(0, 1) \leq D_i(1, 1).$ (15)

If i satisfies (14) then by Table 3, their group is $G_i \in \{\text{at, nt, 1c, ec, rc}\}$; that is, something other than a Z_2 complier. If i satisfies (15), then their group is something other than a Z_1 complier. In either case, they must belong to one of the six groups listed in Table 3. *Q.E.D.*

Proof of Proposition 5. Label the instrument pairs as $z^1 = (0, 0)$, $z^2 = (0, 1)$, $z^3 = (1, 0)$, and $z^4 = (1, 1)$, and denote their associated probabilities as $q^k \equiv \mathbb{P}[Z_i = z^k]$ for $k = 1, 2, 3, 4$. We will prove the result for the case with $\pi_{1c} \geq \pi_{2c}$, so that the propensity score $p(z^k) \equiv \mathbb{P}[D_i = 1 | Z_i = z^k]$ is increasing in k , due to Assumption PM and (7). A symmetric proof applies to the case with $\pi_{2c} \geq \pi_{1c}$.

Theorem 2 in Imbens and Angrist (1994) shows that the 2SLS estimand is given by a convex weighted average of three Wald (1940) estimands, which we write as

$$\beta_{2\text{sls}} = \lambda_2 w_2 + \lambda_3 w_3 + \lambda_4 w_4, \tag{16}$$

where the Wald estimands are given by

$$w_k \equiv \frac{\mathbb{E}[Y_i | Z_i = z^k] - \mathbb{E}[Y_i | Z_i = z^{k-1}]}{p(z^k) - p(z^{k-1})},$$

and the weights are defined by

$$\lambda_k \equiv \frac{(p(z^k) - p(z^{k-1})) \sum_{\ell=k}^4 q^\ell (p(z^\ell) - \mathbb{E}[p(Z_i)])}{\sum_{j=2}^4 [(p(z^j) - p(z^{j-1})) \sum_{\ell=j}^4 q^\ell (p(z^\ell) - \mathbb{E}[p(Z_i)])]}.$$

Theorem 1 of Imbens and Angrist (1994) shows that each Wald estimand, w_k , gives the average treatment effect for individuals who change treatment status in response to a change in the instrument from z^{k-1} to z^k . Using the group definitions in Proposition 4, this implies that w_2 represents the average treatment effect for both the Z_2 compliers and eager compliers. Similarly, w_4 reflects the average treatment effect for the Z_2 compliers and the reluctant compliers. So,

$$w_2 = \left(\frac{\pi_{2c}}{\pi_{2c} + \pi_{ec}} \right) \Delta_{2c} + \left(\frac{\pi_{ec}}{\pi_{2c} + \pi_{ec}} \right) \Delta_{ec},$$

and

$$w_4 = \left(\frac{\pi_{2c}}{\pi_{2c} + \pi_{rc}} \right) \Delta_{2c} + \left(\frac{\pi_{rc}}{\pi_{2c} + \pi_{rc}} \right) \Delta_{rc}.$$

However, w_3 is different, since a shift from $z^2 \equiv (0, 1)$ to $z^3 \equiv (1, 0)$ creates two-way flows. In particular, such a shift induces Z_1 compliers to take treatment, but Z_2 compliers to exit treatment. Using a minor modification of the argument in Imbens and Angrist (1994), it follows that

$$\begin{aligned} w_3 &= \frac{\mathbb{E} [(Y_i(1) - Y_i(0))(D_i(1, 0) - D_i(0, 1))]}{p(z^3) - p(z^2)} \\ &= \left(\frac{\pi_{1c}}{p(z^3) - p(z^2)} \right) \Delta_{1c} - \left(\frac{\pi_{2c}}{p(z^3) - p(z^2)} \right) \Delta_{2c} \\ &= \left(\frac{\pi_{1c}}{\pi_{1c} - \pi_{2c}} \right) \Delta_{1c} - \left(\frac{\pi_{2c}}{\pi_{1c} - \pi_{2c}} \right) \Delta_{2c}, \end{aligned}$$

where the last equality used $p(z^3) - p(z^2) = (\pi_{1c} + \pi_{ec}) - (\pi_{2c} + \pi_{ec}) = \pi_{1c} - \pi_{2c}$.

Substituting the expressions for the Wald estimands into (16), we have

$$\beta_{2sls} = \sum_{g \in \{1c, 2c, ec, rc\}} \omega_g \Delta_g,$$

where

$$\begin{aligned} \omega_{ec} &\equiv \frac{\lambda_2 \pi_{ec}}{\pi_{ec} + \pi_{2c}}, & \omega_{1c} &\equiv \frac{\lambda_3 \pi_{1c}}{\pi_{1c} - \pi_{2c}}, \\ \omega_{rc} &\equiv \frac{\lambda_4 \pi_{rc}}{\pi_{rc} + \pi_{2c}}, & \text{and } \omega_{2c} &\equiv \left(\frac{\lambda_2 \pi_{2c}}{\pi_{ec} + \pi_{2c}} \right) - \left(\frac{\lambda_3 \pi_{2c}}{\pi_{1c} - \pi_{2c}} \right) + \left(\frac{\lambda_4 \pi_{2c}}{\pi_{rc} + \pi_{2c}} \right). \end{aligned}$$

It is straightforward to verify that ω_{ec}, ω_{1c} , and ω_{rc} are each non-negative, and that

$$\sum_{g \in \{1c, 2c, ec, rc\}} \omega_g = \lambda_2 + \lambda_3 + \lambda_4 = 1.$$

For ω_{2c} , note first that

$$\pi_{ec} + \pi_{2c} = p(z^2) - p(z^1),$$

and that, similarly,

$$\pi_{1c} - \pi_{2c} = p(z^3) - p(z^2) \quad \text{and} \quad \pi_{rc} - \pi_{2c} = p(z^4) - p(z^3).$$

Substituting this observation and the definition of λ_k into the expression for ω_{2c} and simplifying, we have

$$\omega_{2c} = \pi_{2c} \times \frac{q^2 (p(z^2) - \mathbb{E}[p(Z_i)]) + q^4 (p(z^4) - \mathbb{E}[p(Z_i)])}{\sum_{j=2}^4 \left[(p(z^j) - p(z^{j-1})) \sum_{\ell=j}^4 q^\ell (p(z^\ell) - \mathbb{E}[p(Z_i)]) \right]}.$$

The denominator of this expression is always positive and π_{2c} is always non-negative. For the numerator, notice that since $Z_{i,2} = 1$ if and only if $Z_i \in \{z^2, z^4\}$,

$$\begin{aligned} & q^2 (p(z^2) - \mathbb{E}[p(Z_i)]) + q^4 (p(z^4) - \mathbb{E}[p(Z_i)]) \\ &= \mathbb{E} [Z_{i,2} (p(Z_i) - \mathbb{E}[p(Z_i)])] \\ &= \mathbb{E} [Z_{i,2} (D_i - \mathbb{E}[D_i])] \equiv \text{Cov}(D_i, Z_{i,2}), \end{aligned}$$

where the second equality follows by iterating expectations. Thus, the sign of ω_{2c} is the same as that of $\text{Cov}(D_i, Z_{i,2})$, which is in turn the same as the sign of $\mathbb{E}[D_i|Z_{i,2} = 1] - \mathbb{E}[D_i|Z_{i,2} = 0]$, since

$$\text{Cov}(D_i, Z_{i,2}) = (\mathbb{E}[D_i|Z_{i,2} = 1] - \mathbb{E}[D_i|Z_{i,2} = 0]) \mathbb{P}[Z_{i,2} = 1] \mathbb{P}[Z_{i,2} = 0].$$

Q.E.D.

Proof of Proposition 6. Since $Z_{i,1}$ and $Z_{i,2}$ are binary,

$$\begin{aligned} & \mathbb{P}[D_i = 1|Z_{i,2} = 1] - \mathbb{P}[D_i = 1|Z_{i,2} = 0] \\ &= p(1, 1) \mathbb{P}[Z_{i,1} = 1|Z_{i,2} = 1] + p(0, 1) \mathbb{P}[Z_{i,1} = 0|Z_{i,2} = 1] \\ & \quad - p(1, 0) \mathbb{P}[Z_{i,1} = 1|Z_{i,2} = 0] - p(0, 0) \mathbb{P}[Z_{i,1} = 0|Z_{i,2} = 0]. \end{aligned}$$

Assumption PM with (7) together with E imply that

$$p(1, 1) \equiv \mathbb{P}[D_i = 1 | Z_{i,1} = 1, Z_{i,2} = 1] = \mathbb{P}[D_i(1, 1) = 1] \geq \mathbb{P}[D_i(0, 1) = 1] = p(0, 1),$$

and similarly that $p(1, 0) \geq p(0, 0)$ and $p(0, 1) \geq p(0, 0)$. If $\text{Cov}(Z_{i,1}, Z_{i,2}) \geq 0$, then also $\mathbb{P}[Z_{i,1} = 1 | Z_{i,2} = 1] \geq \mathbb{P}[Z_{i,1} = 1] \geq \mathbb{P}[Z_{i,1} = 1 | Z_{i,2} = 0]$. Thus,

$$\begin{aligned} & \Pr[D_i = 1 | Z_{i,2} = 1] - \Pr[D_i = 1 | Z_{i,2} = 0] \\ & \geq p(1, 1) \mathbb{P}[Z_{i,1} = 1] + p(0, 1) (1 - \mathbb{P}[Z_{i,1} = 1]) \\ & \quad - p(1, 0) \mathbb{P}[Z_{i,1} = 1] - p(0, 0) (1 - \mathbb{P}[Z_{i,1} = 1]) \\ & = [p(1, 1) - p(1, 0)] \mathbb{P}[Z_{i,1} = 1] + [p(0, 1) - p(0, 0)] (1 - \Pr[Z_{i,2} = 1]) \geq 0. \end{aligned}$$

By Proposition 5, this implies that $\omega_{2c} \geq 0$. A symmetric argument shows that $\mathbb{P}[D_i = 1 | Z_{i,1} = 1] - \mathbb{P}[D_i = 1 | Z_{i,1} = 0] \geq 0$, so that $\omega_{1c} \geq 0$ as well. *Q.E.D.*

Proof of Proposition 7. By Theorem 2 in Imbens and Angrist (1994),

$$\beta_{2\text{sls}} = \sum_{k=2}^K \lambda_k w_k, \tag{17}$$

where the Wald estimands are

$$w_k \equiv \frac{\mathbb{E}[Y_i | Z_i = z^k] - \mathbb{E}[Y_i | Z_i = z^{k-1}]}{p(z^k) - p(z^{k-1})},$$

and the weights are

$$\lambda_k \equiv \frac{(p(z^k) - p(z^{k-1})) \sum_{\ell=k}^K q^\ell (p(z^\ell) - \mathbb{E}[p(Z_i)])}{\sum_{j=2}^K [(p(z^j) - p(z^{j-1})) \sum_{\ell=j}^K q^\ell (p(z^\ell) - \mathbb{E}[p(Z_i)])]}. \tag{18}$$

By Assumption E,

$$\begin{aligned} w_k &= \frac{\mathbb{E}[Y_i(D_i(z^k)) - Y_i(D_i(z^{k-1}))]}{p(z^k) - p(z^{k-1})} \\ &= \frac{\sum_{g \in \mathcal{G}} \mathbb{E}[Y_i(D_i(z^k)) - Y_i(D_i(z^{k-1})) | G_i = g] \pi_g}{p(z^k) - p(z^{k-1})} \\ &= \frac{\sum_{g: k \in \mathcal{C}_g} \Delta_g \pi_g - \sum_{g: k \in \mathcal{D}_g} \Delta_g \pi_g}{p(z^k) - p(z^{k-1})}, \end{aligned} \tag{19}$$

since $Y_i(D_i(z^k)) - Y_i(D_i(z^{k-1})) = 0$ except when $k \in \mathcal{C}_{G_i}$ or $k \in \mathcal{D}_{G_i}$. Substituting

(19) into (17),

$$\begin{aligned}
\beta_{2\text{sIs}} &= \sum_{k=2}^K \lambda_k \left(\frac{\sum_{g:k \in \mathcal{C}_g} \Delta_g \pi_g - \sum_{g:k \in \mathcal{D}_g} \Delta_g \pi_g}{p(z^k) - p(z^{k-1})} \right) \\
&= \sum_{k=2}^K \lambda_k \left(\frac{\sum_{g \in \mathcal{G}} (\mathbb{1}[k \in \mathcal{C}_g] - \mathbb{1}[k \in \mathcal{D}_g]) \Delta_g \pi_g}{p(z^k) - p(z^{k-1})} \right) \\
&= \sum_{g \in \mathcal{G}} \left(\pi_g \sum_{k=2}^K (\mathbb{1}[k \in \mathcal{C}_g] - \mathbb{1}[k \in \mathcal{D}_g]) \left(\frac{\lambda_k}{p(z^k) - p(z^{k-1})} \right) \right) \Delta_g \equiv \sum_{g \in \mathcal{G}} \omega_g \Delta_g. \quad (20)
\end{aligned}$$

Substituting the definition of λ_k from (18) and simplifying,

$$\begin{aligned}
\omega_g &= \pi_g \sum_{k=2}^K (\mathbb{1}[k \in \mathcal{C}_g] - \mathbb{1}[k \in \mathcal{D}_g]) \left(\frac{\sum_{\ell=k}^K q^\ell (p(z^\ell) - \mathbb{E}[p(Z_i)])}{\sum_{j=2}^K [(p(z^j) - p(z^{j-1})) \sum_{\ell=j}^K q^\ell (p(z^\ell) - \mathbb{E}[p(Z_i)])]} \right) \\
&= \pi_g \sum_{k=2}^K (\mathbb{1}[k \in \mathcal{C}_g] - \mathbb{1}[k \in \mathcal{D}_g]) \left(\frac{\text{Cov}(D_i, \mathbb{1}[p(Z_i) \geq p(z^k)])}{\text{Var}(p(Z_i))} \right), \quad (21)
\end{aligned}$$

where in the numerator we used

$$\begin{aligned}
\text{Cov}(D_i, \mathbb{1}[p(Z_i) \geq p(z^k)]) &= \mathbb{E} \left[\mathbb{1}[p(Z_i) \geq p(z^k)] (D_i - \mathbb{E}[D_i]) \right] \\
&= \mathbb{E} \left[\mathbb{1}[p(Z_i) \geq p(z^k)] (p(Z_i) - \mathbb{E}[p(Z_i)]) \right] \\
&= \sum_{\ell=k}^K q^\ell \left(p(z^\ell) - \mathbb{E}[p(Z_i)] \right),
\end{aligned}$$

and in the denominator we used

$$\begin{aligned}
\text{Var}(p(Z_i)) &= \mathbb{E} [D_i (p(Z_i) - \mathbb{E}[p(Z_i)])] \\
&= \sum_{\ell=1}^K p(z^\ell) \left(p(z^\ell) - \mathbb{E}[p(Z_i)] \right) q^\ell \\
&= \sum_{\ell=1}^K \left(\sum_{j=2}^K \mathbb{1}[j \leq \ell] (p(z^j) - p(z^{j-1})) \right) \left(p(z^\ell) - \mathbb{E}[p(Z_i)] \right) q^\ell \\
&= \sum_{j=2}^K \left[(p(z^j) - p(z^{j-1})) \sum_{\ell=j}^K \left(p(z^\ell) - \mathbb{E}[p(Z_i)] \right) q^\ell \right],
\end{aligned}$$

where the third equality follows from a telescoping sum identity,²¹ together with the

²¹ In particular, that $a^\ell = a^1 + \sum_{j=2}^K \mathbb{1}[j \leq \ell] (a^j - a^{j-1})$ for any scalars $\{a^\ell\}_{\ell=1}^K$.

fact that $\sum_{\ell=1}^K (p(z^\ell) - \mathbb{E}[p(Z_i)]) q^\ell = 0$. Examining the expression for the weights in (21), we have that

$$\text{sgn}(\omega_g) = \mathbb{1}[\pi_g > 0] \times \text{sgn} \left(\sum_{k=2}^K (\mathbb{1}[k \in \mathcal{C}_g] - \mathbb{1}[k \in \mathcal{D}_g]) \text{Cov} \left(D_i, \mathbb{1}[p(Z_i) \geq p(z^k)] \right) \right).$$

It remains to show that $\omega_g = 0$ when $\mathcal{C}_g = \emptyset$, so that only groups that comply with at least one instrument contrast receive weight in the 2SLS estimand. To see that this is so, suppose to the contrary that there is a group g with $\pi_g > 0$ for which $\mathcal{C}_g = \emptyset$, while $\omega_g \neq 0$. Given the structure of ω_g , such a group must have $\mathcal{D}_g \neq \emptyset$. That is, this group must defy at some instrument contrast, even though they do not comply at any other instrument contrasts. We will prove that such a “pure defier” group cannot exist under Assumption PM by establishing a contradiction.

Let $j_0 \in \mathcal{D}_g$ be the instrument contrast at which the “pure defier” group g defies. By definition, $D_i(z^{j_0}) = 0$, while $D_i(z^{j_0-1}) = 1$. Since $\mathcal{C}_g = \emptyset$, it follows that for any i with $G_i = g$,

$$D_i(z^j) = \mathbb{1}[j < j_0]. \quad (22)$$

In particular, $D_i(z^1) = 1$, while $D_i(z^K) = 0$.

To proceed, it will be helpful to use the following terminology. We call two vectors z^j and z^k *pm-comparable* if they differ in only one component. That is, z^j and z^k are pm-comparable if there exists an $\ell' \in \{1, \dots, L\}$ such that $z_{\ell'}^j = z_{\ell'}^k$ for all $\ell \neq \ell'$. If z^j and z^k are pm-comparable, then Assumption PM requires that either $D_i(z^j) \leq D_i(z^k)$ for all $i \in \mathcal{I}$ or that $D_i(z^j) \geq D_i(z^k)$ for all $i \in \mathcal{I}$. Moreover, as we show in Lemma 1, if $j \leq k$, then there cannot exist a group g^* with $\pi_{g^*} > 0$ for which individuals i in group g^* have $D_i(z^j) > D_i(z^k)$. We will now use this result to show that the existence of the “pure defier” group g defined above creates a contradiction.

Let z^{j_1} be the vector whose first component is the same as that of the largest propensity-score instrument value, z^K , while all other components are the same as the smallest, z^1 . That is,

$$z^{j_1} \equiv (z_1^K, z_{-1}^1).$$

Then z^{j_1} and z^1 are pm-comparable, and $z^{j_1} \in \text{supp}(Z_i)$, which we have assumed is rectangular. Since $D_i(z^1) = 1$ for any i with $G_i = g$ and $p(z^1)$ is the smallest propensity score value, it follows from Lemma 1 that $D_i(z^{j_1}) = 1$ for these individuals as well. Thus by (22), it must be that $j_1 < j_0$.

Now let z^{j_2} be the same as z^{j_1} except with its second component replaced by z_2^K . That is,

$$z^{j_2} \equiv (z_2^K, z_{-2}^{j_1}) \equiv (z_1^K, z_2^K, z_3^1, \dots, z_L^1).$$

Then z^{j_2} is pm-comparable to z^{j_1} , and $z^{j_2} \in \text{supp}(Z_i)$. If it were the case that $j_2 \geq j_0$, then $p(z^{j_2}) \geq p(z^{j_0}) \geq p(z^{j_1})$, so that Lemma 1 would imply that $D_i(z^{j_2}) = 1$ for individuals with $G_i = g$. At the same time, (22) would imply that $D_i(z^{j_2}) = 0$ for these individuals, yielding a contradiction. Thus, it must be that $j_2 < j_0$.

Continuing in this way, we find a sequence of vectors $z^{j_1}, z^{j_2}, z^{j_3}, \dots, z^{j_L}$ that each differ from z^K in one component less than its predecessor, and such that $j_\ell < j_0$ for each ℓ . This process ends once we reach j_L , at which point $z^{j_L} = z^K$ is the instrument value corresponding to the largest propensity score value. However, this implies a contradiction, because $j_L < j_0$ while at the same time $j_L = K \geq j_0$. We conclude that a “pure defier” group cannot exist under Assumption PM, and therefore that the sum in (20) only needs to be indexed over $g \in \mathcal{G}$ for which $\mathcal{C}_g \neq \emptyset$. *Q.E.D.*

Lemma 1. *Suppose that Assumption E holds. Let \mathcal{G} denote the set of all realizations of $\{D_i(z)\}_{z \in \mathcal{Z}}$ that are consistent with Assumption PM. Suppose that z and z' are pm-comparable and that $p(z) \leq p(z')$. Then there does not exist a group $g^* \in \mathcal{G}$ such that $\pi_{g^*} > 0$ and $D_i(z) > D_i(z')$ for all i with $G_i = g^*$.*

Proof of Lemma 1. Since z and z' are pm-comparable, Assumption PM requires that $D_i(z) \leq D_i(z')$ for all $i \in \mathcal{I}$, or $D_i(z) \geq D_i(z')$ for all $i \in \mathcal{I}$. If such a group g^* did exist, then the latter case would need to hold. However, this would imply that

$$\begin{aligned} p(z) &\equiv \mathbb{P}[D_i = 1 | Z_i = z] \\ &= \mathbb{P}[D_i(z) = 1] \\ &= \mathbb{P}[D_i(z) = 1 | G_i = g^*] \pi_{g^*} + \mathbb{P}[D_i(z) = 1 | G_i \neq g^*] (1 - \pi_{g^*}) \\ &> \mathbb{P}[D_i(z') = 1 | G_i = g^*] \pi_{g^*} + \mathbb{P}[D_i(z') = 1 | G_i \neq g^*] (1 - \pi_{g^*}) \\ &= \mathbb{P}[D_i = 1 | Z_i = z'] \equiv p(z'), \end{aligned}$$

which contradicts the assumption that $p(z) \leq p(z')$.

Q.E.D.

B Additional Examples of Assumptions IAM and PM

In this appendix, we consider three example papers from our survey, each of which combines multiple economically distinct IVs using 2SLS. In each example, we discuss

the content of both Assumptions IAM and PM. We conclude in each that Assumption IAM is unlikely to hold, whereas the weaker Assumption PM is more plausible.

B.1 Thornton (2008)

The author evaluates an experiment in rural Malawi in which individuals who were screened for HIV were then randomly assigned incentives to receive their results. Two incentives were used: a cash-redeemable voucher and the distance to the nearest results center. Thornton uses these randomly assigned incentives as instruments for the decision to obtain the results, with the goal of estimating the causal effect of learning HIV status on the demand for condoms.

Assumption IAM would require an individual's decision to obtain their HIV test results to either be influenced more by the voucher, or by the distance to the results center. It is not difficult to imagine two individuals who similarly value learning their HIV status, but who differ in their preferences over distance versus a monetary incentive. For example, a low-wage worker with a lower opportunity cost of time might be more affected by a voucher than by the distance to the center. On the other hand, a high-wage worker with a higher opportunity cost might value the reduced travel time to a closer center more than a voucher.

While Assumption IAM monotonicity is hard to justify, Assumption PM is likely to hold. Keeping distance fixed, a voucher incentive should increase the likelihood that any individual obtains his or her results. Keeping the voucher incentive fixed, being randomly assigned a closer results facility should decrease the opportunity cost and thus also increase the likelihood of obtaining the test results.

B.2 Currie and Moretti (2003)

The authors examine the impact of a mother's educational attainment on her infant's health. As instruments for mother's education, the authors use the number of two- and four-year colleges present in the mother's county of residence when she was 17 years old. The argument is that the availability of college may have induced some women to obtain more education.

Assumption IAM would require all mothers to be influenced more by the presence of a two-year college or by the presence of a four-year college when deciding how much education to obtain. This is unlikely. For example, two-year colleges may provide a cheaper option, while four-year colleges may provide a better education. There is likely to be substantial heterogeneity in preferences over the cost and quality of education.

However, Assumption PM is a more reasonable assumption. A mother's educa-

tional attainment should increase as her opportunities for education increase. If this is true, then the *ceteris paribus* impact of the presence of a four-year college on attained education would be positive regardless of whether there is also a two-year college, and vice-versa. This is all that is required for Assumption PM to be satisfied.

B.3 Dippel (2014)

The author considers the long-term effects on economic growth of forcibly integrating Native American communities in the United States. As instruments for integration, the author uses the value of gold and silver mining activity in these communities. The rationale behind these instruments is that when land was found to be more valuable, the federal government would free up larger portions of the land by forming fewer, more concentrated reservations.

Assumption IAM would require the likelihood of forced integration across all tribes and reservations to either be affected more by the value of silver activity or by the value of gold activity. This would not hold if the officials responsible for forced integration had different beliefs about the value of each metal. This could occur if these beliefs vary by the location of the reservation or by the official responsible, among other reasons. In contrast, Assumption PM follows the logic of the author’s argument: holding the value of one metal fixed, increasing the value of the other increases the likelihood of forced integration.

C Tests about the Signs of the 2SLS Weights

C.1 Estimation

Suppose that we exclude a priori the possibility that $\pi_{1c} = 0$ or $\pi_{2c} = 0$. Then Proposition 5 implies that the signs of the 2SLS weights in the case with two binary instruments are determined by the signs of

$$\theta_j(x) \equiv \mathbb{P}[D_i = 1 | Z_{i,j} = 1, X_i = x] - \mathbb{P}[D_i = 1 | Z_{i,j} = 0, X_i = x] \quad \text{for } j = 1, 2,$$

where we have conditioned on covariates, X_i , since they are included in our application in Section 5. To develop our tests, we assume that the conditional probability of treatment given $Z_{i,j}$ and X_i is additively separable in $Z_{i,j}$. This implies that $\theta_j(x) \equiv \theta_j$ does not depend on x , and that θ_j is identified as the population regression coefficient on $Z_{i,j}$ in a regression of D_i on $Z_{i,j}$ and X_i .²²

²² It is straightforward to extend the following to test joint hypotheses about $\theta_j(x)$ across different pre-specified x values. It is also possible to test joint hypotheses about $\theta_j(X_i)$ as a random variable using tools

We estimate $\theta \equiv (\theta_1, \theta_2)$ with two ordinary least squares estimators, $\hat{\theta} \equiv (\hat{\theta}_1, \hat{\theta}_2)$, based on a sample of size n . We assume that these estimators are jointly asymptotically normal and denote the limiting variance–covariance matrix of $\sqrt{n}(\hat{\theta} - \theta)$ by Σ . Let $\hat{\Sigma}$ be a consistent estimator of Σ . Let σ_1^2 and σ_2^2 be the diagonal components of Σ , let σ_{12} be the off–diagonal component, and denote the corresponding components of $\hat{\Sigma}$ by $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$, and $\hat{\sigma}_{12}$, respectively.

C.2 Testing the Null Hypothesis of Positive Weights

We first consider tests of the null hypothesis that the weights are positive (non-negative), that is of

$$H_0^+ : \theta_1 \geq 0 \quad \text{and} \quad \theta_2 \geq 0,$$

against the complementary alternative.

The first and simplest approach is to treat $\theta_1 \geq 0$ and $\theta_2 \geq 0$ as separate hypotheses and then apply a Bonferroni correction. Letting \hat{p}_1 and \hat{p}_2 denote the p –values from the corresponding one–sided t –tests, the Bonferroni–corrected p –value is then

$$\hat{p}_0 \equiv \min\{2\hat{p}_1, 2\hat{p}_2, 1\}.$$

This test will typically be conservative.

The second approach is to consider the test statistic

$$\hat{T} \equiv \min_{j=1,2} \frac{\sqrt{n}\hat{\theta}_j}{\hat{\sigma}_j}, \tag{23}$$

that is, the minimum of the individual t –statistics, and reject H_0^+ if this quantity is too small. If H_0^+ is true, then the distribution of \hat{T} has the following lower bound asymptotically:

$$\mathbb{P} \left[\hat{T} \leq t \right] = \mathbb{P} \left[\min_{j=1,2} \frac{\sqrt{n}(\hat{\theta}_j - \theta_j)}{\hat{\sigma}_j} + \frac{\sqrt{n}\theta_j}{\hat{\sigma}_j} \leq t \right] \leq \mathbb{P} \left[\min_{j=1,2} W_j \leq t \right], \tag{24}$$

where $W \equiv (W_1, W_2)$ is a bivariate normal distribution with mean zero, unit variances, and correlation $\sigma_{1,2}\sigma_1^{-1}\sigma_2^{-1}$. Thus, the test that rejects H_0^+ when \hat{T} is smaller than the

from the literature on conditional moment inequalities (e.g. Andrews and Shi, 2013; Chernozhukov, Lee, and Rosen, 2013; Armstrong, 2015; Chetverikov, 2018). We focus on the separable case as it enables the construction of simple tests which can be implemented easily in statistical software, and which require no additional choices or tuning parameters on the part of the researcher.

α quantile of the distribution of $\min_{j=1,2} W_j$ has size no greater than α . In the Monte Carlo simulation in Section C.4, we refer to this test as the “Mintest.” Implementing it requires simulating the distribution of $\min_{j=1,2} W_j$ using its estimated correlation, $\hat{\sigma}_{12}\sigma_1^{-1}\hat{\sigma}_2^{-1}$.

The third approach uses the quasi-likelihood ratio statistic

$$\hat{Q} = \min_{t \geq 0} n \left(\hat{\theta} - t \right)' \hat{\Sigma}^{-1} \left(\hat{\theta} - t \right),$$

and rejects if \hat{Q} is too large. Let \hat{t}^* denote the minimizer of this problem and let $\hat{k}^* \in \{0, 1, 2\}$ denote the number of components of \hat{t}^* that are zero, that is, where the non-negativity constraint is binding. Cox and Shi (2019) show that the test that rejects when \hat{Q} is larger than the $1 - \alpha$ quantile of a chi-squared distribution with \hat{k}^* degrees of freedom controls size at level α . We call this the Cox-Shi test in Section C.4.

The fourth test is from Romano et al. (2014, “RSW”), and also uses the test statistic \hat{T} from (23). Their approach improves on the Mintest described above by estimating a $1 - \beta$ joint confidence interval for $\min_{j=1,2} \sqrt{n}(\hat{\theta}_j - \theta_j)$ and using this to improve the coarse bound used in (24) to obtain a critical value. Both this first step and their resulting critical value requires bootstrapping the linear regression estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$. For a level α test, RSW recommend setting $\beta = \alpha/10$, which means that the number of bootstraps used in the first step confidence interval needs to be rather large to get an accurate approximation of the $1 - \beta$ quantile. This can make the RSW test somewhat computationally demanding compared to the other three tests.

C.3 Testing the Null Hypothesis of Negative Weights

We also consider the opposite null hypothesis that one or more 2SLS weight is negative, that is, of

$$H_0^- : \theta_1 \leq 0 \quad \text{or} \quad \theta_2 \leq 0,$$

against the complementary alternative. We use an intersection-union test (IUT) based on the theory in Berger (1982). In the current context, the IUT argument is simple: reject H_0^- at level α if both $\theta_1 \leq 0$ and $\theta_2 \leq 0$ are rejected at level α using one-sided t -tests. This controls size because the probability that both $\theta_1 \leq 0$ and $\theta_2 \leq 0$ are rejected under the null is by construction smaller than the probability that either $\theta_1 \leq 0$ or $\theta_2 \leq 0$ are rejected. Perhaps more surprisingly, Berger (1982, Theorem 2) provides conditions under which the IUT test is size-correct, which is confirmed in our simulations. See Berger and Hsu (1996, Section 3) and Casella and Berger (2002,

Section 8.3) for more detail.

C.4 A Monte Carlo Simulation

The Monte Carlo simulation has the following data generating process. The group shares are set at $\pi_{\text{at}} = 2/12$, $\pi_{\text{ec}} = 1/12$, $\pi_{\text{rc}} = 1/12$, $\pi_{\text{nt}} = 2/12$, $\pi_{1\text{c}} = 5/12$, and $\pi_{2\text{c}} = 1/12$. First, $Z_{i,2} \in \{0, 1\}$ is drawn with probability 1/2. Then, $Z_{i,1}$ is drawn conditional on $Z_{i,2}$ with probability

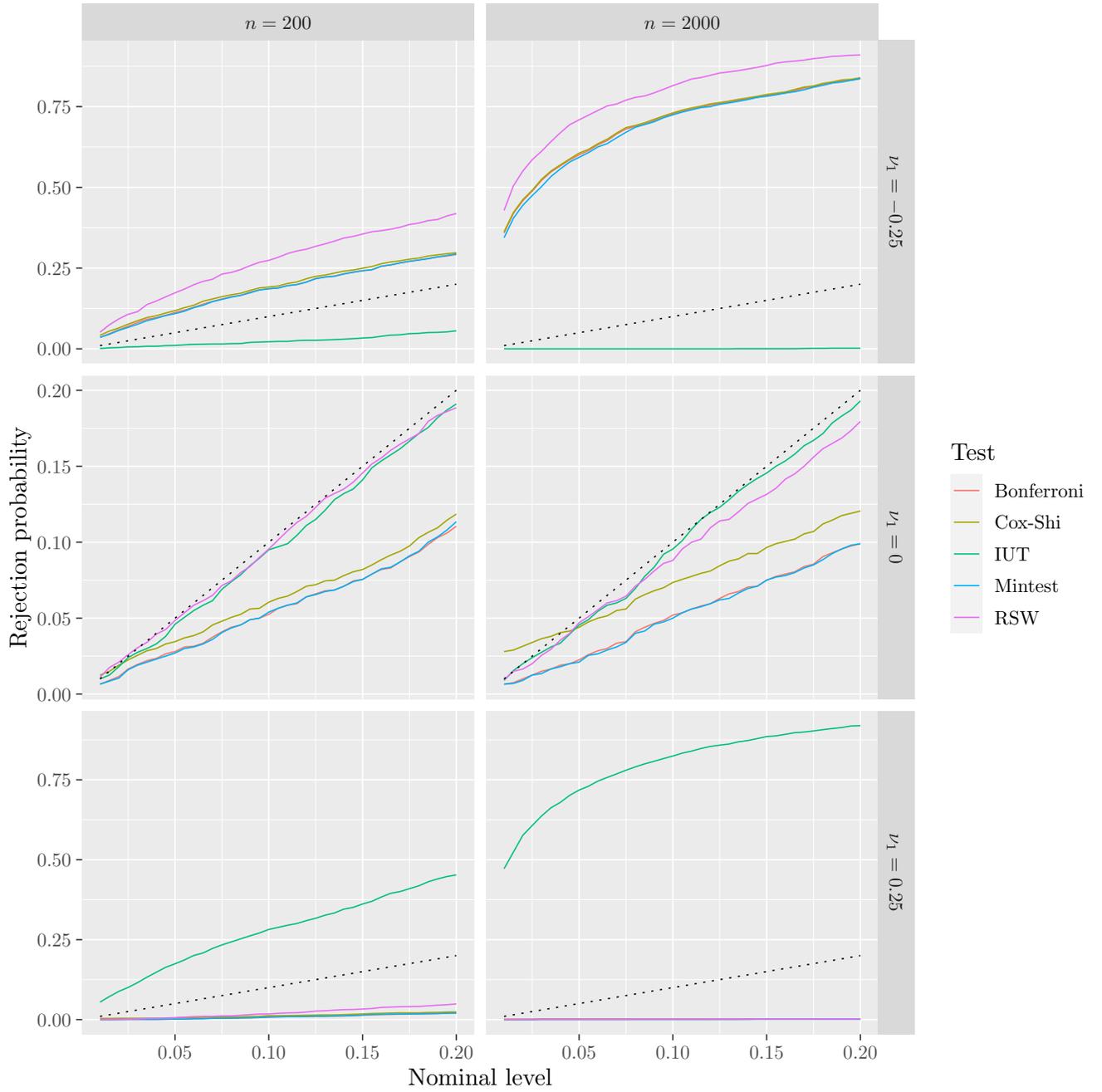
$$\mathbb{P}[Z_{i,1} = 1 | Z_{i,2} = z_2] = \Phi(\nu_0(1 - z_2) + \nu_1 z_2),$$

where Φ is the standard normal distribution function, and ν_0, ν_1 are design parameters. The parameter ν_0 is set such that H_0^+ is true if and only if $\nu_1 \geq 0$, H_0^- is true if and only if $\nu_1 \leq 0$, and both null hypotheses are true when $\nu_1 = 0$.

Figure C.1 shows QQ-plots of the p-values from our various tests against the uniform distribution for three values of ν_1 .²³ The middle row with $\nu_1 = 0$ shows that all tests control size when both H_0^+ and H_0^- are true, with the IUT and RSW tests being closest to size-correct. When $\nu_1 = -.25$, so that H_0^- is true, the three simpler-to-implement tests (Bonferroni, Mintest, Cox-Shi) all have roughly the same power, while the RSW test is substantially more powerful. Power for the IUT test when $\nu_1 = .25$ is difficult to gauge, since there is no point of comparison, but one would expect that it is quite good given its performance at the boundary of the null hypothesis ($\nu_1 = 0$). Figure C.2 reports power curves for a 5% level test which confirm the superior power of the RSW test.

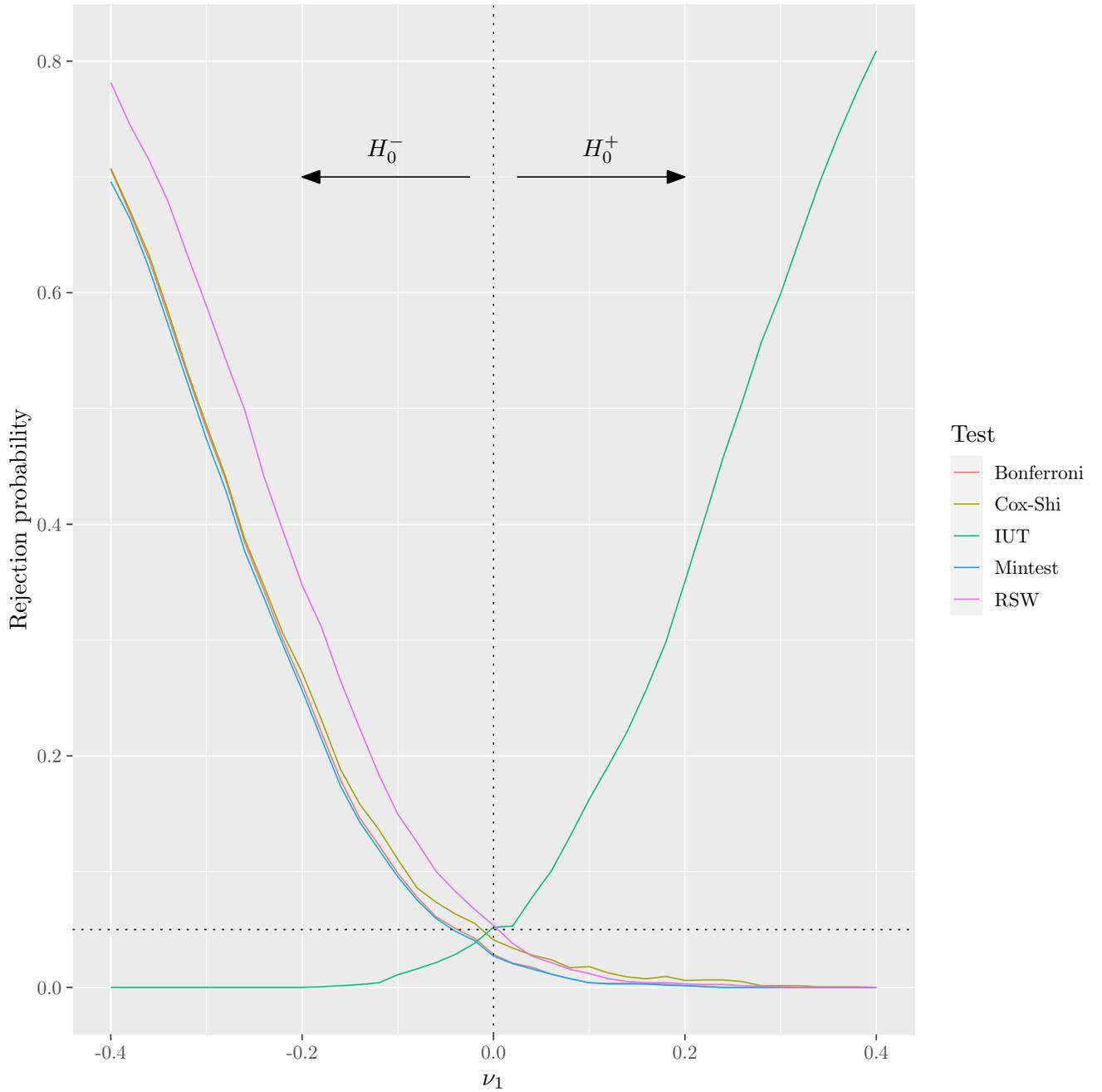
²³ All simulations are based on 2,000 replications. We used 2,000 bootstrap draws for the RSW test.

Figure C.1: Size and power for five tests



The dotted line is the 45 degree line. The top row ($\nu_1 = -0.25$) is where H_0^- is true, the bottom row is where H_0^+ is true, and both hypothesis are true in the middle row.

Figure C.2: Power curves for five tests



The dotted horizontal line indicates the nominal level of .05. The dotted vertical line indicates the boundary between where H_0^+ and H_0^- are true. The sample size is $n = 1000$.

References

- ANDREWS, D. W. K. AND X. SHI (2013): “Inference Based on Conditional Moment Inequalities,” *Econometrica*, 81, 609–666. 41
- ANGRIST, J. D. AND W. N. EVANS (1998): “Children and Their Parents’ Labor Supply: Evidence from Exogenous Variation in Family Size,” *American Economic Review*, 88, 450–477. 11
- ANGRIST, J. D. AND G. W. IMBENS (1995): “Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity,” *Journal of the American Statistical Association*, 90, 431–442. 5
- ANGRIST, J. D., G. W. IMBENS, AND D. B. RUBIN (1996): “Identification of Causal Effects Using Instrumental Variables,” *Journal of the American Statistical Association*, 91, 444–455. 13
- ANGRIST, J. D. AND A. B. KRUEGER (1991): “Does Compulsory School Attendance Affect Schooling and Earnings?” *The Quarterly Journal of Economics*, 106, 979–1014. 4
- ANGRIST, J. D. AND J.-S. PISCHKE (2009): *Mostly Harmless Econometrics*, Princeton University Press. 1, 29
- ARMSTRONG, T. B. (2015): “Asymptotically Exact Inference in Conditional Moment Inequality Models,” *Journal of Econometrics*, 186, 51–65. 41
- BERGER, R. L. (1982): “Multiparameter Hypothesis Testing and Acceptance Sampling,” *Technometrics*, 24, 295–300. 42
- BERGER, R. L. AND J. C. HSU (1996): “Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets,” *Statistical Science*, 11, 283–319. 42
- BITLER, M., H. HOYNES, AND T. DOMINA (2014): “Experimental Evidence on Distributional Effects of Head Start,” Tech. rep. 1
- BITLER, M. P., J. B. GELBACH, AND H. W. HOYNES (2006): “What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments,” *The American Economic Review*, 96, 988–1012. 1
- BRINCH, C. N., M. MOGSTAD, AND M. WISWALL (2017): “Beyond LATE with a Discrete Instrument,” *Journal of Political Economy*, 125, 985–1039. 1, 27
- CAMERON, S. V. AND J. J. HECKMAN (1998): “Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males,” *Journal of Political Economy*, 106, 262–333. 21
- CAMERON, S. V. AND C. TABER (2004): “Estimation of Educational Borrowing Constraints Using Returns to Schooling,” *Journal of Political Economy*, 112, 132–182. 21
- CARD, D. (1995): “Using Geographic Variation in College Proximity to Estimate the Return to Schooling,” in *Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp*, ed. by L. N. Christofides, E. K. Grant, and R. Swidinsky, University of Toronto Press, 201–222. 8, 21

- (1999): “The causal effect of education on earnings,” in *Handbook of Labor Economics*, ed. by O. Ashenfelter and D. Card, Elsevier Science, vol. 3, 1801–1863. 24
- (2001): “Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems,” *Econometrica*, 69, 1127–1160. 23
- CARNEIRO, P., J. J. HECKMAN, AND E. J. VYTLACIL (2011): “Estimating Marginal Returns to Education,” *American Economic Review*, 101, 2754–81. 1, 2, 4, 19, 20
- CARNEIRO, P. AND S. LEE (2009): “Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality,” *Journal of Econometrics*, 149, 191–208. 1
- CARNEIRO, P., M. LOKSHIN, AND N. UMAPATHI (2016): “Average and Marginal Returns to Upper Secondary Schooling in Indonesia,” *Journal of Applied Econometrics*, 32, 16–36. 1
- CASELLA, G. AND R. L. BERGER (2002): *Statistical Inference*, Pacific Grove, Calif. [u.a.]: Duxbury/Thomson Learning. 42
- CHERNOZHUKOV, V., S. LEE, AND A. M. ROSEN (2013): “Intersection Bounds: Estimation and Inference,” *Econometrica*, 81, 667–737. 41
- CHETVERIKOV, D. (2018): “ADAPTIVE TESTS OF CONDITIONAL MOMENT INEQUALITIES,” *Econometric Theory*, 34, 186–227. 41
- CORNELISSEN, T., C. DUSTMANN, A. RAUTE, AND U. SCHÖNBERG (forthcoming): “Who benefits from universal childcare? Estimating marginal returns to early childcare attendance,” *Journal of Political Economy*. 1
- COX, G. AND X. SHI (2019): “A Simple Uniformly Valid Test for Inequalities,” *arXiv:1907.06317 [econ, math, stat]*. 42
- CURRIE, J. AND E. MORETTI (2003): “Mother’s Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings*,” *The Quarterly Journal of Economics*, 118, 1495–1532. 39
- DIPPEL, C. (2014): “Forced Coexistence and Economic Development: Evidence From Native American Reservations: Forced Coexistence and Economic Development,” *Econometrica*, 82, 2131–2165. 40
- DOYLE JR., J. J. (2007): “Child Protection and Child Outcomes: Measuring the Effects of Foster Care,” *The American Economic Review*, 97, 1583–1610. 1
- FELFE, C. AND R. LALIVE (2014): “Does Early Child Care Help or Hurt Children’s Development?” Tech. Rep. 8484. 1
- FIRPO, S., N. M. FORTIN, AND T. LEMIEUX (2009): “Unconditional Quantile Regressions,” *Econometrica*, 77, 953–973. 1
- FRENCH, E. AND J. SONG (2014): “The Effect of Disability Insurance Receipt on Labor Supply,” *American Economic Journal: Economic Policy*, 6, 291–337. 1
- HAVNES, T. AND M. MOGSTAD (2015): “Is universal child care leveling the playing field?” *Journal of Public Economics*, 127, 100–114. 1

- HECKMAN, J. J. (2001): “Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture,” *The Journal of Political Economy*, 109, 673–748. 1
- HECKMAN, J. J., S. URZUA, AND E. VYTLACIL (2006): “Understanding Instrumental Variables in Models with Essential Heterogeneity,” *Review of Economics and Statistics*, 88, 389–432. 1, 5, 10
- HECKMAN, J. J. AND E. VYTLACIL (2005): “Structural Equations, Treatment Effects, and Econometric Policy Evaluation,” *Econometrica*, 73, 669–738. 1, 6, 25
- HECKMAN, J. J. AND E. J. VYTLACIL (1999): “Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects,” *Proceedings of the National Academy of Sciences of the United States of America*, 96, 4730–4734. 25
- (2007): “Chapter 71 Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments,” in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. Volume 6, Part 2, 4875–5143. 25
- IMBENS, G. W. AND J. D. ANGRIST (1994): “Identification and Estimation of Local Average Treatment Effects,” *Econometrica*, 62, 467–475. 1, 6, 12, 14, 18, 19, 25, 32, 33, 35
- KANE, T. J. AND C. E. ROUSE (1993): “Labor Market Returns to Two- and Four-Year Colleges: Is a Credit a Credit and Do Degrees Matter?” Working Paper 4268, National Bureau of Economic Research. 8
- KIRKEBOEN, L. J., E. LEUVEN, AND M. MOGSTAD (2016): “Field of Study, Earnings, and Self-Selection,” *The Quarterly Journal of Economics*, 131, 1057–1111. 1
- KLINE, P. AND C. WALTERS (2019): “On Heckits, LATE, and Numerical Equivalence,” *Econometrica*, 87, 677–696. 27
- KLINE, P. AND C. R. WALTERS (2016): “Evaluating Public Programs with Close Substitutes: The Case of Head Start*,” *The Quarterly Journal of Economics*, 131, 1795–1848. 1
- KLING, J. R. (2001): “Interpreting Instrumental Variables Estimates of the Returns to Schooling,” *Journal of Business and Economic Statistics*, 19, 358–364. 21
- MAESTAS, N., K. J. MULLEN, AND A. STRAND (2013): “Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt,” *The American Economic Review*, 103, 1797–1829. 1
- MILGROM, P. AND C. SHANNON (1994): “Monotone comparative statics,” *Econometrica*, 157–180. 11
- MOFFITT, R. (2008): “Estimating Marginal Treatment Effects in Heterogeneous Populations,” *Annales d’Economie et de Statistique*, 239–261. 1
- MOGSTAD, M., A. SANTOS, AND A. TORGOVITSKY (2018): “Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters,” *Econometrica*, 86, 1589–1619. 27, 29
- MOGSTAD, M. AND A. TORGOVITSKY (2018): “Identification and Extrapolation of Causal Effects with Instrumental Variables,” *Annual Review of Economics*, 10. 27

- MOGSTAD, M., A. TORGOVITSKY, AND C. WALTERS (2020): “Policy Evaluation with Multiple Instrumental Variables,” Working paper. 19, 25, 29
- MOUNTJOY, J. (2019): “Community Colleges and Upward Mobility,” *Working paper*. 10
- NYBOM, M. (2017): “The Distribution of Lifetime Earnings Returns to College,” *Journal of Labor Economics*, 000–000. 1
- ROMANO, J. P., A. M. SHAIKH, AND M. WOLF (2014): “A Practical Two-Step Method for Testing Moment Inequalities,” *Econometrica*, 82, 1979–2002. 17, 28, 42
- THORNTON, R. L. (2008): “The Demand for, and Impact of, Learning HIV Status,” *American Economic Review*, 98, 1829–63. 39
- VYTLACIL, E. (2002): “Independence, Monotonicity, and Latent Index Models: An Equivalence Result,” *Econometrica*, 70, 331–341. 7, 25
- WALD, A. (1940): “The Fitting of Straight Lines if Both Variables are Subject to Error,” *The Annals of Mathematical Statistics*, 11, 284–300. 32
- WALTERS, C. (2014): “The Demand for Effective Charter Schools,” Tech. rep. 1