Post-Instrument Bias[*]

Julian Schuessler[†] Adam N. Glynn[‡] Miguel R. Rueda[§]

March 2021

Abstract

When using instrumental variables, researchers often assume that causal effects are only identified conditional on covariates. We show that the role of these covariates in applied research is often unclear and that there exists confusion regarding their ability to mitigate violations of the exclusion restriction. We explain when and how existing adjustment strategies may lead to bias. We then discuss assumptions that are sufficient to identify various treatment effects, some of which are new, when the exclusion restriction only holds conditionally. In general, these assumptions are highly restrictive, albeit they sometimes are testable. We also show that other existing tests are generally misleading. Then, we introduce an alternative sensitivity analysis that uses information on variables influenced by the instrument to gauge the effect of potential violations of the exclusion restriction. We illustrate it in two replications of existing analyses and summarize our results in easy-to-understand guidelines.

Keywords: Causal inference, instrumental variables, exclusion restriction, sensitivity analysis
Word count: 10,448

[*] We would like to thank Winston Chou for valuable input and the idea for the sensitivity analysis, and Phillip Heiler, Caleb Koch, Sebastian Koehler, Max Reinwald, Peter Selb, Anselm Hager, and Macartan Humphreys for comments. Participants at the University of Konstanz, ETH Zurich, modSim 2017, DVPW Methods Section 2017, and EPSA 2018 provided helpful feedback. We also want to thank Abigail Heller and Elisha Cohen for excellent research assistance. Julian Schuessler acknowledges financial support from the German National Academic Foundation, the Graduate School of Decision Sciences, and the Department for Quantitative Theory and Methods.

[†] Corresponding author. Post-Doc, Institute for Political Science, Aarhus University. Email: julians@ps.au.dk.

[‡] Associate Professor. Department of Political Science and Quantitative Theory and Methods, Emory University. Email: adam.glynn@emory.edu

[§] Assistant Professor. Department of Political Science, Emory University. Email: miguel.rueda@emory.edu.

1 Introduction

Identification of causal effects using instrumental variables is a popular approach in both experimental and observational research, and recent decades have seen an increasingly sophisticated understanding of what effects such instruments may identify. Based on the seminal work by Angrist, Imbens and Rubin (1996), social scientists are nowadays aware of the role that assumptions such as the exclusion restriction or first-stage monotonicity play (Betz, Cook and Hollenbach 2018; Marshall 2016; Sovey and Green 2011). However, we contend that the choice of covariates in instrumental variable (IV) identification is not well-understood and leads to biases in applied research. Of special interest is the widespread adjustment for "post-instrument" variables to address a violation of the exclusion restriction, on which existing guidelines are either silent or contradictory.

In this paper, we give straightforward advice for researchers on how to think about covariates in the context of IV analysis and which of these need to be controlled for. To this end, we uncover significant new identification results and subtleties, including with regards to (partial) tests of identifying assumptions. Furthermore, we develop a semi-parametric sensitivity analysis that aids applied researchers when there is a direct effect of an instrument that runs over measured variables.

Our contribution is motivated by both the widespread practice and voiced concerns of researchers that use instrumental variables. We have identified 154 papers published since 2010 in top political science journals[1] that use IV and explicitly discuss the exclusion restriction. Among those, 39 use post-instrument covariates to justify the exclusion restriction. Such practice appears to have increased over time: while from 2010 to 2014 14.1% of the papers using IV applied this adjustment, 22.05% do so from 2015 onward.[2]

[1] The American Political Science Review, the American Journal of Political Science, and the Journal of Politics.

[2] The total number of papers that use instrumental variables is 205. For more details, see

However, some researchers are aware that adjustment for variables on other paths from instrument to outcome may not always lead to identification. For example, both Kern and Hainmueller (2009) and Carnegie and Marinov (2017) use instrumental variables and two-stage least-squares regression where they choose not (or not always) to control for such variables to avoid what they call "post-treatment bias". But there seems to be no justification for this in the literature, which uses this term for biases that are introduced in standard adjustment identification strategies, where instruments play no role (Rosenbaum 1984; Angrist and Pischke 2009; Montgomery, Nyhan and Torres 2018). On the other hand, Wucherpfennig, Hunziker and Cederman (2016), for example, claim that "the instrumental variable logic is immune to any correlation (and even causation) between the instruments and the covariates". A leading econometrics textbook similarly advises to simply control for covariates influenced by the instrument (Wooldridge 2010, 94, 938). Other standard textbooks like Angrist and Pischke (2009) and reader's guides like Sovey and Green (2011) do not discuss such issues.

To fix ideas, consider an example from Angrist (1990), whose identification strategy has inspired several studies of political behavior (see Berinsky and Chatfield 2015 for an overview). The author is interested in estimating the effect of serving in the Vietnam war on earnings. The draft was largely determined by a randomized lottery, and Angrist notes that men who have a low draft lottery number were more likely to serve in the war. He uses functions of this number as instruments for military service.

There could be some concerns about the validity of the exclusion restriction. For example, those who received a low lottery number could have chosen to stay in school to obtain a deferment (Angrist 1990, 330). This creates a link between the lottery and earnings via education. So if information on post-lottery education was available, should we control for it?

In this paper, we answer this question and discuss various related problems. We use both

Appendix A.

potential outcomes and directed acyclic graphs (Pearl 2009) in our formal analysis. This allows us to give advice to applied researchers that is both easy to formulate and understand. We first make clear the asymmetric role of pre- and post-instrumental variables. Then, we illustrate how adjustment for variables influenced by the instrument may not always be successful, and that adjustment for variables influenced by the *treatment* will lead to biases in IV identification even when the IV is unconditionally valid. The mechanics behind these phenomena resemble the better-known "post-treatment" bias in adjustment strategies (Montgomery, Nyhan and Torres 2018), although additional, more subtle problems occur. However, we also show, perhaps to the surprise of some researchers, that adjustment for variables influenced by the *instrument* is sometimes *necessary* for successful identification. In some cases, we show that this identifies the well-known "local" or a weighted average treatment effect. For other cases, we propose to identify a new, different treatment effect. In sum, "post-instrument bias" is quite different from "post-treatment bias".

The assumptions for valid post-instrument adjustment are highly restrictive, although we also prove that they are testable under some circumstances. In this context, we discuss the evidential value and implicit causal assumptions of other informal tests and robustness checks that are prevalent in the applied literature. We show that these tests are generally *misleading*. For example, checking the association between the instrument and the post-instrument variable is not sufficient, as the bias induced by the post-instrument variable also depends on the strength of the instrument with respect to the treatment. Therefore, we also add to the theory of robustness checks, which so far has concentrated on regression adjustment strategies (Lu and White 2014; Chen and Pearl 2015).

What if the strong assumption necessary for identification are not plausible or rejected by the data? We propose that researchers utilize measures of the variable on the pathway from the instrument to the outcome for a semi-parametric sensitivity analysis. Our approach generalizes previous approaches (Conley, Hansen and Rossi 2012; Van Kippersluis and Rietveld 2018) that operate under a strong effect homogeneity assumption and cannot use sample information to bound biases. Moreover, our approach also works if there is measurement error in the post-instrument variable. This will often be the case when potential violations of the exclusion restriction are uncovered only after initial data collection and intense scrutiny of an IV strategy. We illustrate our approach by reanalyzing the data of Spenkuch and Tillmann (2017) on the causal effect of Catholicism on the Nazi vote share at the end of the Weimar Republic, as well as the data of Carnegie and Marinov (2017) on the effect of foreign aid on human rights. The applications highlight the need to relax stringent linearity assumptions and to account for potential heterogeneity in causal effects.

A formal analysis of violations to the exclusion restriction was already provided in the seminal paper by Angrist, Imbens and Rubin (1996), but similar to Conley, Hansen and Rossi (2012) and Van Kippersluis and Rietveld (2018), this did not incorporate post-instrument variables. A more closely related paper is Deuchert and Huber (2017). They point out that investigating instruments that may affect more than one variable is also highly relevant because oftentimes the same instrument is used to study causal effects of different treatment variables so that researchers might be tempted to adjust for these other treatments. For example, Bazzi and Clemens (2013) discuss the "origin of a country's legal system" instrument that has been used for at least seven different treatments. Mellon (2020) points out that weather-related variables like measures of rainfall are often used as instruments for various relationships, but have been linked empirically to more than 100 other variables, each of which constitutes a potential measured violation to the exclusion restriction. Similar to our approach, Deuchert and Huber (2017) also use causal graphs. However, they use these for illustrative purposes only and prove their main results under a strong linearity assumption. In contrast, we discuss these issues in a completely nonparametric framework and integrate causal graphs with the popular potential outcomes approach. Importantly, we discuss additional identification assumptions, prove that these are sometimes testable, introduce a new causal estimand, and propose a new sensitivity analysis. Finally, some of the problems that we discuss are similar to what Elwert and Winship (2014) and Elwert and Segarra (2020) call "endogenous selection bias", and both Betz, Cook and Hollenbach (2018) and Imai and Kim (2019) also use causal graphs to illustrate failures of IV identification.

2 Understanding Conditional IV Identification Using Causal Graphs

In this section, we present a series of causal graphs that allow for the identification of various treatment effects when the key "ignorability" assumption only holds conditionally. We use causal graphs because they offer a straightforward formalization of the language already used by many researchers to communicate assumptions about the causal ordering of variables, direct and indirect effects, confounding, etc. Additionally, they can be integrated with the popular potential outcomes approach to causality, and allow for a derivation of assumptions on the distribution of these potential outcomes. Specifically, we interpret graphs as nonparametric structural equation models, as in Imai and Kim (2019). We expand on such formal aspects in Appendix B. For other recent uses of causal graphs in political science, see Imai et al. (2011), Imai and Yamamoto (2013), and Glynn and Kashin (2017).

2.1 A First Causal Graph for our Running Example

Consider again our example from Angrist (1990)'s seminal analysis. Angrist is interested in the causal effect of serving as a soldier in the Vietnam war (D_i) on later earnings Y_i. The draft lottery leads to a binary instrument Z_i that indicates draft eligibility.

The "ceiling" for the draft varied by year due to fluctuating demands by the military. Therefore, the cohort X_i of a man influenced the probability that he would be drafted. At the same time, birth year is clearly causally prior to the draft and might have other effects on the outcome. This can easily be depicted in a causal graph such as figure 1.

The dashed arrows emanating from the U_i-variable indicate that it stands for unobserved variables that may (directly) influence treatment, outcome, and covariates X_i, but not the instrument. In the Vietnam draft example, U_i may contain variables describing the socio-economic status of one's parents. These will impact the decision to enlist in the military, and on later socio-economic outcomes. They may also affect the timing of birth. The existence of such unobserved confounders is the central motivation for employing IV identification because they make identification of the effect of D_i on Y_i via regression or matching impossible. With this first example in mind, we now discuss basic quantities of interests and identification assumptions in the potential outcomes framework.

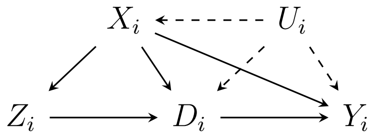

Figure 1: Benchmark graph. In this graph, Z_i is an instrument for the effect of D_i on Y_i conditional on X_i, but not unconditionally.

2.2 Basic IV Identification in the Potential Outcomes Framework

Generally, we will discuss the identification of variants of a local average treatment effect (LATE):

$$E[Y_i(D = 1) - Y_i(D_i = 0)|D_i(Z_i = 1) > D_i(Z_i = 0), X_i]$$

Here $Y_i(D = d)$ is the potential outcome of Y in unit i when D_i is set to d, and $D_i(Z_i = z)$ is the potential outcome of D in unit i when Z_i is set to z. Therefore, this expression defines the average causal effect of a binary treatment D_i on outcome Y_i among those individuals 1) for which an instrument Z_i changes treatment status (compliers) and 2) which are characterized by covariate profile X_i. Implicitly and throughout this paper, we assume that there are no spillovers, i.e., the treatment or instrument of one unit does not affect other unit's variables.

What if treatment is continuous, as is the case in our two application studies? First write the causal effect of instrument on treatment as $D_i(Z = 1) - D_i(Z = 0) = \alpha_i$. If the causal ("structural") equation of interest has heterogeneous effects, but otherwise is linear, as in

$$Y_i = \mu_Y + \beta_i D_i + \epsilon_i,$$

then the parameter of interest is usually (e.g., Angrist and Pischke (2009, 186–187))

$$\frac{E[\alpha_i \beta_i]}{E[\alpha_i]} = E\left[\frac{\alpha_i}{E[\alpha_i]}\beta_i\right]. \tag{1}$$

Here, $\frac{\alpha_i}{E[\alpha_i]}$ can be understood as individual-level weights of the treatment effects β_i.

Conventionally, three assumptions are used to identify such treatment effects. These are often discussed for the case of binary instrument and treatment, although they easily generalize. The first assumption, monotonicity, assumes that

$$P(D_i(Z_i = 1) \geq D_i(Z_i = 0)) = 1.$$

That is, the instrument has a causal effect on the treatment that pushes every unit in the same direction, and there are no "defiers". If this holds, $\alpha_i \geq 0$ so that the expression in equation 1 is a weighted average of individual-treatment effects β_i, where the weights are all greater than or equal to zero.

Secondly, it is assumed that Z_i and D_i are dependent ("relevance"):

$$E[D_i|Z_i = 1, X_i] - E[D_i|Z_i = 0, X_i] \neq 0$$

which is directly testable. In this paper, we will focus on understanding the crucial conditional independence assumption (CIA)

$$Y_i(D_i = d), D_i(Z_i = z) \perp\!\!\!\perp Z_i | X_i$$

In words, this assumptions states that the potential outcome of outcome Y_i when treatment D_i is set to d and the potential outcome of D_i when instrument Z_i is set to z are jointly independent from Z_i, given covariates X_i.

If these assumptions - CIA, monotonicity, and relevance - hold, two-stage least squares with saturated models in both stages estimates a weighted average of these X_i-specific LATEs, and this or linear unsaturated models are dominant in applied research (Angrist and Imbens 1995; Angrist and Pischke 2009, 177). Notably, the CIA subsumes both the exclusion restriction and the more opaque "ignorability" requirement. We use graphs to illustrate when this latter assumption hold, and will usually discuss the "causal first-stage" assumption $D_i(Z = z) \perp\!\!\!\perp Z_i | X_i$ separately from the $Y_i(D_i = d) \perp\!\!\!\perp Z_i | X_i$ requirement, since this is more intuitive. Formal derivations of the joint independence and other proofs are in Appendix C.

2.3 Identification with Pre-Instrument Covariates

We start with Figure 1 as a benchmark graph. In this graph, the treatment and outcome are driven by unobserved confounders U_i, while there are also observed confounders X_i that may influence the instrument, treatment, and outcome. A first important insight that this will not be the case when Z_i is physically and unconditionally randomized, because this precludes the $X_i \rightarrow Z_i$ path. However, if there are such observed confounders, adjustment for them is necessary. Intuitively, a first-stage regression of D_i on Z_i only would not give the causal effect of Z_i on D_i because of the open "back-door" paths $Z_i \leftarrow X_i \rightarrow D_i$ and $Z_i \leftarrow X_i \leftarrow U_i \rightarrow D_i$. Similarly, the instrument and the outcome would be connected through a path other than the effect going through D_i. Conditioning on X_i solves both problems, because X_i "blocks" these spurious paths.

Here, the CIA would not hold if at least one of two key conditions are violated. First, it may be that the confounders U_i also influence the instrument Z_i. In this case, Z_i and U_i are dependent (d-connected), and conditioning on X_i does not break this dependence. This is the problem of "back-door paths" which has found extensive treatment in the graphical literature (Shpitser, VanderWeele and Robins 2010). In fact, it suffices to have unobserved confounders that influence Z_i and D_i or Z_i and Y_i, but not necessarily all three variables, to invalidate the CIA. This fact has been overlooked even by very careful applied researchers that made their thinking about potential confounders explicit (e.g. Stanig 2015, 188).

Second, Z_i may have an effect on Y_i going not through D_i, which violates the "exclusion restriction". In this case, one can think of the potential outcomes as being determined by the equation (see Appendix B for a detailed discussion)

$$Y_i(D_i = d) = f_y(d, Z_i, X_i, U_i)$$

which clearly depends on Z_i, so that the CIA is violated.[3] Note, however, that $D_i(Z_i = z) \perp\!\!\!\perp Z_i | X_i$ may still hold if the exclusion restriction is violated, so that the effect of Z_i on D_i remains identified.

In the following, we will assume that observed pre-instrument covariates X_i may exist, and that conditioning on them solves the "back-door" problem. Specifically, this will even hold if U_i influences X_i (so that the effects of variables in X_i are not identified). This relaxes the common restriction for all X_i variables to be "exogenous" (e.g. Wooldridge 2010, 110), and differentiates such control variables from the post-instrument variables we discuss next. For ease of visual presentation, we will not depict the X_i nodes in the causal graphs that we discuss in the remainder of this article.

[3] Accordingly, the exclusion restriction is part of the CIA in our formulation, a point also made by Angrist and Pischke (2009, 132). In their guide to IV, Sovey and Green (2011, 191) state that the exclusion restriction is needed in addition to this CIA, but this is not accurate. It becomes clear when one thinks of potential outcomes as a summary of variables that influence Y_i when D_i if physically fixed.

2.4 Identification with Post-Instrument Covariates

We now discuss a variety of situations in which researchers measure covariates M_i that are influenced by the instrument, that influence the outcome, and that may also influence or be influenced by the treatment.[4] Our main result is that identification of a local average treatment effect is possible in some cases under strong assumptions. It turns out that identification relies on adjustment for the M_i covariates, even if they also influence the treatment. For the latter case, we introduce a new causal estimand and show how it is identified. Accordingly, "post-instrument" bias does not generally occur but depends on the causal model. Additionally, ruling out causation between D_i and M_i allows for a test of the identification assumptions which is easy to implement. We discuss other, informal tests in the literature and show that these are generally misleading.

In the Vietnam draft example, a potential M_i variable is college education, because the latter may have been used to avoid the draft, and because it plausibly affects earnings. The textbook by Wooldridge (2010, 938) discusses this complication and claims that statistical adjustment for such a variable M_i "effectively solves this problem". In the following, we show that this statement needs considerable qualification.

The most simple case is shown in graph (a) in figure 2, where the variable M_i is influenced by the instrument Z_i and in turn is a cause of Y_i. However, neither does D_i drive M_i, nor does M_i influence D_i, nor is U_i influencing M_i. Can we then simply control for the "post-instrument" variable M_i to make the instrumental variable approach work?

It turns that under the restrictive assumptions visualized in graph (a), this conditioning strategy indeed identifies an (X_i, M_i)-specific LATE or weighted ATE as in equation 1, since the CIA holds with conditioning set (X_i, M_i). To see why, consider the first-stage effect of Z_i on D_i. Although M_i is "post-instrument" - i.e., influenced by Z_i - conditioning on it does

[4] Our results only hold for *acyclic* graphs. This means that researchers need to rule out mutual causality between variables a priori.

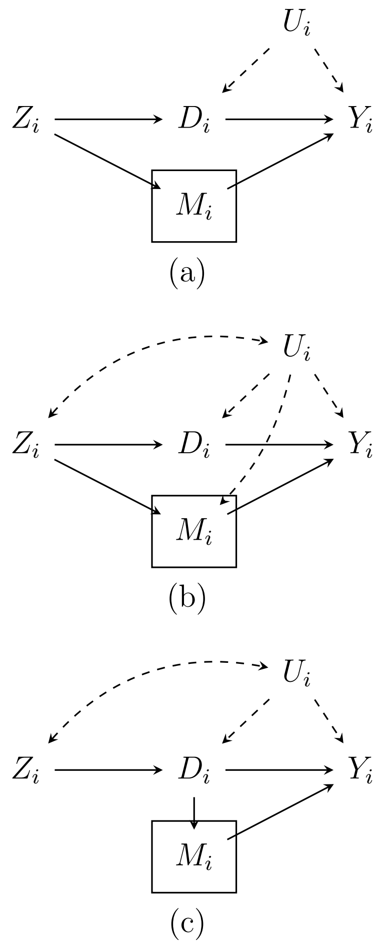

Figure 2: Boxes indicate the conditioning on M_i, and bi-directed arrows indicate dependencies created by such conditioning. In graph (a), conditioning on M_i is required and identifies the M_i-specific local effect of D_i on Y_i. In graph (b), conditioning on the collider M_i opens a non-causal path between U_i and Z_i. In graph (c), M_i is a descendant of collider D_i, and the same dependence by Z_i and U_i is created.

not invalidate the ignorability of Z_i with regards to D_i, i.e. $D_i(Z_i = z) \perp\!\!\!\perp Z_i | X_i, M_i$ holds. Intuitively, there is no "back-door" path from Z_i to D_i not blocked by X_i, and conditioning on M_i does not block any genuinely causal paths, nor does it open up any new spurious paths, since it is not a "collider". In a similar vein, the potential outcome $Y_i(D_i = d)$ is now determined by M_i, X_i, U_i as

$$Y_i(D_i = d) = f_y(d, M_i, X_i, U_i),$$

and is independent from Z_i conditional on M_i and X_i. This is because the direct path through M_i is blocked while no other paths are opened up.[5]

There are two crucial assumptions for the validity of this approach that may be violated. First, it may be that M_i is also driven by the unobserved confounder U_i. This situation is depicted in graph (b) of figure 2. In our running example, it is quite easy to imagine that unobserved parental SES positively influences the choice to go to college directly. In this case, M_i becomes a "collider", and conditioning on it (indicated by the box around it) opens up an unblockable path (indicated by the dashed by-directed arrow) between Z_i and U_i.

Specifically, in the "reduced-form" part of the two-stage least squares regression, we would compare draftees ($Z_i = 1$) to non-draftees ($Z_i = 0$), given the same college decision $M_i = m$. If Z_i affects the college decision, then the fact that the latter is observed to be constant in such a group must be due to individual differences in U_i, which then affect Y_i irrespective of an actual treatment effect. E.g., draftees that did not attend college to avoid the draft probably had lower parental SES than non-draftees, and lower wages Y_i for that reason alone - even if neither treatment nor college affected earnings.

This open "non-causal" path then actually invalidates both the first-stage $D_i(Z_i = z) \perp\!\!\!\perp Z_i | X_i, M_i$ assumption due to post-treatment selection bias,[6] as well as the $Y_i(D_i = d) \perp\!\!\!\perp Z_i | X_i, M_i$ assumption.

Second, even if Z_i does not *directly* drive M_i, the latter may be influenced by the treatment D_i, as in graph (c) of figure 2. In this case, M_i is a mediator of the $D_i \rightarrow Y_i$ relationship, and is also influenced by Z_i indirectly through D_i. Wooldridge (2010, 95) suggests that on-the-job training might be such a variable in the Vietnam draft application. In this case, Z_i is a valid instrument when one does *not* adjust for M_i. This is because the exclusion restriction obviously holds, and there are also no other back-door paths which connect Z_i and Y_i.

[5] See Appendices B and C for a more detailed explanation of this formal argument.

[6] For an in-depth analysis of this phenomenon in standard adjustment strategies in political science, see Montgomery, Nyhan and Torres (2018).

However, adjusting for M_i introduces a severe, but more subtle problem. D-separation—explained in more detail in Appendix B—does not only prohibit to condition on "colliders" to block paths, but also to condition on *descendants* of such variables. Since Z_i and U_i collide in D_i, conditioning on its "child" M_i has the same qualitative consequences as in graph (b), making it impossible to identify the ATE of Z_i on D_i or the LATE of D_i on Y_i. In the "reduced-form" regression of Y_i on Z_i controlling for M_i, we would again compare individuals with different values for Z_i, but the same M_i. Then, observed differences in Y_i may be due to differences in unobserved U_i that are mediated through D_i, and not due to a causal effect of D_i.

This subtle problem went unnoticed by Deuchert and Huber (2017, 416), who discuss a similar graph and state that conditioning on a mediator satisfies the CIA and identifies a "partial direct effect". As we hope we have made clear, this is not the case, because conditioning on a mediator renders Z_i correlated with U_i, which prohibits any identification.[7] We return to these graphs again when we discuss the possibility of testing which of the assumptions hold.

An interesting special case of graph (c) of figure 2 is when M_i stands for the inclusion of an observation in the dataset (or, reversely, for attrition). In both observational and experimental studies, participants often drop out based on the realization of their treatment or their data is selectively reported due to administrative reasons (Aronow, Baron and Pinson 2019; Elwert and Winship 2014; Knox, Lowe and Mummolo 2020). Researchers are then forced to condition on M_i. In IV settings, even if M_i is not directly driven by U_i and does not influence Y_i, it is a descendant of the collider D_i, so that the instrumental variable becomes invalid. Similarly, in Angrist (1990), it is noted that reported earnings are censored at a maximum l, so that the whole sample is conditional on $Y \leq l$. This means one conditions on a descendant of the true unobserved earnings so that the IV becomes invalid,

[7] Frölich and Huber (2017) propose to identify mediation effects in such a setting using an instrument influencing D_i and a separate instrument influencing M_i.

a fact acknowledged by Angrist (1990, 334). Berinsky and Chatfield (2015) discuss this and related selection problems that may occur for the draft lottery instrument.[8]

A final possible set of causal assumptions is depicted in graph 3. In this graph, M_i is not influenced by the confounder U_i, but affects D_i. Again, the no-confounding assumption is crucial. If it is violated, a collider phenomenon would occur as in the previous cases, making Z_i an invalid instrument. However, if such confounding can be ruled out, one can identify a local ATE:

$$E[Y_i(D_i = 1) - Y_i(D_i = 0)|D_i(Z_i = 1, M_i = m) > D_i(Z_i = 0, M_i = m), X_i]$$

This estimand has not been discussed before. It is the average causal effect of a binary treatment for the latent subpopulation of units which 1) change treatment status as a response to the instrument Z_i, *while fixing* M_i *at* m and 2) which are characterized by covariates X_i.[9]

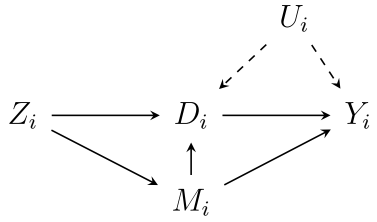

Figure 3: Graph where adjustment for M_i is necessary to identify a local average treatment effect.

The intuition behind this identification result is that under the assumptions in graph 3, one can actually identify the joint effect of Z_i and M_i on D_i, which is what Pearl (2001) and Acharya, Blackwell and Sen (2016) call the "controlled direct effect". For those individuals that shift their treatment uptake as a result of this hypothetical joint intervention, the

[8] See Elwert and Segarra (2020) for an analysis of this problem under a linearity assumption.

[9] Blackwell (2017) discusses similar quantities where M_i would be a second randomized instrument that does not affect Y_i directly.

effect of D_i on Y_i is then also identified. There are additional relevance and monotonicity assumptions needed, which are very similar to the usual LATE assumptions. We discuss these in more detail in Appendices C and E.

We summarize all of these identification results in the following proposition:

Proposition Under the assumptions in graph (a) of figure 2, the CIA

$$D_i(Z_i = z), Y_i(D_i = d) \perp\!\!\!\perp Z_i | X_i, M_i$$

holds and under the usual monotonicity and relevance assumption, the LATE estimand

$$E[Y_i(D_i = 1) - Y_i(D_i = 0) | D_i(Z_i = 1) > D_i(Z_i = 0), X_i, M_i]$$

is identified.

Under the assumptions depicted in graphs (b) of figure 2, the CIA does not hold with any conditioning set.

Under the assumptions depicted in graphs (c) of figure 2, the CIA does hold conditional on X_i, but not conditional on M_i.

Under the assumptions depicted in figure 3, the CIA

$$D_i(Z_i = z, M_i = m), Y_i(D_i = d) \perp\!\!\!\perp Z_i | X_i, M_i$$

holds. If additionally $P(D_i(Z_i = 1, M_i = m) \geq D_i(Z_i = 0, M_i = m) | X_i) = 1$ ("partial" monotonicity) and $E[D_i | Z_i = 1, M_i = m, X_i] - E[D_i | Z_i = 0, M_i = m, X_i] \neq 0$ (relevance) hold, the LATE estimand

$$E[Y_i(D_i = 1) - Y_i(D_i = 0) | D_i(Z_i = 1, M_i = m) > D_i(Z_i = 0, M_i = m), X_i]$$

is identified.

Proof: See Appendix C.

2.5 Judging and Testing the Causal Assumptions

In sum, what are the implications of these results for applied researchers if they suspect that Z_i influences M_i? We emphasize that only the restrictive sets of assumptions in figure 2 (a) and figure 3 allow for IV identification by conditioning on X_i and M_i. Again, if researchers think that the instrument may influence Y_i through variables M_i, they need to rule out confounders that may affect M_i and Y_i either directly or through D_i. We also emphasize that researchers must not condition on mediators of the $D_i \rightarrow Y_i$ relationship. This causes inconsistencies even when instruments are unconditionally valid. We now return to some of the empirical applications that motivated our research and focus on the validity of various tests that were proposed to scrutinize instrumental validity in the face of variables influenced by the instrument.

In general, robustness tests rely on determining "core" and additional control variables such that 1) identification holds with core controls, but also with additional controls and 2) there must be a chance that the robustness test fails if the assumptions are incorrect (Chen and Pearl 2015; White and Lu 2011). Regarding condition 1), if one knew that one of the sets is incorrect a priori, then there would be no point in testing, as one would have to stick to the other, correct set of controls anyways. The problem with many applied papers using IV and post-instrument variables is that they violate this condition. If one allows for the fact that the IV impacts on Y_i directly over M_i, then either the instrument is completely invalid – but one can engage in a sensitivity analysis, as shown below –, or one needs to adjust for M_i.

For example, Wucherpfennig, Hunziker and Cederman (2016) acknowledge the possibility of various post-instrument variables, and try to mitigate such concerns by adjusting for these as a robustness test. They report that estimates under either adjustment set are similar. Such a strategy is also undertaken by Kern and Hainmueller (2009) and Spenkuch and

Tillmann (2017). It turns out that this testing strategy is misleading. To see why, consider first graph (a) in figure 2. In this situation, M_i-adjusted IV estimation identifies a LATE, whereas unadjusted estimates will be different and will exhibit asymptotic bias. In situations like graph (b) in figure 2, Z_i is not a valid instrument *under either adjustment*, and there are no sets of observed variables that are d-separated, so there is no way to empirically test this graph. In graph (c), M_i-adjusted IV estimates will differ, just like in graph (a), but now the *unadjusted* estimators converge to a LATE, whereas the adjusted estimates are biased. Accordingly, researchers cannot circumvent to commit themselves to causal assumptions a priori in situations like these. Comparing adjusted and unadjusted estimates is, in general, misleading: Both equal and unequal estimates may come from a real-world process where the variable Z_i is a valid instrument unconditionally, conditional on M_i, or in neither case.

This phenomenon is directly related to approaches that directly inspect the correlation between Z_i and M_i that would lead to differences in point estimates (conditional on X_i). Researchers often report that this association is not significant, that the instrument is therefore valid, and that M_i can effectively be ignored. But when one acknowledges the possibility of a non-zero effect of Z_i on M_i, the Null hypothesis to be tested should be a composite Null of there being an effect (Hartman and Hidalgo 2018). Additionally, the bias introduced through post-instrument variables increases as the instrument becomes weaker (as discussed below), which such tests do not address. Our sensitivity analysis can be seen as an alternative approach to such a testing strategy because it incorporates the statistical uncertainty associated with the estimate of the effect of Z_i on M_i. This is the price researchers have to pay when they are not willing to definitely assume a zero effect. Finally, even small effects of Z_i on M_i may be relevant when the effect of M_i on Y_i is large.

There is a much more sensible testing strategy for variables that researchers think are influenced by the instrument. One situation in which causal assumptions we have proposed are sharp enough that they allow a test is graph (a) of figure 2. In this graph, D_i and M_i are connected via the $D_i \leftarrow Z \rightarrow M_i$ path, and additional blocked paths running over the collider

Y_i. Accordingly, Z_i (and X_i, as usual) d-separate D_i and M_i, and these two variables should therefore be conditionally independent in the population. This can be tested by estimating $E[D_i|M_i, Z_i, X_i]$ as a function of M_i, which is simply the first-stage that is often reported by researchers. However, the focus normally rests on the partial association between the instrument Z_i and D_i (e.g., for testing whether the instrument is weak), while the test we propose rests on the partial association between the post-instrument variable M_i and D_i. Specifically, graph (a) of figure 2 suggests that the coefficient of a linear regression of D_i on M_i, controlling for Z_i and X_i, is zero (under a correct regression specification and appropriate standard errors). This test may seem unintuitive at first glance because it does not directly check for associations between the instrument and other variables. However, it is the only test that can be justified by relatively weak assumptions. We note that tests for ignorability of the treatment using proxies of unobserved confounders take a similar indirect route (White and Chalak 2010; Pei, Pischke and Schwandt 2017).[10]

What if the test fails, i.e., the independence relationship is empirically violated? In this case, at least one open path between D_i and M_i must exist, like in graphs (b) and (c) of figure 2, or as in figure 3. Accordingly, researchers should consider a priori which of these paths may exist. Again, the possibility of causal cycles must be ruled out beforehand to ensure that any of the conclusions we presented are valid.

[10] Graph (c) of 2 also has a testable implication: $Z_i \perp\!\!\!\perp M_i | D_i, X_i$. This again is a highly non-standard test (as explained, conditioning on D_i leads to misleading inferences in all other situations). We discussed this graph to illustrate the mechanics of conditioning on a descendant of a collider. In most situations, M_i will also be driven by U_i, and then the graph does not have a testable implication.

3 A New Sensitivity Analysis

We have shown that instruments for a causal effect may not be valid when they affect other variables that affect the outcome of interest and are also driven by unobserved confounders. Specifically, conditioning on these other variables M_i oftentimes will not achieve identification. In this section, we propose a new semi-parametric sensitivity analysis for such situations. Our approach is based on the fact that we can often assess the effect of the instrument on the M_i variable, which provides useful information to bound the bias introduced by the direct effect of the instrument. This goes beyond other recent approaches (Conley, Hansen and Rossi 2012; Van Kippersluis and Rietveld 2018) that rely completely on researchers' judgments on the sign and magnitude of the direct effect of the instrument. In contrast, our approach can use sample information. Furthermore, we relax parametric assumptions (e.g., constant effects) that are often made in the literature, and that are often implausible on substantive grounds. We present two different models with different assumptions that nonetheless lead to similar estimation approaches: First, a model for situations where instrument, treatment, and post-instrument variable are binary. Second, a model for a binary instrument, but possibly continuous treatment and post-instrument variable.

3.1 Model 1: Binary Variables

When Z_i, D_i, and M_i are all binary, one can perform sensitivity analysis under relatively weak parametric restrictions. At the same time, the resulting estimation approach is a special case of our second approach that we discuss in the next section.

Here, our model for Y_i looks as follows:

$$Y_i = \mu_Y + \beta_i D_i + \gamma_i M_i + \lambda_{1i}^{'} X_i + \epsilon_{1i}. \tag{2}$$

In this model, all causal effects vary across individuals in a fairly unrestricted fashion, and so are random variables (see Imai and Yamamoto (2013) for a similar setup). X_i is a vector of controls. We assume $E[\epsilon_{1i}] = 0$ without loss of generality. In Appendix D, we show that when D_i and M_i are binary and further exogeneity and monotonicity assumptions discussed below hold, the standard LATE conditional on X_i can be expressed as

$$\begin{aligned}&\frac{E[Y_i|Z_i = 1, X_i] - E[Y_i|Z_i = 0, X_i]}{E[D_i|Z_i = 1, X_i] - E[D_i|Z = 0, X_i]} - \\ &E[\gamma_i|M_i(Z_i = 1) > M_i(Z_i = 0)] \times \frac{E[M_i|Z_i = 1, X_i] - E[M_i|Z = 0, X_i]}{E[D_i|Z_i = 1, X_i] - E[D_i|Z = 0, X_i]}.\end{aligned} \tag{3}$$

In this expression, the first term can be estimated by a standard two-stage least squares regression that completely ignores M_i, with outcome Y_i, treatment D_i, instrument Z_i, and controls X_i. The second term is the asymptotic bias introduced by direct effects of the instrument through M_i. It consists of the average causal effect of M_i on Y_i (γ_i) for units for which Z_i has an effect on M_i. This is the unknown sensitivity parameter. It is multiplied by a term that can be estimated via another standard two-stage least squares regression, but now with outcome M_i. Here, the numerator equals the average effect of Z_i on M_i, which (under monotonicity) is equal to the share of units for which Z_i has an effect on M_i. The larger this effect, the larger the bias. The denominator is the first-stage of the main regression and equals the share of units for which the instrument has an effect on the treatment. The smaller this quantity, the weaker the instrument is for D_i, and the larger the bias through direct effects is.

An important insight from this bias decomposition is that the association between Z_i and M_i may be small, but the bias nonetheless large if the instrument is (also) weakly associated with D_i. This is on top of other problems associated with weak instruments which occur in finite samples (Bound, Jaeger and Baker 1995).

While the causal model for Y_i in equation 2 restricts interactions between the observed variables, we make no assumption on the causal models for D_i and M_i, except that the effect of Z_i is “monotone” in both.[11] Therefore, this approach is quite general, although with continuous X modeling will be necessary.

3.2 Model 2: Binary IV, Continuous Treatment and Post-Instrument Variable

Many applications (including our two empirical analyses further below) deal with continuous D_i or M_i, in which our bias decomposition is not valid. Here, one must instead make further assumptions on the causal models for D_i and M_i. Consistent with our model for Y_i, we assume that

$$D_i = \mu_D + \alpha_i Z_i + \pi_i M_i + \lambda_{2i}^{'} X_i + \epsilon_{2i} \tag{4}$$

$$M_i = \mu_M + \delta_i Z_i + \lambda_{3i}^{'} X_i + \epsilon_{3i}. \tag{5}$$

Importantly, the causal model defined by all three equations is consistent with graphs (a) and (b) graphs in figure 2 and additionally allows for M_i to affect D_i.[12]

Beyond the causal models, we make a series of further assumptions which are enumerated

[11] One could in fact allow for interactions between D_i and M_i in the model in equation 2. The interaction term would be a second sensitivity parameter that is multiplied with the estimable share of “joint compliers”, $P(D_i(Z_i = 1)M_i(Z_i = 1) > D_i(Z_i = 0)M_i(Z_i = 0))$. See Blackwell (2017) for related estimation strategies when there are two IVs. Since applied researchers using IV regressions rarely specify interactions between treatment and covariates and allowing for them in our second sensitivity model increases complexity even more, we do not pursue this here.

[12] In graph (c), a sensitivity analysis would only be necessary if Z_i affected M_i directly. However, β_i would then no longer describe the total effect of D_i, which is of primary interest in most analyses.

in Appendix D. We here give an intuitive summary. The first assumption follows from graphs (a) and (b) in figure 2. It requires that there are no unblocked back-door paths from Z_i to any of D_i, M_i, Y_i, and that there is no direct effect of Z_i on Y_i save for the effects through D_i and M_i. The second assumption states that Z_i affects D_i monotonically, which again is a standard assumption. The third assumption requires Z_i to also affect M_i monotonically. Both monotonicity assumptions restrict π_i, so that in most situations arguments for one of these to be plausible also make the other plausible. However, they are logically independent (we expand on this in Appendix E). Finally, for our second sensitivity model, we assume that the covariance of the potential outcomes $(M(0), M(1))$ is non-negative. This assumption allows us to use the data to bound a parameter and effectively decreases the width (but not the midpoint) of the resulting bounds. If analysts are not willing to impose this assumption and they find a large mean effect of Z_i on M_i, we suggest that they allow for larger values of the second sensitivity parameter σ_{γ_i} than is otherwise plausible. We discuss this in more detail in Appendix D.3.

Under these assumptions and the causal models for Y_i, M_i, and D_i, we show in Appendix D that one can bound the weighted causal effect of D_i on Y_i, $E\left[\frac{\alpha_i + \delta_i \pi_i}{E[\alpha_i + \delta_i \pi_i]} \beta_i\right]$. The bias term becomes

$$E[\delta_i \gamma_i] = E[\delta_i] E[\gamma_i] + cov(\delta_i, \gamma_i). \tag{6}$$

Here, $E[\delta_i]$ is the average causal effect of Z on M (equal to the share of M_i-compliers), which can be estimated from the data. $E[\gamma_i]$ is the direct effect of M on Y, which is the first sensitivity parameter.[13] If treatment effects were constant, it would be the only unknown. However, if treatment effects vary and unobserved confounders impact on both M and Y, the individual-level effects δ_i and γ_i will be correlated, and the covariance term will be different

[13] To connect this to the first sensitivity model, note that with M_i continuous, δ_i is possibly continuous as well so that $P(\delta_i = 0) = 0$, and, due to monotonicity, $E[\gamma_i] = E[\gamma_i | \delta_i > 0]$.

from zero (Glynn 2012).

For example, in the Vietnam draft study, if unobserved parental SES U_i influences the decision to attend college (M_i) as well as later wages (Y_i), it is plausible that lower parental SES makes both effects in question larger, and thereby creates a positive covariance between them. For example, for men with low parental SES, the effect of the draft on attending college (δ_i) will be relatively large (because they are more likely to be at the margin when it comes to deciding for or against college, Card 1999). And we would expect the effect of college on earnings (γ_i) in this group also to be relatively large because it has a higher potential to benefit (Brand and Xie 2010). Accordingly, $cov(\delta_i, \gamma_i)$ would be positive. Taken together, this could lead to large bias, even if the constituent average causal effects are small. Previous approaches to sensitivity analysis (Conley, Hansen and Rossi 2012; Van Kippersluis and Rietveld 2018) assume that all causal effects are constants and therefore cannot address biases that arise from such scenarios.

We show in Appendix D that one can use the data to bound this covariance term. Intuitively, the bounds increase when the standard deviation of M and the effect of Z on M's standard deviation gets larger. The second sensitivity parameter then is the standard deviation of γ_i, σ_{γ_i}. This quantity is in the same units as $E[\gamma_i]$, and describes how much γ_i typically varies. As illustrated, this standard deviation may be quite large even if mean effects are thought to be small.

Finally, we can extend this sensitivity model to situations where the post-instrument variable M may be measured with error. This is of special interest because often researchers are made aware of potential violations of the exclusion restriction after initial data collection. Although they then might gather some measure of a candidate M_i variable, it may well be affected by measurement error. It turns out that such an error-ridden measure is still informative and can be used for sensitivity analysis.

We formalize this by complementing the model in equations 2 - 5 with a model for $M_i^{\star}$, the observed measure of the now unobserved M_i:

$$M_i^\star = M_i + \eta_i \tag{7}$$

and by assuming $Z_i, M_i \perp\!\!\!\perp \eta_i$ and $E[\eta_i] = 0$. This is "classical" measurement error. In Appendix D, we show that the resulting estimator for the bounds stays the same, although measurement error does indeed widen the bounds compared to a situation without measurement error.

4 An Illustration of the Proposed Methodology

We illustrate our new sensitivity analysis using data from Spenkuch and Tillmann (2017) as well as from Carnegie and Marinov (2017). In the case of Spenkuch and Tillmann (2017), we find that their original estimate is very robust to a large mean effect of the post-instrument variable on the outcome as well as to all but implausibly large variances of this effect. In the case of Carnegie and Marinov (2017), we find a more nuanced picture: While their estimate is nominally not very robust against negative effects of M_i on Y_i, it actually increases in size and significance for substantively more plausible positive effects. Then, it is also robust to some heterogeneity in these effects.

4.1 The Effect of Religious Composition on Support for National Socialists

In Spenkuch and Tillmann (2017), one aim is to estimate the effect of Catholicism on the vote share of the national socialists (NSDAP) in Germany in 1932. The data used is on the county-level and comprises official election results and census data on the share of Catholics, protestants, and other religions, as well as extensive socio-economic information like unemployment rates in various demographic subgroups. Since the authors cannot rule out unobserved confounders between religious composition and the Nazi vote share, they suggest using a county's official religion measured in 1624 as an instrument for the effect of religion on the propensity to vote for the NSDAP. They discuss evidence that the historical county denomination was largely idiosyncratic, except for a few observable factors for which they adjust in their statistical analysis. In our framework, these variables correspond to pre-instrument covariates X_i.

Spenkuch and Tillmann (2017, p. 9) then further assert that for this historical variable to be a valid instrument for the effect of interest, "it may influence voters' decisions to support the NSDAP only through its impact on covariates that are included in the regression". We take this as an indication that past religious composition Z_i may have affected, for example, the economic situation in counties in the 1930s, which we conceptualize as M_i variables. One well-known mechanism for such an effect is Max Weber's hypothesis of a "protestant work ethic". Furthermore, it is plausible that such economic variables also exerted a strong influence on the Nazi vote share. Accordingly, the instrument would be valid if we faced the situation of graph (a) in figure 2. Assuming the IV is valid, Spenkuch and Tillmann (2017) estimate that a one percentage point increase in the share of Catholics in a county decreased the NSDAP votes share by about 0.27 percentage points. This estimate is very large and also quite precise (its estimated standard error is about 0.03). It is substantively and statistically indistinguishable from OLS estimates, where Catholicism explains about 40% of the variance in NSDAP votes.

From this alone, it is clear that only strong deviations from the IV assumption can change the substantive conclusions. We concentrate on one single M_i variable measuring a highly relevant economic fundamental: The county-level unemployment rate among blue-collar workers. Figure 4 plots our sensitivity analysis for the causal effect of Catholicism on NSDAP vote. The X-axis depicts the first sensitivity parameter, the mean effect of unemployment on the NSDAP vote share. The solid lines depict 95% confidence intervals as a function of this sensitivity parameter when the heterogeneity in this effect (σ_{γ_i}), the second sensitivity parameter, is set to 0. It is clear that due to a weak association of Z_i with

M_i, the inference is virtually unchanged.

The grey dashed lines visualize confidence intervals when there is more heterogeneity in the causal effect of M_i on Y_i. Here, we see that when σ_{γ_i} exceeds 0.9, the confidence interval covers 0, so that the original inference is not robust. Is such a heterogeneous effect reasonable to expect? Due to severe data limitations, the empirical literature on Weimar elections focuses on descriptive inferences (King et al. 2008), so that it cannot directly inform our assessment on the magnitude of σ_{γ_i}. Spenkuch and Tillmann (2017)'s own estimates for mean effects of unemployment, which they do not claim are causal effects, are negative and at most as large as the effect of Catholicism. The contemporary literature on the causes of extreme-right voting (Jackman and Volpert 1996; Arzheimer 2009), finds positive effects of both individual unemployment and aggregate unemployment rates. This suggests some variability in effects. However, even if the effect of unemployment rates would vary uniformly between, say, -0.75 and 0.5 percentage point increases across counties, the implied standard deviation would only be about 0.36. While this introduces some additional uncertainty, the main inference is robust.

4.2 The Effect of Foreign Aid on Human Rights

In Carnegie and Marinov (2017), the authors exploit the "essentially random" rotation of the presidency of the Council of the European Union across member states as a variable Z_i that impacts on foreign aid D_i transferred to countries that used to be colonies of that presiding state. They use this variable as an instrument for aid to address the longstanding and contested question of whether aid impacts human rights and democratic institutions of a receiver country (Y_i).

The paper carefully adjusts for institutional reforms that can be interpreted as pre-instrument X_i variables and, since the data set is a panel of countries over years, lags variables to avoid "posttreatment bias" (Carnegie and Marinov 2017, 680). This can be interpreted as trying to avoid situations such as in graph (c) in Figure 2. In their main analysis, they do

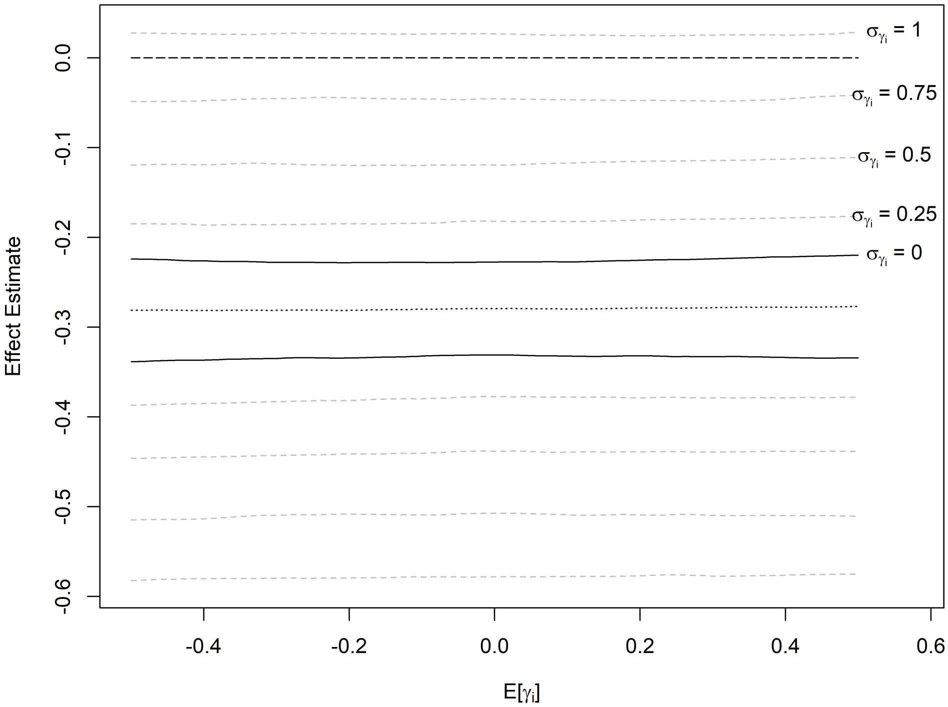

Figure 4: 95% confidence intervals for the effect of Catholicism on NSDAP vote shares, as a function of the mean effect of unemployment on NSDAP vote shares ($E[\gamma_i]$) and its variability (σ_{γ_i}). Original estimates are replicated by the solid interval at $E[\gamma_i] = 0$. Based on data from Spenkuch and Tillmann (2017).

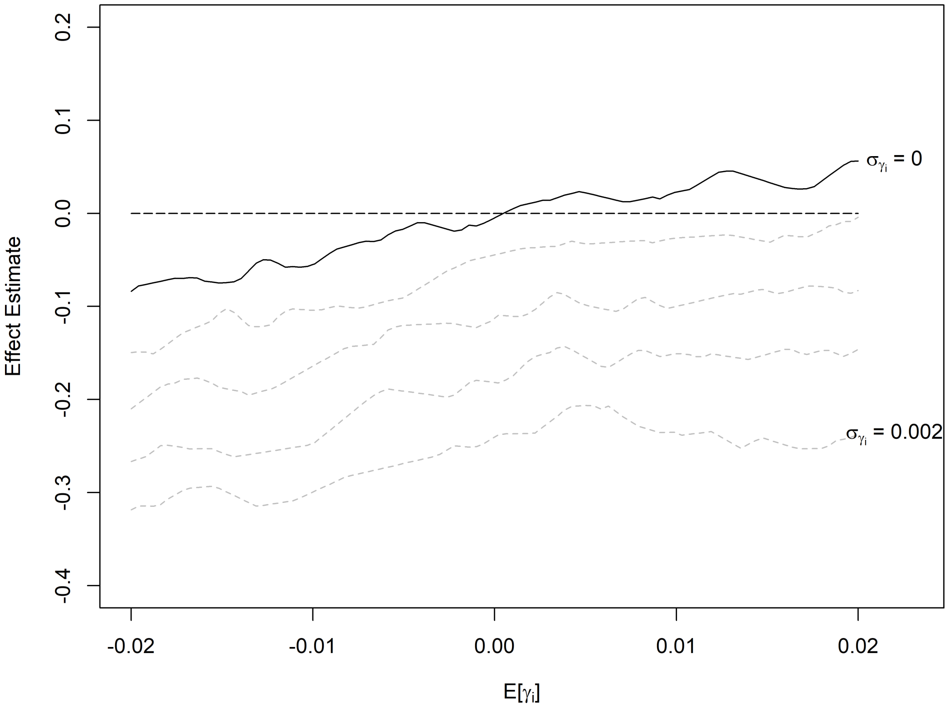

Figure 5: Lowest ends of 95% confidence intervals for the effect of (logged) foreign aid in $ million on human rights (0-14 scale), as a function of the mean effect of economic openness (in percentage points) on human rights ($E[\gamma_i]$) and its variability (σ_{γ_i}). The original lowest end is replicated by the solid line at $E[\gamma_i] = 0$. The original confidence interval was approx. $[0.03, 3.74]$. Based on data from Carnegie and Marinov (2017).

not adjust for any plausible post-instrument variables and find a very large, but short-lived effect of foreign aid on a Human Rights index: An increase by \$5 million increases the 0-14 index by 0.4 points, with an associated confidence interval $[0.01, 0.8]$.

However, in their discussion of the exclusion restriction, Carnegie and Marinov (2017, A5, Table A18) mention several observed variables that may be influenced by the instrument and test for their association with the instrument, assuming a Null of zero association. As discussed before, such an approach does not incorporate appropriately additional uncertainty and mean biases that may occur.

In Figure 5, we plot the sensitivity of their main inference with respect to "economic openness" as one of such M_i-variables (defined as the sum of national export and imports in terms of GDP per capita). The plot zooms in on the Y-axis around 0 and shows the lower end of confidence intervals for the effect of (logged) foreign aid on human rights.

The solid line indicates the endpoints of the confidence intervals as a function of varying mean effects of economic openness. The literature suggests that this effect is likely to be positive, but may also be highly heterogeneous (Hafner-Burton 2005; Hill Jr. and Jones 2014). We see here that while the effect of foreign aid on human rights becomes insignificant for negative values of $E[\gamma_i]$, it actually becomes larger and more significant for substantively more plausible positive effects of M_i on Y_i. This is true for values of $E[\gamma_i]$ up to 0.02, which would correspond to a fairly large effect.[14] In such situations, the effect is also robust against small heterogeneity in the effect of M_i on Y_i (σ_{γ_i}). Here, the lowest dashed gray line corresponds to situations where $\sigma_{\gamma_i} = 0.002$, and the lines in between have it decreasing by 0.0005. Such magnitudes seem plausible, if small. For example, if the effect varies uniformly between 0 and 0.04, the induced standard deviation is already about 0.012. In sum, the main inference becomes actually stronger for plausible values of $E[\gamma_i]$, but then is sensitive

[14] In the data, median economic openness is around 70% and the 75% percentile is about 100%. Therefore, an effect of 0.02 would mean that moving from the median to the 75% percentile translates into an 0.6 increase in the 0-14 Human Rights index.

to plausible heterogeneity in γ_i.

5 Conclusion

Many applied researchers use instrumental variables in settings where they try to "control away" a direct effect of the instrument on the outcome by measuring other variables M. In this paper, we explained why this strategy only works under specific, restrictive assumptions. Using causal graphs and potential outcomes, we highlighted the asymmetric role of pre- and post-instrument covariates: While adjustment for the former is often necessary and unproblematic, statistical control for the latter has to be taken with extreme caution. We showed that with direct effects of the instrument through M_i, some local average treatment effects may be identified, but we also highlighted various sources of asymptotic bias. We discussed the limited value of existing robustness tests and provided a more suitable test of a specific set of identification assumptions. Finally, we introduced a sensitivity analysis as an alternative and illustrated it using the IV analysis in Spenkuch and Tillmann (2017) and Carnegie and Marinov (2017). Here, it became clear that both mean effects, as well as the variability of causal effects, may play an important role in the sensitivity of estimates.

We conclude by providing a checklist for applied researchers that want to use a (potential) instrumental variable that may have a direct effect on the outcome through another variable:

1. Based on substantive knowledge, determine which of the graphs discussed in this paper seems plausible for your research design. Specifically, be clear about which variables are confounders X_i that influence Z_i, D_i, and Y_i. and which variables M_i are driven by Z_i or D_i.

2. If M_i is a mediator and not directly driven by Z_i (as in graph (c) of figure 2), proceed with standard estimation routines like 2SLS, where you condition only on X_i.

3. If your assumptions are equivalent to those in graph (a) of figure 2, implement the

diagnostic test by checking whether D_i and M_i are independent conditional on Z_i. If they are, condition on X_i and M_i in your statistical analysis.

4. If the test fails, reconsider your assumptions. Only the assumptions in figure 3 allow for conditional dependency between D_i and M_i and identification based on adjustment for X_i and M_i.

5. If prior knowledge or the diagnostic test leads to the conclusion that Z_i directly influences M_i and that the unobserved confounder also influences M_i (as in graph (b) of figure 2), identification is not possible. Perform estimation conditional only on X_i and then use our sensitivity analysis to assess whether substantive conclusions still hold.

Finally, we reiterate a point made, inter alia, by Conley, Hansen and Rossi (2012): A strong but imperfect instrument may be preferable to an exogenous, but weak instrument. The strength of an instrument is, of course, estimable. When a central post-instrument variable M_i is measured, our method also allows researchers to better assess the consequences of imperfections of their instrument, without the need to rely completely on a priori judgments about exogeneity.

References

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. "Explaining causal findings without bias: Detecting and assessing direct effects." *American Political Science Review* 110(3):512–529.

Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review* 80(3):313–336. **URL:** *https://ideas.repec.org/a/aea/aecrev/v80y1990i3p313-36.html*

Angrist, Joshua D and Guido 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(430):431–442.

Angrist, Joshua D, Guido W Imbens and Donald B Rubin. 1996. "Identification of causal effects using instrumental variables." *Journal of the American statistical Association* 91(434):444–455.

Angrist, Joshua David and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: an empiricist's companion*. Princeton University Press.

Aronow, Peter M, Jonathon Baron and Lauren Pinson. 2019. "A note on dropping experimental subjects who fail a manipulation check." *Political Analysis* 27(4):572–589.

Arzheimer, Kai. 2009. "Contextual factors and the extreme right vote in Western Europe, 1980–2002." *American Journal of Political Science* 53(2):259–275.

Bazzi, Samuel and Michael A Clemens. 2013. "Blunt instruments: avoiding common pitfalls in identifying the causes of economic growth." *American Economic Journal: Macroeconomics* 5(2):152–186.

Berinsky, Adam J and Sara Chatfield. 2015. "An empirical justification for the use of draft lottery numbers as a random treatment in political science research." *Political Analysis* 23(3):449–454.

Betz, Timm, Scott J Cook and Florian M Hollenbach. 2018. "On the use and abuse of spatial instruments." *Political Analysis* pp. 1–6.

Blackwell, Matthew. 2017. "Instrumental variable methods for conditional effects and causal interaction in voter mobilization experiments." *Journal of the American Statistical Association* 112(518):590–599.

Bound, John, David A Jaeger and Regina M Baker. 1995. "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak." *Journal of the American statistical association* 90(430):443–450.

Brand, Jennie E and Yu Xie. 2010. "Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education." *American sociological review* 75(2):273–302.

Card, David. 1999. The causal effect of education on earnings. In *Handbook of labor economics*. Vol. 3 Elsevier pp. 1801–1863.

Carnegie, Allison and Nikolay Marinov. 2017. "Foreign Aid, Human Rights, and Democracy Promotion: Evidence from a Natural Experiment." *American Journal of Political Science* 61(3):671–683.

Chen, Bryant and Judea Pearl. 2015. Exogeneity and robustness. Technical report Tech. Rep.

Conley, Timothy G, Christian B Hansen and Peter E Rossi. 2012. "Plausibly exogenous." *Review of Economics and Statistics* 94(1):260–272.

Deuchert, Eva and Martin Huber. 2017. "A cautionary tale about control variables in IV estimation." *Oxford Bulletin of Economics and Statistics* 79(3):411–425.

Elwert, Felix and Christopher Winship. 2014. "Endogenous selection bias: The problem of conditioning on a collider variable." *Annual Review of Sociology* 40:31–53.

Elwert, Felix and Elan Segarra. 2020. "Instrumental Variables with Treatment-Induced Selection: Exact Bias Results." *arXiv preprint arXiv:2005.09583* .

Frölich, Markus and Martin Huber. 2017. "Direct and indirect treatment effects–causal chains and mediation analysis with instrumental variables." *Journal of the Royal Statistical*

Society: Series B (Statistical Methodology) .
URL: *http://dx.doi.org/10.1111/rssb.12232*

Glynn, Adam N. 2012. "The product and difference fallacies for indirect effects." *American Journal of Political Science* 56(1):257–269.

Glynn, Adam N. and Konstantin Kashin. 2017. "Front-Door Difference-in-Differences Estimators." *American Journal of Political Science* .

Hafner-Burton, Emilie M. 2005. "Right or robust? The sensitive nature of repression to globalization." *Journal of Peace Research* 42(6):679–698.

Hartman, Erin and F Daniel Hidalgo. 2018. "An equivalence approach to balance and placebo tests." *American Journal of Political Science* 62(4):1000–1013.

Hill Jr., Daniel W and Zachary M Jones. 2014. "An empirical evaluation of explanations for state repression." *American Political Science Review* pp. 661–687.

Imai, Kosuke and In Song Kim. 2019. "When should we use unit fixed effects regression models for causal inference with longitudinal data?" *American Journal of Political Science* 63(2):467–490.

Imai, Kosuke, Luke Keele, Dustin Tingley and Teppeiarlarl Yamamoto. 2011. "Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies." *American Political Science Review* 105(4):765–789.

Imai, Kosuke and Teppei Yamamoto. 2013. "Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments." *Political Analysis* 21(2):141–171.

Jackman, Robert W and Karin Volpert. 1996. "Conditions favouring parties of the extreme right in Western Europe." *British Journal of Political Science* 26(4):501–521.

Kern, Holger Lutz and Jens Hainmueller. 2009. "Opium for the masses: How foreign media can stabilize authoritarian regimes." *Political Analysis* 17(4):377–399.

King, Gary, Ori Rosen, Martin Tanner and Alexander F Wagner. 2008. "Ordinary economic voting behavior in the extraordinary election of Adolf Hitler." *The Journal of Economic History* 68(4):951–996.

Knox, Dean, Will Lowe and Jonathan Mummolo. 2020. "Administrative Records Mask Racially Biased Policing." *American Political Science Review* p. 1–19.

Lu, Xun and Halbert White. 2014. "Robustness checks and robustness tests in applied economics." *Journal of econometrics* 178:194–206.

Marshall, John. 2016. "Coarsening Bias: How Coarse Treatment Measurement Upwardly Biases Instrumental Variable Estimates." *Political Analysis* 24(2):157–171.

Mellon, Jonathan. 2020. "Rain, Rain, Go away: 137 potential exclusion-restriction violations for studies using weather as an instrumental variable." *Available at SSRN* .

Montgomery, Jacob M, Brendan Nyhan and Michelle Torres. 2018. "How conditioning on posttreatment variables can ruin your experiment and what to do about it." *American Journal of Political Science* 62(3):760–775.

Pearl, Judea. 2001. Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc. pp. 411–420.

Pearl, Judea. 2009. *Causality*. Cambridge university press.

Pei, Zhuan, Jörn-Steffen Pischke and Hannes Schwandt. 2017. "Poorly Measured Confounders are More Useful on the Left Than on the Right." *NBER Working Paper* 23232.

Rosenbaum, Paul R. 1984. "The consquences of adjustment for a concomitant variable that has been affected by the treatment." *Journal of the Royal Statistical Society. Series A (General)* pp. 656–666.

Shpitser, Ilya, Tyler VanderWeele and James M Robins. 2010. On the validity of covariate adjustment for estimating causal effects. In *26th Conference on Uncertainty in Artificial Intelligence, UAI 2010.* pp. 527–536.

Sovey, Allison J and Donald P Green. 2011. "Instrumental variables estimation in political science: A readers' guide." *American Journal of Political Science* 55(1):188–200.

Spenkuch, Jörg L and Philipp Tillmann. 2017. "Elite Influence? Religion and the Electoral Success of the Nazis." *American Journal of Political Science* .

Stanig, Piero. 2015. "Regulation of speech and media coverage of corruption: An empirical analysis of the Mexican Press." *American Journal of Political Science* 59(1):175–193.

Van Kippersluis, Hans and Cornelius A Rietveld. 2018. "Pleiotropy-robust Mendelian randomization." *International Journal of Epidemiology* 47(4):1279–1288.

White, Halbert and Karim Chalak. 2010. "Testing a conditional form of exogeneity." *Economics Letters* 109(2):88–90.

White, Halbert and Xun Lu. 2011. "Causal diagrams for treatment effect estimation with application to efficient covariate selection." *Review of Economics and Statistics* 93(4):1453–1459.

Wooldridge, Jeffrey M. 2010. *Econometric analysis of cross section and panel data.* MIT press.

Wucherpfennig, Julian, Philipp Hunziker and Lars-Erik Cederman. 2016. "Who inherits the state? Colonial rule and postcolonial conflict." *American Journal of Political Science* 60(4):882–898.

Online Appendix

"Post-Instrument Bias"

Table of Contents

A Papers Using Instrumental Variables (Survey)

Table A1 presents the counts of articles taken from the American Political Science Review, the American Journal of Political Science, and the Journal of Politics that use instrumental variables in their empirical analyses for the period from 2010 to the present. For each of the papers found, we have coded whether there is an explicit discussion regarding the exclusion restriction and among those where there is, whether there is a covariate being included as a control to satisfy such restriction. The table shows that 75.12% of the papers discuss the exclusion restriction and 19.02% include a covariate to address potential violations to this assumption. When dividing the sample into two periods, one starting in 2010 up to 2014 and a second one for papers published in 2015 and after, we see that the percentage of papers that apply the fix has increased, from 14.1% to 22.05%.

Table A1: Exclusion Restriction and Added Covariates (Counts)

	Exclusion restriction	Added covariate	Total articles
2010-2014	58	11	78
	[74.36]	[14.10]	[100]
2015-2020	96	28	127
	[75.59]	[22.05]	[100]
2010-2020	154	39	205
	[75.12]	[19.02]	[100]

Exclusion restriction denotes the number of articles that explicitly discuss exclusion restrictions as identification assumptions in the instrumental variable analysis. Added covariate denotes articles that include a control variable to address a violation of the exclusion restriction. Total articles is the number of articles using instrumental variable techniques. Percentages are taken over total articles in the period and are in brackets.

B Causal Graphs and IV Identification using Potential Outcomes

Causal graphs, specifically *directed acyclic graphs*, consist of *nodes*, which visualize variables, and *edges*, which are usually directed arrows from one node to another. A *path* is any consecutive sequence of edges. In line with Pearl (2009), we view causal graphs as depictions of a nonparametric system of structural equations that describes cause-effect relationships. That is, nodes stand for observable or unobservable features of the units of interests, and an edge or arrow from one such node to the other communicates the assumption that the one variable causally affects the other variable in the population of interest. To be precise, a causal model G consists of exogenous background variables U_i, usually assumed to be unobserved, observed endogenous[15] variables V_i, and structural (causal) functions f_v for each endogenous variable. These functions are deterministic in the sense that if we knew all relevant inputs of f_v for an endogenous variable, we could determine the value of this variable exactly. Since U_i is assumed to be unknown, the observable variables V_i become random variables. Whenever we want to indicate that observable variables are driven by an unobserved confounder, we will use dashed nodes for edges emanating from this confounder. This is equivalent to assuming that the "structural errors" U_i (i.e., all unobserved causes) of the confounded variables are dependent. Throughout, we discuss *acyclic* graphs, that is, graphs where no variable may have an effect on itself. Finally, we use upper-case letters do denote random variables, and lower-case letters to denote realized or fixed values of these variables.

[15] Here, the word "endogenous" simply means "explained in the model".

B.1 Deriving Independencies from Causal Graphs

To understand in which situations an instrument is (conditionally) valid, it is necessary to derive independence relationships from the causal graph the researcher assumes. Throughout, we do so by using an easy yet powerful tool called *d-separation* (Geiger, Verma and Pearl 1990). In a given graph, a path p is said to be d-separated (or *blocked*) by a set of nodes Z_i if and only if

1. p contains a chain $X_i \rightarrow M_i \rightarrow Y_i$ or a fork $X_i \leftarrow M_i \rightarrow Y_i$ such that the middle node M_i is in Z_i, or

2. p contains an inverted fork (or *collider*) $X_i \rightarrow M_i \leftarrow Y_i$ such that the middle node M_i is not in Z_i and such that no descendant of M_i is in Z_i.

A set of variables Z_i is then said to d-separate X_i from Y_i if and only if Z_i blocks every path from a node in X_i to a node in Y_i. Importantly, d-separation implies conditional independence, which we write as $X_i \perp\!\!\!\perp Y_i | Z_i$. This means that, once we know the value of Z_i, X_i does not predict Y_i and vice versa. In addition, we employ graphoid axioms (Dawid 1979) to prove our results. We expand on these more technical aspects and give proofs in the appendix. In the main body of this article, we stick as closely as possible to intuitive explanations.

The fact that conditioning on a collider of two variables (or its descendant) makes these variables dependent is central to understanding the failure of certain IV strategies, but may be counterintuitive, so that an example is helpful. Consider two independent binary variables A and B and a random variable C that is the sum of A and B. Accordingly, C can take on the values $\{0, 1, 2\}$, and is a collider variable, with A and B pointing into it. A and B may be random coin flips, so clearly knowing the value of A does not help in predicting B. However, conditioning on the collider C means that we are told its value, for example 1. The question then is whether A and B have become dependent, that is, whether knowing C and A now tells us anything about B. The answer is a clear yes: Knowing the result C is 1 and, for example, that A is 0, we know for sure that B has to be 1. Put differently, knowing the result of a process (C) and the value of one of its independent inputs (A) also lets us predict the value of the other input (B). The same mechanics apply if we happen to know the realization of a descendant of C. For example, let D be a variable that takes on the value 1 when C equals 1, and is 0 otherwise (so that it is a binary proxy for C). Knowing that D equals 1 and that A equals 0 also leads to the prediction that B equals 1.

To give a more elaborate example of d-separation, consider Figure 1 in the main text. Let us assume for the moment that we could measure U_i and we were interested in its dependency with Z_i . In this case, one would find four paths between the instrument Z_i and U_i: $Z_i \rightarrow D_i \leftarrow U_i$, $Z_i \leftarrow X_i \rightarrow D_i \leftarrow U_i$, $Z_i \rightarrow D_i \rightarrow Y_i \leftarrow U_i$, and $Z_i \leftarrow X_i \rightarrow Y_i \leftarrow U_i$. The first two paths contain the variable D_i as a collider and so are unconditionally blocked. The last two paths contain Y_i as a collider and therefore are blocked as well. In summary, all paths between Z_i and U_i are blocked unconditionally, so that Z_i and U_i are d-separated and $Z_i \perp\!\!\!\perp U_i$ holds. Put informally, this conveys the notion that a valid instrument needs to be independent from unmeasured causes of Y_i. Accordingly, if one could measure U_i for each individual, a linear regression of it on Z_i should yield a coefficient of zero (asymptotically and under the assumption that the regression is correctly specified).

B.2 From Graphs to Potential Outcomes

Having discussed the basic properties of causal graphs, we now introduce potential outcomes and the causal effects of interests. As usual, the identification assumptions need to be stated in independence relationships of observed and counterfactual variables. Following Pearl (2009), we connect causal graphs and potential outcomes by defining the latter quite naturally as solutions to the structural model that researchers assume. The potential outcome of variables $Y_i \in V_i$ when variables $X_i \in V_i$ are set to x is denoted $Y_i(X = x)$ and is given by $Y_i(G_x)$. G_x stands for a manipulated version of the original causal model G in which all functions f_{X_i} are deleted and replaced by constants x (Pearl 2009, 204).

To give a simple example, consider the graph $D_i \to Y_i \leftarrow U_i$. In this graph, the potential outcome of Y_i in unit i when D_i is set to d is

$$Y_i(D = d) = f_y(d, U_i)$$

which, since d is fixed, is a random variable only because it is a function U_i, which stands for all unobserved causes of Y_i. It follows immediately that $D_i \perp\!\!\!\perp Y_i(D_i = d)$ ("ignorability") holds, because D_i and U_i are d-separated unconditionally (since Y_i is a collider that blocks the only path between D_i and U_i). In DAGs, ignorability of the treatment can also be evaluated by simple graphical criteria like the adjustment criterion (Shpitser, VanderWeele and Robins 2010). However, we resort to this structural definition of counterfactuals to make explicit the exact reasons for why IV identification may fail, and because such general graphical criteria for IV problems do not exist.

Our approach is fully compatible with previous results that used counterfactuals to communicate causal assumptions. Approaches that define potential outcomes as byproducts of structural equation are also becoming standard in econometrics, see for example Imbens and Newey (2009), Chernozhukov et al. (2013), and especially White and Lu (2011), who also employ causal graphs. It should also become clear that potential outcomes are indeed a generalization and refinement of the "structural error" that plays a central role in econometrics. Again, this error term in a structural or causal equation stands for all unobserved factors that influence the outcome when observed determinants are held fixed, and it should not be confused with the regression error. The latter stands for unit's deviations in Y_i from its conditional mean.[16]

[16] See Imbens (2014) for a discussion of this issue in an IV context.

C Proof of the Proposition

For ease of exposition, we first introduce some useful properties of conditional independence:

Lemma 1. *(Dawid 1979) If $X_i \perp\!\!\!\perp Y_i | Z_i$ and U_i is a function of X_i, then 1) $U_i \perp\!\!\!\perp Y_i | Z_i$ and 2) $X_i \perp\!\!\!\perp Y_i | Z_i, U_i$.*

Lemma 2. *(Contraction, Pearl (2009)) $X_i \perp\!\!\!\perp Y_i | Z_i$ and $X_i \perp\!\!\!\perp W_i | Z_i, Y_i$ imply $X_i \perp\!\!\!\perp Y_i, W_i | Z_i$.*

Lemma 3. *$Z_i \perp\!\!\!\perp U_i | X_i$ implies $Z_i \perp\!\!\!\perp f(U_i), g(U_i) | X_i$, where f, g are arbitrary functions.*

Proof. $Z_i \perp\!\!\!\perp U_i | X_i$ implies $Z_i \perp\!\!\!\perp f(U_i) | X_i$ as well as $Z_i \perp\!\!\!\perp U_i | X_i, f(U_i)$ by lemma 1. The latter then similarly implies $Z_i \perp\!\!\!\perp g(U_i) | X_i, f(U_i)$. By contraction, we then have $Z_i \perp\!\!\!\perp f(U_i), g(U_i) | X_i$. □

We can now prove the statements in the main text. Throughout, we will assume there are additional observed confounders X_i influencing all observed variables.

Proof of the Proposition. In graph (a) of figure 2, we have $Y_i(D_i = d) = f_y(d, M_i, X_i, U_i)$ and $D_i(Z_i = z) = f_d(z, X_i, U_i)$. By d-separation, the graph implies $Z_i \perp\!\!\!\perp U_i | X_i, M_i$. By Lemma 3, this implies $Y_i(D_i = d), D_i(Z_i = z) \perp\!\!\!\perp Z_i | X_i, M_i$. Identification of the X_i, M_i-specific LATE then follows as in Angrist, Imbens and Rubin (1996).

In graph (b) of figure 2, $Y_i(D_i = d) = f_y(d, M_i, X_i, U_i) = f_y(d, f_m(Z_i, X_i, U_i), X_i, U_i)$, which depends on Z_i. Conditioning on X_i does not block this dependency. Conditioning on X_i, M_i makes Z_i and U_i dependent, so the CIA is generally violated. However, $D_i(Z_i = z) = f_d(z, X_i, U_i)$, and $D_i \perp\!\!\!\perp U_i | X_i$ by d-separation, so $Z_i \perp\!\!\!\perp D_i(Z_i = z) | X_i$ holds and the ATE of Z_i on D_i is identified.

In graph (c) of figure 2, $Y_i(D_i = d) = f_y(d, X_i, U_i)$ and $D_i(Z_i = z) = f_d(z, X_i, U_i)$. d-separation implies $Z_i \perp\!\!\!\perp U_i | X_i$, so by lemma 3, $Y_i(D_i = d), D_i(Z_i = z) \perp\!\!\!\perp Z_i | X_i$. However, conditioning on M_i makes Z_i and U_i dependent, because we are conditioning on a descendant of a collider.

In figure 3, we have

$$Y_i(D_i = d), D_i(Z_i = z, M_i = m) \perp\!\!\!\perp Z_i | X_i, M_i$$

(CIA.2)

First, in this graph, $Y_i(D_i = d) = f_y(d, M_i, X_i, U_i)$ and $D_i(Z_i = z, M_i = m) = f_d(z, m, X_i, U_i)$. By d-separation, we have $Z_i \perp\!\!\!\perp U_i | X_i, M_i$. Lemma 3 then implies CIA.2. Additionally, we assume

$P(D_i(Z_i = 1, M_i = m) \geq D_i(Z_i = 0, M_i = m)) = 1$ for all m (partial monotonicity)

$E[D_i|Z_i = 1, M_i = m, X_i] - E[D_i|Z_i = 0, M_i = m, X_i] \neq 0$ for all m (relevance)

Consider the X_i, M_i-adjusted Wald estimator

$$\frac{E[Y_i|Z_i = 1, M_i = m, X_i] - E[Y_i|Z_i = 0, M_i = m, X_i]}{E[D_i|Z_i = 1, M_i = m, X_i] - E[D_i|Z = 0, M_i = m, X_i]}$$

Under the above assumptions, the numerator evaluates to

$$E[Y_i|Z_i = 1, M_i = m, X_i] - E[Y_i|Z_i = 0, M_i = m, X_i] =$$

$$E[(Y_i(D = 1) - Y_i(D = 0))(D_i(Z_i = 1, M_i = m) - D_i(Z_i = 0, M_i = m))|M_i = m, X_i] =$$

$$E[Y_i(D = 1) - Y_i(D = 0)|D_i(Z_i = 1, M_i = m, X_i) > D_i(Z_i = 0, M_i = m, X_i)] \times$$

$$P(D_i(Z_i = 1, M_i = m) > D_i(Z_i = 0, M_i = m)|X_i).$$

The first step follows from

$$E[Y_i|Z_i = z, M_i = m, X_i] =$$

$$E[Y_i(D_i = 0) + (Y_i(D = 1) - Y_i(D = 0))D_i(Z_i = z, M_i = m)|Z_i = z, M_i = m, X_i],$$

for $z = 0, 1$ and CIA.2. The second uses the fact that $D_i(Z_i = 1, M_i = m) - D_i(Z_i = 0, M_i = m)$ is either one or zero by partial monotonicity.

The denominator is

$$E[D_i(Z_i = 1, M_i = m)|M_i = m, X_i] - E[D_i(Z_i = 0, M_i = m)|M_i = m, X_i] =$$

$$P(D_i(Z_i = 1, M_i = m) > D_i(Z_i = 0, M_i = m)|M_i = m, X_i)$$

The first step follows from consistency and CIA.2, and the second step follows from partial monotonicity. Accordingly, the Wald estimator evaluates to

$$E[Y_i(D = 1) - Y_i(D = 0)|D_i(Z_i = 1, M_i = m) > D_i(Z_i = 0, M_i = m)|X_i].$$

□

D Derivation of the Sensitivity Analysis

The structural models in equations 2–5 suggest estimation of all regression functions using linear models where the control variables X_i enter separately. Therefore, we leave the conditioning on X_i implicit in the following; all variables can be thought of as having partialled out their correlation with X_i. Consistent with this, we also assume that our sensitivity parameters are independent from X_i (see Knox, Lowe and Mummolo (2020, p. 11) for a similar approach).

Sensitivity model 1, in contrast to model 2, implies no assumptions on the functional form of $E[D_i|Z_i, X_i]$ and $E[M_i|Z_i, X_i]$. Then, two-stage least squares regression nonetheless is robust (at least if the true values of the sensitivity parameter were known) (Vansteelandt and Didelez 2018, Proposition 3).

D.1 Model 1: Binary Z_i, M_i, D_i

In addition to the model in equation 2, we here assume

$$Z_i \perp\!\!\!\perp Y_i(d,m), D_i(z), M_i(z) \text{ for all } z, d, m \tag{8}$$

$$P(D_i(Z_i = 1) \geq D_i(Z_i = 0)) = 1 \tag{9}$$

$$P(M_i(Z_i = 1) \geq M_i(Z_i = 0)) = 1 \tag{10}$$

Under these assumptions, we have

$$\begin{aligned}
&E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] = \\
&\qquad E[\beta_i(D_i(Z_i = 1) - D_i(Z_i = 0))] + E[\gamma_i(M_i(Z_i = 1) - M_i(Z_i = 0))] = \\
&\qquad E[\beta_i|D_i(Z_i = 1) > D_i(Z_i = 0)]P(D_i(Z_i = 1) > D_i(Z_i = 0))+ \\
&\qquad E[\gamma_i|M_i(Z_i = 1) > M_i(Z_i = 0)]P(M_i(Z_i = 1) > M_i(Z_i = 0)).
\end{aligned} \tag{11}$$

The first equality follows from the model equation and assumption 8. The second equality follows the monotonicity assumptions 9 and 10.

By the exogeneity assumption 8, $P(M_i(Z_i = 1) > M_i(Z_i = 0))$ and $P(M_i(Z_i = 1) > M_i(Z_i = 0))$ are identified as $E[D_i|Z_i = 1] - E[D_i|Z_i = 1]$ and $E[M_i|Z_i = 1] - E[M_i|Z_i = 1]$. Combining this, we have that

$$\begin{aligned}
&E[\beta_i|D_i(Z_i = 1) > D_i(Z_i = 0)] = \\
&\frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z = 0]} - \frac{E[\gamma_i|M_i(Z_i = 1) > M_i(Z_i = 0)](E[M_i|Z_i = 1] - E[M_i|Z_i = 1])}{E[D_i|Z_i = 1] - E[D_i|Z_i = 1]}.
\end{aligned} \tag{12}$$

Here, $E[\beta_i|D_i(Z_i = 1) > D_i(Z_i = 0)]$ is the LATE of interest, $\frac{E[Y_i|Z_i=1]-E[Y_i|Z_i=0]}{E[D_i|Z_i=1]-E[D_i|Z=0]}$ is a standard Wald (two-stage least squares) estimator with outcome Y_i, treatment D_i, and instrument Z_i, $E[\gamma_i|M_i(Z_i = 1) > M_i(Z_i = 0)]$ is the sensitivity parameter, and

$$\frac{E[M_i|Z_i = 1] - E[M_i|Z_i = 1]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 1]}$$

can be estimated by a two-stage least squares regression with outcome M_i, treatment D_i, and instrument Z_i.

D.2 Model 2: Binary Z_i, Continuous M_i, D_i

Here, our assumptions in addition to the model in equations 2–5 are

$$Z_i \perp\!\!\!\perp (\beta_i, \gamma_i, \alpha_i, \pi_i, \delta_i, \epsilon_{1i}, \epsilon_{2i}, \epsilon_{3i}) \tag{13}$$

$$P(\alpha_i + \delta_i \pi_i \geq 0) = 1 \tag{14}$$

$$P(\delta_i \geq 0) = 1 \tag{15}$$

$$cov(M_i(0), M_i(1)) \geq 0. \tag{16}$$

Under these assumptions, we have

$$\begin{aligned} E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] = \\ E[\beta_i(\alpha_i + \delta_i \pi_i)] + E[\delta_i \gamma_i]. \end{aligned} \tag{17}$$

This holds because Z_i is independent from all causal effects and the error terms.

$$E[\delta_i \gamma_i]$$

is the bias term we need to bound.

Note that with Model 1 (with binary D_i, M_i), we would have $E[\delta_i \gamma_i] = E[\gamma_i|\delta_i = 1]P(\delta_i =$

$1) = E[\gamma_i|\delta_i = 1](E[M_i|Z_i = 1] - E[M_i|Z_i = 1])$. This explains why we have only one sensitivity parameter in Model 1, whereas the next section shows that we have two unknown parameters in Model 2.

Using similar reasoning as before, we also have

$$E[D_i|Z_i = 1] - E[D_i|Z_i = 1] = E[\alpha_i + \delta_i\pi_i] \tag{18}$$

and

$$E[M_i|Z_i = 1] - E[M_i|Z_i = 0] = E[\delta_i]. \tag{19}$$

D.2.1 With Measured M_i

Rewrite the bias term as

$$E[\delta_i\gamma_i] = cov(\delta_i, \gamma_i) + E[\delta_i]E[\gamma_i]. \tag{20}$$

In the second term, $E[\delta_i]$ is point-identified as $E[M_i|Z_i = 1] - E[M_i|Z_i = 0]$, while $E[\gamma_i]$ will be a sensitivity parameter.

Further rewrite

$$cov(\delta_i, \gamma_i) = cor(\delta_i, \gamma_i)\sigma_{\delta_i}\sigma_{\gamma_i}. \tag{21}$$

In this latter term, we can decompose σ_{δ_i} as

$$\sqrt{var(M_i(1)) + var(M_i(0)) - 2cov(M_i(1), M_i(0))}. \tag{22}$$

The variance terms are nonparametrically point-identified as $var(M_i|Z_i = z)$. Regarding the covariance, intuition might suggest that monotonicity ($M_i(1) \geq M_i(0)$) implies that it is positive, but one can create joint distributions of $(M_i(1), M_i(0))$ where this is not the case. However, the Frechét-Hoeffding bounds (e.g. Aronow et al. (2014)) for this quantity using the marginals are not sharp, because the monotonicity does in fact improve the lower bound. Very recent work characterizes this lower bound under monotonicity (Nutz and Wang 2020). Since we are not aware of research on how to estimate this bound, especially with covariates, we make the simplifying assumption that $cov(M_i(1), M_i(0)) \geq 0$. We evaluate the consequences and the plausiblity of this assumption in section D.3. Using this assumption, an upper bound for equation 22 is

$$\sqrt{var(M_i|Z_i = 1) + var(M_i|Z_i = 0)}. \tag{23}$$

Further using $-1 \leq cor(\delta_i, \gamma_i) \leq 1$, we can bound equation 21 as

$$\begin{aligned} -\sqrt{(var(M_i|Z_i = 1) + var(M_i|Z_i = 0))}\sigma_{\gamma_i} \\ \leq cov(\delta_i, \gamma_i) \leq \\ \sqrt{(var(M_i|Z_i = 1) + var(M_i|Z_i = 0))}\sigma_{\gamma_i}, \end{aligned} \tag{24}$$

where σ_{γ_i}, the standard deviation of the direct causal effect of M_i on Y_i, is the second sensitivity parameter.

Collecting terms and rearranging, we have

$$\begin{aligned} \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} - \frac{1}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} \times \\ \{(E[M_i|Z_i = 1] - E[M_i|Z_i = 0])E[\gamma_i] + \sqrt{var(M|Z = 1) + var(M|Z = 0)}\sigma_{\gamma_i}\} \\ \leq E\left[\frac{\alpha_i + \delta_i\pi_i}{E[\alpha_i + \delta_i\pi_i]}\beta_i\right] \leq \\ \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} - \frac{1}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} \times \\ \{(E[M_i|Z_i = 1] - E[M_i|Z_i = 0])E[\gamma_i] - \sqrt{var(M|Z = 1) + var(M|Z = 0)}\sigma_{\gamma_i}\}, \end{aligned} \tag{25}$$

if $\dfrac{E[M_i|Z_i=1]-E[M_i|Z_i=0]}{E[D_i|Z_i=1]-E[D_i|Z_i=0]}$ is positive. If it is negative, the inequality signs reverse.

D.2.2 With Mismeasured M_i

As before, we want to gain information on the bias term (equation 20) from the data. $E[\delta_i]$ remains identified under the measurement model in equation 7 and the stated assumptions on the measurement error: $E[M_i^{\star}|Z=1]-E[M_i^{\star}|Z=0]=E[M_i+\eta_i|Z=1]-E[M_i+\eta_i|Z=0]=E[M_i|Z=1]-E[M_i|Z=0]=E[\delta_i]$.

It further turns out that the variances $var(M_i(z))$ are not point-identified anymore, although they can be bounded from above by the same quantities as in the case without measurement error. Accordingly, the resulting bounds for the sensitivity analysis do not change. To see why, consider

$$\begin{aligned} var(M_i(z)) = var(M_i|Z_i=z) = var(M_i^{\star}-\eta_i|Z=z) = \\ var(M_i^{\star}|Z=z)+var(\eta_i|Z=z)-2cov(M_i^{\star},\eta_i|Z=z) = \\ var(M_i^{\star}|Z=z)+var(\eta_i)-2cov(M_i^{\star},\eta_i|Z=z). \end{aligned} \tag{26}$$

Regarding this last term, we have

$$\begin{aligned} cov(M_i^{\star},\eta_i|Z=z) = cov(M_i+\eta_i,\eta_i|Z=z) = \\ cov(M_i,\eta_i|Z=z)+var(\eta_i|Z=z) = var(\eta_i). \end{aligned} \tag{27}$$

Accordingly,

$$var(M_i(z)) = var(M_i^{\star}|Z=z)-var(\eta_i) \leq var(M_i^{\star}|Z=z). \tag{28}$$

This bound could be improved upon if we could improve the trivial zero lower bound for $var(\eta_i)$. However, it is only possible to bound $var(\eta_i)$ from above using $var(M_i)$.

In sum, equation 28 shows that the observed conditional variance of the measurement is equal to or larger than the marginal variance of the potential outcome of the actual M_i variable. If measurement error is large, the empirical estimate will be far away from zero, even though the true marginal variance might be close or equal to zero. This is the information loss incurred by the measurement error.

Accordingly, the bounds in equation 25 remain valid, substituting $M_i^\star$ for M_i.

D.3 Understanding $cov(M_i(1), M_i(0)) > 0$

We here show how to understand the assumption that $cov(M_i(1), M_i(0)) \geq 0$, how to detect possible violations to it, and how to incorporate those into the sensitivity analysis.

First, the assumption that $cov(M_i(1), M_i(0)) \geq 0$ decreases the width of the bounds for the causal effect of interest, but has no effect on the location of the bounds. To see why, consider again our expression for $cov(\delta_i, \gamma_i)$, which is one part of the bias term:

$$cov(\delta_i, \gamma_i) = cor(\delta_i, \gamma_i)\sigma_{\delta_i}\sigma_{\gamma_i}$$

The standard deviations are always non-negative. The correlation is unknown and between -1 and 1. Therefore, this covariance between the causal effects is always in the interval $[-\sigma_{\delta_i}\sigma_{\gamma_i}, \sigma_{\delta_i}\sigma_{\gamma_i}]$. Our analysis bounds σ_{δ_i} from above using the data. Given values of the sensitivity parameter σ_{γ_i}, this results in bounds centered at 0 that are "added" to the mean estimate (which already may include bias adjustments from the first sensitivity parameter).

The empirical bound for σ_{δ_i} is based on writing it as

$$\sqrt{var(M_i|Z=1) + var(M_i|Z=1) - 2cov(M_i(1), M_i(0))}.$$

Clearly, when the covariance is positive, this term becomes smaller, and the width of the resulting bound $[-\sigma_{\delta_i}\sigma_{\gamma_i}, \sigma_{\delta_i}\sigma_{\gamma_i}]$ becomes smaller, too.

Second, to illustrate the relationship between the monotonicity assumption $M_i(1) \geq$

$M_i(0)$ and bounds on $cov(M_i(1), M_i(0))$, consider Figure A1. On the X- and Y-axis, we have values for potential outcomes $M_i(0)$ and $M_i(1)$, respectively. Without loss of generality, we assume here that these are between 0 and 1.

The dashed diagonal line graphs the monotonicity constraint $M_i(1) \geq M_i(0)$. We then plot the domains of two different joint distributions for $M_i(1), M_i(0)$. In both cases, $M_i(0)$ is uniformly distributed on $[0, 0.3]$, and therefore has a mean of 0.15. The domain of $M_i(1)$ differs between the two distributions, but it is always a finite closed interval. The dotted squares indicate the domains of all possible joint distributions given the domains of the marginal distributions.

The solid, piecewise linear function in the bottom left corner determines $M_i(1)$ as follows:

$$M_i(1) = \begin{cases} 0.4 - M_i(0) \text{ if } 0 \leq M_i(0) \leq 0.2 \\ M_i(0) \text{ if } 0.2 \leq M_i(0) \leq 0.3. \end{cases}$$

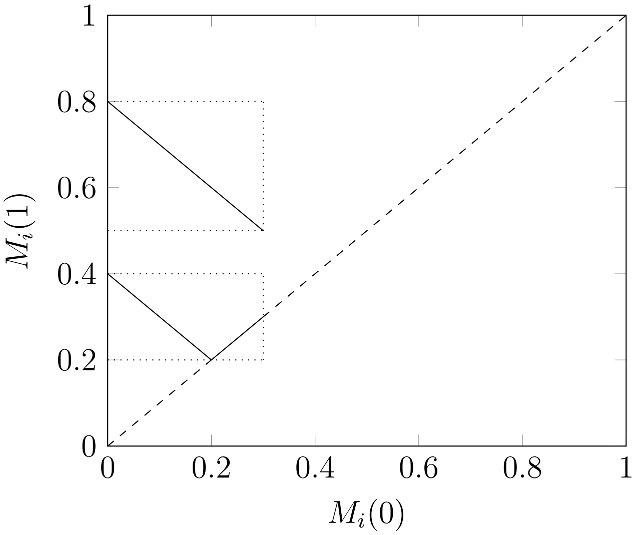

Figure A1: Understanding the relationship between the monotonicity constraint and the covariance between potential outcomes. Solid lines are the domains of two different joint distributions of $M_i(0), M_i(1)$ with negative covariance. Dotted lines indicate the domain of all possible joint distributions associated with each of these two cases. The dashed diagonal graphs the monotonicity constraint.

In this model, the average causal effect of Z_i on M_i is $\frac{1}{15}$. While the downward sloping part of the function contributes to a negative covariance, it cannot cross the monotonicity constraint, and the upwarding sloping part of the function then increases the covariance. Clearly, the monotonicity constraint restricts the covariance from becoming very negative.

To make the covariance more negative, one could shift $M_i(1)$ upwards so that the monotonicity constraint is without consequence. The second line towards the top plots such a function ($M_i(1) = 0.8 - M_i(0)$). Since the distribution of $M_i(0)$ does not change, the average causal effect here is much larger (0.55)

This suggests that while the monotonicity condition does not ensure that $cov(M_i(1), M_i(0))$ is actually positive, it suggests that a negative covariance is associated with large positive mean effects of Z_i on M_i.[17]

In sum, while the $cov(M_i(1), M_i(0)) \geq 0$ assumption used to bound σ_{δ_i} from above may not automatically hold under our monotonicity assumption, violations of it are likely to occur together with a large mean effect of Z_i on M_i. The latter is identified from the data and directly incorporated into our sensitivity analysis. If analysts are not willing to impose $cov(M_i(1), M_i(0)) \geq 0$ and they find a large mean effect of Z_i on M_i, we therefore suggest that they allow for larger values of the second sensitivity parameter σ_{γ_i} than is otherwise plausible. This will increase the width of the bounds $[-\sigma_{\delta_i}\sigma_{\gamma_i}, \sigma_{\delta_i}\sigma_{\gamma_i}]$ and can therefore to some degree address concerns stemming from the fear that the covariance between the potential outcomes is negative.

[17] If the mean effect of Z_i were negative, the monotonicity constraint would reverse and would restrict the covariance from becoming too positive when mean effects are small.

E Relationship Between Different Monotonicity Assumptions

To assess the relationship between the traditional monotonicity assumption and partial monotonicity, consider the case of binary Z_i and binary M_i, and no covariates. In this case, a saturated structural model for D_i without any functional-form assumptions can be written

$$D_i = \alpha + \beta_{1i} Z_i + \beta_{2i} M_i + \beta_{3i} Z_i M_i + \epsilon_i$$

where $\alpha = E[D_i(Z_i = 0, M_i = 0)]$, $\beta_{1i} = D_i(Z_i = 1, M_i = 0) - D_i(Z_i = 0, M_i = 0)$, $\beta_{2i} = D_i(Z_i = 0, M_i = 1) - D_i(Z_i = 0, M_i = 0)$, and $\beta_{3i} = D_i(Z_i = 1, M_i = 1) - D_i(Z_i = 0, M_i = 1) - (D_i(Z_i = 1, M_i = 0) - D_i(Z_i = 0, M_i = 0))$.

Monotonicity requires $D_i(Z_i = 1) \geq D_i(Z_i = 0)$ for all i, which restricts the total effect of Z_i on D_i. This is equivalent to stating that $\beta_{1i} + \beta_{2i} M_i(Z_i = 1) + \beta_{3i} M_i(Z_i = 1) \geq \beta_{2i} M_i(Z_i = 0)$ for all i. This restricts the joint distribution of $(\beta_{1i}, \beta_{2i}, \beta_{3i}, M_i(Z_i = 0), M_i(Z_i = 1))$. Note that the $M_i(z)$ will generally be associated with the coefficients when M_i and D_i are confounded, but this is ruled out by the assumptions we present to identify the new LATE.

Partial monotonicity is equivalent to the requirement that $\beta_{1i} + \beta_{3i} m \geq 0$ for all m and i, where m is constant. This restricts the direct effect of Z_i on D_i not going through M_i to be in the same direction for all m. This restricts the distribution of (β_{1i}, β_{3i}). In theory, there could be fine-tuned distributions of $(\beta_{1i}, \beta_{2i}, \beta_{3i}, M_i(Z_i = 0), M_i(Z_i = 1))$ where monotonicity holds but partial monotonicity does not. However, it seems natural to assume that the restrictions on β_{1i}, β_{3i} also hold when suitable restrictions on $(\beta_{1i}, \beta_{2i}, \beta_{3i}, M_i(Z_i = 0), M_i(Z_i = 1))$ are plausible.

F Implementation & Statistical Inference in the Sensitivity Analysis

For implementing the sensitivity analysis, we need to make a number of choices for estimation and inference. As stated before, and consistent with most IV applications, estimation of the mean differences in equation 25 can be pursued using two-stage least squares. For the variance terms, we pick corresponding linear conditional variance models (Shalizi 2019, 217). We first estimate auxiliary mean regressions

$$E[M_i|Z_i, X_i] = \zeta_1 + \zeta_2 Z_i + (X_i - \bar{X}_i)\zeta_3$$

where ζ_1, ζ_2 are scalars and ζ_3 is a vector, and generate residuals $r_i = M_i - \widehat{E}[M_i|Z_i, X_i]$. We then estimate $var(M|Z = z)$ via

$$E[r_i^2|Z_i, X_i] = \zeta_4 + \zeta_5 Z_i + (X_i - \bar{X}_i)\zeta_6$$

where ζ_4, ζ_5 are scalars and ζ_6 is a vector. Under this model, we have

$$var(M|Z = z) = \int_x \zeta_4 + \zeta_5 z + (x - \bar{X}_i)\zeta_6 dx = \zeta_4 + \zeta_5 z,$$

so that our estimate for $var(M|Z = 1) + var(M|Z = 0)$ is $2\zeta_4 + \zeta_5 z$.

Finally, we use the nonparametric (paired) bootstrap to estimate the sampling distribution of the resulting estimator for the bounds. Young (2019) has recently documented the widespread presence of non-IID errors in two-stage least squares regressions and suggests the bootstrap to improve statistical inference relative to standard robust covariance estimators.

Specifically, we implement the "bootstrap-c", which involves drawing from the original sample with resampling and storing point estimates $\widehat{\beta}$ of the bounds based on this sample. We base statistical inference on both "percentile" and the adjusted, "basic" bootstrap confidence interval (Davison and Hinkley 1997, 193–202), using the `R` package `boot`.

We find that the percentile method very closely replicates the results of Spenkuch and Tillmann (2017). In our replication of Carnegie and Marinov (2017), we find that an approach that uses a simple (non-clustered) bootstrap, percentile confidence intervals, and that removes outliers from the bootstrap distribution according to the standard boxplot definition (i.e., samples that are more than 1.5 inter-quartile ranges below the 25% or above the 75% quantile) also closely replicates the original results. Some remaining minor discrepancies may be due to slight changes in the sample composition, since we have to delete some observations for which measures of M_i are not available (as explained in the main text, Carnegie and Marinov (2017) did not adjust for M_i in their main analysis, but they evaluated its association with the instrument). All in all, this suggests that our approach can closely replicate results from standard two-stage least squares implementations, and may be additionally more robust to outliers as suggested by Young (2019).

G References

Angrist, Joshua D, Guido W Imbens and Donald B Rubin. 1996. "Identification of causal effects using instrumental variables." *Journal of the American statistical Association* 91(434):444–455.

Aronow, Peter M, Donald P Green, Donald KK Lee et al. 2014. "Sharp bounds on the variance in randomized experiments." *The Annals of Statistics* 42(3):850–871.

Carnegie, Allison and Nikolay Marinov. 2017. "Foreign Aid, Human Rights, and Democracy Promotion: Evidence from a Natural Experiment." *American Journal of Political Science* 61(3):671–683.

Chernozhukov, Victor, Iván Fernández-Val, Jinyong Hahn and Whitney Newey. 2013. "Average and quantile effects in nonseparable panel models." *Econometrica* 81(2):535–580.

Davison, Anthony Christopher and David Victor Hinkley. 1997. *Bootstrap methods and their application.* Number 1 Cambridge university press.

Dawid, A Philip. 1979. "Conditional independence in statistical theory." *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 1–31.

Geiger, Dan, Thomas Verma and Judea Pearl. 1990. "Identifying independence in Bayesian networks." *Networks* 20(5):507–534.

Imbens, Guido W. 2014. "Instrumental Variables: An Econometrician's Perspective." *Statistical Science* 29(3):323–358.

Imbens, Guido W and Whitney K Newey. 2009. "Identification and estimation of triangular simultaneous equations models without additivity." *Econometrica* 77(5):1481–1512.

Knox, Dean, Will Lowe and Jonathan Mummolo. 2020. "Administrative Records Mask Racially Biased Policing." *American Political Science Review* p. 1–19.

Nutz, Marcel and Ruodu Wang. 2020. "The Directional Optimal Transport.".

Pearl, Judea. 2009. *Causality*. Cambridge university press.

Shalizi, Cosma. 2019. "Advanced data analysis from an elementary point of view.".
URL: *http://www.stat.cmu.edu/ cshalizi/ADAfaEPoV/ADAfaEPoV.pdf*

Shpitser, Ilya, Tyler VanderWeele and James M Robins. 2010. On the validity of covariate adjustment for estimating causal effects. In *26th Conference on Uncertainty in Artificial Intelligence, UAI 2010*. pp. 527–536.

Spenkuch, Jörg L and Philipp Tillmann. 2017. "Elite Influence? Religion and the Electoral Success of the Nazis." *American Journal of Political Science* .

Vansteelandt, Stijn and Vanessa Didelez. 2018. "Improving the robustness and efficiency of covariate-adjusted linear instrumental variable estimators." *Scandinavian Journal of Statistics* 45(4):941–961.

White, Halbert and Xun Lu. 2011. "Causal diagrams for treatment effect estimation with application to efficient covariate selection." *Review of Economics and Statistics* 93(4):1453–1459.

Young, Alwyn. 2019. "Consistency without inference: Instrumental variables in practical application.".