Post-Instrument Bias

February 4, 2025

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 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 "post-instrument" bias. We then discuss assumptions that are sufficient to identify various treatment effects when adjustment for post-instrument variables is required. In general, these assumptions are highly restrictive, albeit they sometimes are testable. We also show that other existing tests are possibly misleading. Then, we introduce a 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 using a published study and summarize our results in easy-to-understand guidelines.

Word count: 9,480

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–variables influenced by the instrument–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 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 common practice and voiced concerns of researchers who use instrumental variables. We have identified 154 papers published since 2010 in top political science journals that use IV and explicitly discuss the exclusion restriction.[1] Among those, 39 (25%) use potential post-instrument covariates to justify the exclusion restriction.[2] As we will show, this is a lower bound on the phenomenon: There may be other

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

[2] See Appendix A. 22 of these papers mention the post-instrument nature of controls

papers that did not discuss relevant post-instrument covariates but should have considered them.

Some researchers are aware that adjustment for variables on other paths from instrument to outcome posits a thorny issue. For example, both Kern and Hainmueller (2009) and Carnegie and Marinov (2017) use instrumental variables and two-stage least-squares regression where they choose to not (always) 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 simply controlling 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 (Berinsky and Chatfield 2015). 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 with 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 may be concerns about the validity of the exclusion restriction. For example, those who received a low lottery number may have chosen to stay in school to obtain a deferment

explicitly. E.g., "I also control for possible direct channels"(Boix 2011, 818); "this channel is directly accounted for" (Ahmed 2012, 160). Felton and Stewart (2022) review IV papers published in top sociology journals and assess that 27 out of 34 include potential post-instrument covariates.

(Angrist 1990, 330). This creates a link between the lottery and earnings via education. Therefore, 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 potential outcomes and directed acyclic graphs (Pearl 2009) in our formal analysis. This allows us to provide straightforward advice to applied researchers. First, we clarify 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 arise.

The main intuition is as follows. Instrumental variables estimators adjusted for a post-instrument covariate compare observations with different values on the instrument but the same values on the post-instrument variable, even though the variation in the instrument produces variation in the latter variable. Therefore, other omitted causes of the post-instrumental variable need to vary across these observations. Otherwise, the post-instrumental variable would not have materialized to have the same value. Accordingly, the omitted causes co-vary with the instrument. Insofar as these omitted causes affect the outcome, this creates a non-causal link between instrument and outcome that leads to bias. Notably, this can occur even if the post-instrumental variable is exogenous conditional on the treatment, unlike in the standard post-treatment bias case.

However, we also show that adjustment for variables influenced by the instrument is sometimes *necessary* for successful identification when the post-instrumental variable is not impacted by unobserved confounders. 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", where adjustment can only hurt and never helps.

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 possibly misleading.

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. We illustrate our approach by reanalyzing the data of Hong, Park and Yang (2023) on the long-term effects of a rural development program on voting behavior in South Korea. The application highlights the need to relax stringent linearity assumptions and to account for potential heterogeneity in causal effects. We make our methodology available as an `R` package.

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), it did not incorporate post-instrument variables. Glynn, Rueda and Schuessler (2024) analyze post-instrument bias using linear constant-effect models, as do Deuchert and Huber (2017). In contrast, we discuss these issues in a completely nonparametric framework and integrate causal graphs with the potential outcomes approach. We show that Glynn et al.'s result about the magnitude of biases from the constant-effect case does not generalize once one allows for heterogeneous effects. Furthermore, we discuss additional identification assumptions, prove that these are sometimes testable, introduce a new causal estimand, and develop a new sensitivity analysis.

Deuchert and Huber (2017) 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, Mellon (2024) points out that weather-related variables like measures of rainfall are often used as instruments for various relationships, but have been linked empirically to close to 200 variables, each of which constitutes a potential violation to the exclusion restriction. Some of the problems that we discuss are similar to what Elwert and Segarra (2022) calls "endogenous selection bias", and Betz, Cook and Hollenbach (2018), Imai and Kim (2019) and Eggers, Tuñón and Dafoe (2024) also use causal graphs to illustrate (failures of) IV identification. Our sensitivity analysis complements the approaches by Conley, Hansen and Rossi (2012) and Cinelli and Hazlett (2022) that cannot incorporate information on post-instrument covariates. Among other things, this entails that our sensitivity analysis can make estimates more robust (i.e., move farther away from zero into the direction implied by the original estimate), which we also show in our application.

2 Understanding conditional IV identification using causal graphs

In this section, we present a series of causal graphs that allow for IV 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.

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 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 (SES) of one's parents. These will impact on 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 impossible. With this first example in mind, we now discuss basic quantities of interests and identification assumptions in the potential outcomes framework.

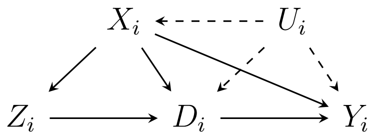

Figure 1: Benchmark graph. In this graph, Z_i is an instrument for the effect of D_i on Y_i conditional on X_i, but not unconditionally.

2.2 IV identification in the potential outcomes framework

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. 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 application study? 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 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 is 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, 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).

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)

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

which clearly depends on Z_i, so that the CIA is violated.

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.

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.[3] 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 possibly 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.

2.4.1 Post-instrument variable not impacted by unobserved confounders

A 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 out 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 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 con-

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

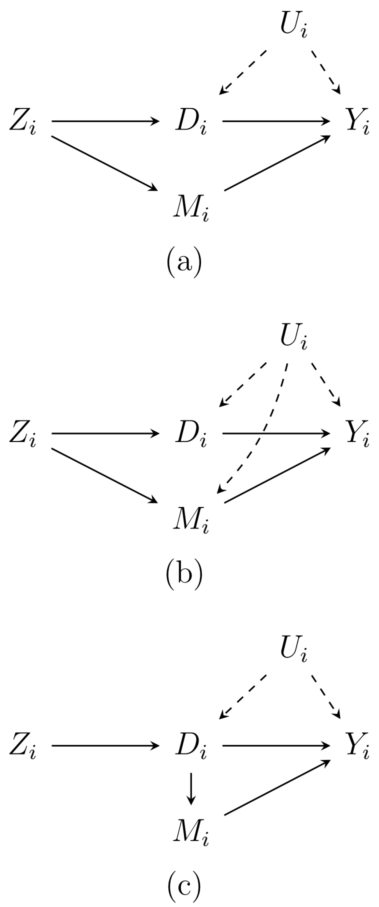

Figure 2: Three scenarios for relationships between candidate instrument Z_i and post-instrument variable M_i. In graph a), conditioning on M_i is required and identifies a local effect of D_i on Y_i. In graph b), neither unadjusted nor adjusted IV estimators identify a causal effect. Without control for M_i, there is a direct effect of the instrument. However, conditioning on the collider M_i opens a non-causal path between U_i and Z_i. In graph c), IV identification is possible only when not conditioning on M_i. M_i is a descendant of collider D_i and conditioning on it creates a non-causal dependence between Z_i and U_i

ditioning 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.[4]

There are two crucial assumptions for the validity of this approach that may be violated which we now discuss in turn.

2.4.2 Post-instrument variable directly impacted by unobserved confounders

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 (U_i) positively influences the choice to go to college directly (M_i). In this case, M_i becomes a "collider", and conditioning on it creates a statistical dependence between Z_i and U_i.

Specifically, in the "reduced-form" regression of Y_i on Z_i, 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 across groups 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,[5] as well as the $Y_i(D_i = d) \perp\!\!\!\perp Z_i | X_i, M_i$ assumption.

In situations described by this graph, both unadjusted and adjusted IV estimators are biased (inconsistent). How do these biases compare? Glynn, Rueda and Schuessler (2024) show, using linear models, that biases are proportional to the strength of the association between the instrument and the post-instrument variable multiplied with the effect of M_i

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

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

on Y_i (unadjusted estimator) or the endogeneity of M_i (adjusted estimator), divided by the strength of the instrument with respect to the treatment D_i.[6] This appears intuitive: If Z_i and M_i are only weakly related even if one adjusts for the strength of the IV (an issue to which we return below), the potential for bias should generally be low. Similarly, if the causal effect of M_i on Y_i is small, the direct effect of Z_i, and hence the bias of the unadjusted estimator, should be small, too. However, these results from the linear case–that imposes an assumption of no effect heterogeneity–do not generalize. We will later show analytically that with heterogeneous effects, the unadjusted estimator may be biased even if the mean effects of Z_i on M_i or of M_i on Y_i (or both) are zero. Furthermore, in Appendix D, we analyze the biases in more detail from a potential outcomes perspective and show that even if the effect of Z_i on M_i is small, the bias of the adjusted estimator can also be large. Consistent with this, we show that, in our application study, unadjusted and adjusted IV estimates barely differ, yet our sensitivity analysis points towards potential biases.

2.4.3 Post-instrument variable directly impacted by treatment

The second crucial identification assumption concerns the relationship between treatment and adjustment variable. 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 $D_i \to Y_i$ relationship, and is also influenced by Z_i indirectly through D_i. For example, M_i could stand for mental and physical health, civilian work experience, or access to veterans' benefits. 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. However, adjusting for M_i introduces a severe, but more subtle problem. In the "reduced-form" regression of Y_i on Z_i controlling for M_i, we would again compare

[6] Accordingly, the biases of the two different estimators are possibly of the same order of magnitude, unlike in the linear "M-bias" case, where the bias of the adjusted estimator is of a higher order (smaller) (Ding and Miratrix 2015).

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 now mediated through D_i, and not due to a causal effect of D_i. E.g., we could compare draftees to non-draftees with equally sound mental health post-Vietnam. Because of the deleterious impact of active military service on mental health, it would appear likely that all of them had in fact not served in Vietnam, despite the differences in draft status. Accordingly, unobserved causes U_i would be lower for those who were drafted ($Z_i = 1$), explaining why they did not serve after all ($D_i = 0$), than for those who were not drafted. Therefore, Z_i and U_i would co-vary, creating a non-causal link between the instrument and the outcome. More formally, 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.

This subtle problem went unnoticed by Deuchert and Huber (2017, 416), who discuss a similar graph and state that conditioning on a mediator identifies a "partial direct effect" (Wooldridge (2010, 95) appears to make a similar suggestion). As we hope we have made clear, this is not the case, because conditioning on a mediator renders Z_i correlated with U_i, invalidating its use as an instrument. Interestingly, this occurs here even though the mediator is exogenous conditional on D_i. That is, a suitable regression of Y_i on D_i and M_i identifies the average causal effect of M_i. This shows that an analyst using such a regression as the baseline model for the effect of D_i and then using an otherwise valid instrument to address concerns about the endogeneity of D_i (but not of M_i) introduces problems by simply mirroring the set of control variables. Whenever analysts consider an IV strategy, they need to reanalyse the choice of control variable based on a causal model.[7]

An interesting special case of graph c) of Figure 2 is when M_i stands for the inclusion

[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.

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 (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. This means one conditions on a descendant of the true unobserved earnings so that the IV becomes invalid, 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]

2.4.4 Post-instrument variable impacts on treatment, but is unrelated to confounders

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

[8] See Elwert and Segarra (2022) for an analysis of this problem under a linearity assumption.

to the instrument Z_i, *while fixing* M_i *at* m and 2) which are characterized by covariates X_i.[9]

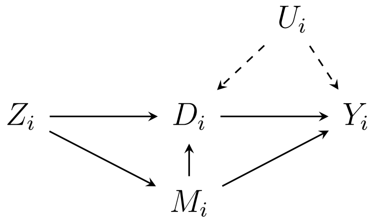

Figure 3: Graph where adjustment for M_i is necessary to identify a local average treatment effect.

The intuition behind this identification result is that under the assumptions in graph 3, one can actually identify the "controlled direct" effect of Z_i on D_i while fixing M_i. For those individuals that shift their treatment uptake as a result of this hypothetical joint intervention, the effect of D_i on Y_i is then also identified. There are additional relevance and monotonicity assumptions needed. We discuss these in more detail in Appendices C and E.6.

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.

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

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 directly affect M_i and Y_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 discuss the validity of robustness and diagnostic tests employed by researchers facing post-instrument variables.

Kern and Hainmueller (2009), Wucherpfennig, Hunziker and Cederman (2016), and Spenkuch and Tillmann (2018), among others, acknowledge the possibility of 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. This is also the case in our application study. Such a testing strategy is informative insofar as it rules out that the biases of the two different estimators are unequal. Finding that estimates are indeed similar, the remaining possibilities then are that the biases happen to be non-zero, but equal, or simply are both zero. Therefore, the probability that biases are both zero increases. Yet we maintain that analysts should never assign a zero probability to a scenario where the different IV estimators have biases of a similar magnitude. Our sensitivity analysis adds evidence in this cases.

Furthermore, a problem with this testing strategy is that if one finds that estimates are not equal, one may erroneously declare the main result to be not robust. This is the case whenever the IV is unconditionally valid but invalid conditional on M_i, as in Figure 2 c), where M_i is a mediator on the pathway from D_i to Y_i. In such cases, the adjusted IV estimator will generally differ, but this is because conditioning on M_i introduces biases where none were before. Accordingly, such robustness tests need to be guided by an analysis of a specific DAG.

A second approach is to inspect the correlation between Z_i and M_i. Researchers often report that this association is not significant and that the instrument is therefore unconditionally valid. However, the bias introduced by direct effects running over post-instrument variables increases as the instrument becomes weaker with respect to the treatment, as discussed below, which such tests do not address. Accordingly, small insignificant correlations may become larger and significant once one adjusts for the first-stage relationship between instrument and treatment. This is the case in our application study. Additionally, even small effects of Z_i on M_i can introduce bias when the effect of M_i on Y_i is large, as shown below.

One situation in which causal assumptions we have proposed are sharp enough that they allow for a valid 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 the first-stage regression$E[D_i|M_i, Z_i, X_i]$. However, the focus normally rests on the partial association between the instrument Z_i and D_i (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 (assuming correct regression specification and standard errors). If researchers commit to this graph, they should use an equivalence test in order to provide evidence for this zero association (Hartman and Hidalgo 2018), for example by determining whether the 90% confidence interval lies entirely within a range of associations that are negligible (at $\alpha = 0.05$). This test (which will we call the "diagnostic test") may seem counter-intuitive 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). If the test fails, at least one open path between D_i and M_i must exist, as in Figure 2 b) and c) or Figure 3.

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. This is the case in Figure 2 b). Identification is possible in the DAGs in Figures 2 a) (with control for M_i, if the diagnostic test passes), 2 c) (without control for M_i), or Figure 3 (with control for M_i). Analysts need to make a theoretical argument for why such a specific DAG appears plausible. If it does not, we propose a new semi-parametric sensitivity analysis for situations such as in Figure 2 b). 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 approaches (Conley, Hansen and Rossi 2012; Van Kippersluis and Rietveld 2018; Cinelli and Hazlett 2022) that cannot use information on post-instrument covariates. An interesting corollary of our approach is that for at least some choices of the sensitivity parameters, estimates are guaranteed to become more robust (i.e., move further away from zero into the direction of the original point estimate). Furthermore, we relax parametric assumptions (e.g., constant effects) that are often made in the literature. We present two different models: First, a model for situations where instrument, treatment, and post-instrument variable are binary. Then, there is only one sensitivity parameter. Second, a model for a binary instrument, but possibly continuous treatment and post-instrument variable. Then, there are two sensitivity parameters. We extend our approach to multiple arbitrarily distributed instruments in Appendix E.3.

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. The resulting estimation approach is a special case of our second approach described in the next section.

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 E, 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 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). However, it is also clear that if one chooses the sign of the sensitivity parameter such that the bias term is of the opposite sign as the first term (the naive estimate), the resulting estimate will actually be in the same direction and larger than the naive estimate. Accordingly, original estimates will become more robust for some choices of the sensitivity parameter.

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.[10] 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

With continuous D_i or M_i, the previous 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.[11]

We make a series of further assumptions, which are enumerated in Appendix E. Here, we give an intuitive summary. The first assumption follows from graphs a) and b) in Figure

[10] 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). 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.

[11] 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.

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.6). 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 E.5.

Under these assumptions, we show in Appendix E 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.[12] 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 from zero (Glynn 2012). This shows that even if there is no mean causal effect of Z_i on M_i or M_i on Y_i, the unadjusted IV estimator may still be biased.

For example, in the Vietnam draft study, if unobserved parental SES U_i influences the

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

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). 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. 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 E 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.

Finally, we can extend this sensitivity model to situations where the post-instrument variable M may be measured with error. We discuss this in Appendix E.

3.3 Assessing values for the sensitivity parameters

To reiterate, the first sensitivity parameter $E[\gamma_i]$ describes the direct effect of M_i on Y_i, fixing D_i. We suggest that researchers reason about the sign and size of this parameter based on the literature studying the effect of the M_i on the Y_i variable, and we illustrate this below.

The second sensitivity parameter, σ_{γ_i}, is the standard deviation of γ_i. This parameter therefore describes the heterogeneity in the effects of M_i on Y_i that the first sensitivity parameters averages.[13] σ_{γ_i} is non-negative and increasing it does not change the mean effect estimate, but rather widens the confidence interval.

[13] The sensitivity analysis developed by Imai and Yamamoto (2013) contains a similar

This parameter is usually not identified in empirical studies. However, existing empirical studies are informative insofar as they document effect heterogeneity. If a study reports the effects of M_i on Y_i to vary in a substantively meaningful way as a function of another covariate, then this suggests that σ_{γ_i} is relatively large, although it is not possible to specify this quantitatively. We therefore suggest to inspect the existing literature for evidence of effect heterogeneity. To get a better quantitative sense of this sensitivity parameter, one can depart from the range the researcher specifies for the first sensitivity parameter (which similarly can be informed by prior literature). If one assumes that these represent the minimum and maximum values for unit-specific causal effects and one further assumes a certain shape for the distribution of these effects (e.g., uniform), then this yields a specific value for σ_{γ_i}. We discuss this issue, including its implementation in our `R` package, in more detail in Appendix E.4.

Quantitative robustness analysis does generally not yield clear qualitative answers on the (non-)robustness of a finding, but rather invites researchers to reason about robustness as a continuous concept. As such, researchers should not blindly insert any default values for the sensitivity parameters, but rather choose them as to change their main inference and then discuss, given substantive judgment and information from the prior literature as discussed above, whether such values for the sensitivity parameters are plausible. We illustrate this below.

3.4 Multiple post-instrument covariates

In some situations, there may be a worry that there are multiple potential post-instrument M_i variables. We analyze this in Appendix F. It turns out that if one is willing to assume that the different post-instrument variables do not causally influence each other, our sensitivity analysis can be extended relatively easily. However, such a causal independence assumption

parameter.

is very strong and untestable given the discussed assumptions. It appears unlikely that there could be cases where one is unwilling to rule out direct effects of the candidate instrument, yet willing to assume that these run over measured variables that happen not to influence each other. If one does not impose such an assumption, the analysis becomes practically intractable, as the number of sensitivity parameters grows very fast.[14] However, in either case, some of the additional sensitivity parameters describe effect heterogeneity (just as σ_{γ_i}) and can only ever widen, but not tighten, confidence intervals. While suitable choices for the other sensitivity parameters could counteract this in theory, it appears improper to put a lot of weight on such specific choices for multiple unknown sensitivity parameters. This implies that if one runs the sensitivity analysis with one post-instrument variable and finds that results are not robust, it is reasonable to infer that this would not be salvaged by extending the analyses to include multiple post-instrument variables. On the other hand, if results appear robust, analysts should point out that this may not hold if there are other post-instrument variables or direct effects.

4 An illustration of the proposed methodology

We illustrate our new sensitivity analysis using data from Hong, Park and Yang (2023). The authors are interested in how a rural development program administered by the South Korean dictator Park Chung-hee in the 1970s affected short-term and long-term election results, including the vote share of Park's daughter Park Geun-hye, who was democratically elected president in 2012. The program involved village members deciding on and investing labor and other assets in development projects. In the following, the government then paid out subsidies depending on the performance of these projects. The treatment variable measures logged subsidies per voter and the outcome is vote shares, both at the township level. The authors note that there may be omitted variables that impacted citizens' efforts and thereby

[14] E.g., with two post-instrument covariates, there are seven sensitivity parameters.

subsidies as well as their political ideology. The authors therefore use an instrumental variable approach based on the observation that villages in disadvantageous terrain had worse baseline infrastructure upon which they could more easily improve, attracting more subsidies. Indeed, the authors find a significant first-stage relationship between both the elevation and the slope of a village's terrain and the amount of subsidies. An estimate based on 2SLS using these two variables as instruments implies that a 1% increase in the subsidies in the 1970s led to a 6-point increase in the vote share of Park Geun-hye in 2012.

Hong, Park and Yang (2023) employ a placebo test using the 2007 vote share of a presidential candidate without a family relationship to Park Chung-hee as an outcome, for which causal effects (driven by nostalgia) should and indeed are estimated to be zero. They therefore argue that there are no direct effects of the instrument. We here provide an additional robustness check using our sensitivity analysis. The IV analyses adjust for the same variables as the OLS analyses, which may pose a problem insofar as they are influenced by the instrument. Since the instrumental variables measure fundamental geographical aspects that are essentially time-constant and may have significant societal downstream consequences, this appears quite likely. We focus on the share of female inhabitants in a given village measured in 1966, as sex ratios vary significantly as a function of urbanization, which is influenced by terrain features, due to differences in disease burden and economic incentives (Courtwright 2008). Using this variable, we can also illustrate some of the potential pitfalls of informal robustness tests discussed above.

Because this variable is measured before the onset of the treatment program, we can rule out that it is a mediator as in Figure 2 c). If we assumed instead it was an unconfounded post-instrument variable as in Figure 2 a), this could be checked by our diagnostic test. We replicate the authors' first-stage regression and find the 90% confidence interval for the association between M_i and D_i to be $[-4.34, -0.14]$. Since this interval includes large associations, we cannot reject the Null of a meaningful association between Z_i and M_i, and therefore the situation in Figure 2 a) appears implausible. It also appears theoretically possible that there are unadjusted confounders of M_i and Y_i. For example, the authors do not adjust for any variables that describe the economic situation of a village. Accordingly, we might face a situation such as Figure 2 b), for which both unadjusted and adjusted IV estimators are invalid.

For illustrative purposes, we estimate a 2SLS model where we leave M_i out. We find that estimates of the treatment effect barely change (adjusted model: b = 0.060, se = 0.026; unadjusted model: b = 0.057, se = 0.025). When we use M_i as the outcome in the authors' first-stage regression, we find that neither instrument is significantly associated with M_i. Accordingly, these informal robustness tests would not indicate any problem. However, when we use M_i as the outcome in the authors' 2SLS specification, which estimates one of the empirical parameters that is relevant for the sensitivity analysis, we find that instruments and post-instrumental variable are significantly associated once one adjusts for the first-stage relationship between instrument and treatment (b = 0.008, se = 0.004).

Consistent with this, the results from the sensitivity analysis in Figure 4 indicate that treatment effect estimates do vary as a function of the average causal effect of M_i on Y_i ($E[\gamma_i]$), the first sensitivity parameter, which is depicted on the X-axis. Since M_i and Y_i are both percentages, a causal effect of 1 indicates a very strong relationship. The different confidence intervals vary as a function of the second sensitivity parameter σ_{γ_i}. They are evenly spaced between the chosen minimum (0) and maximum (0.2) of value σ_{γ_i}.

At $E[\gamma_i] = 0$ and $\sigma_{\gamma_i} = 0$, we obtain a significant point estimate close to the original one. However, for values of $E[\gamma_i]$ below ca. -0.75, keeping σ_{γ_i} at zero, point estimates become smaller and insignificant, as the innermost confidence interval then overlaps with 0. Yet for positive values of this sensitivity parameter, estimates become actually more positive and significant. This is a general feature of our sensitivity analysis: Given a non-zero relationship between Z_i and M_i, point estimates will become larger in absolute terms for either all positive or all negative values of the first sensitivity parameter.

The two sensitivity parameters interact in determining robustness. For example, at

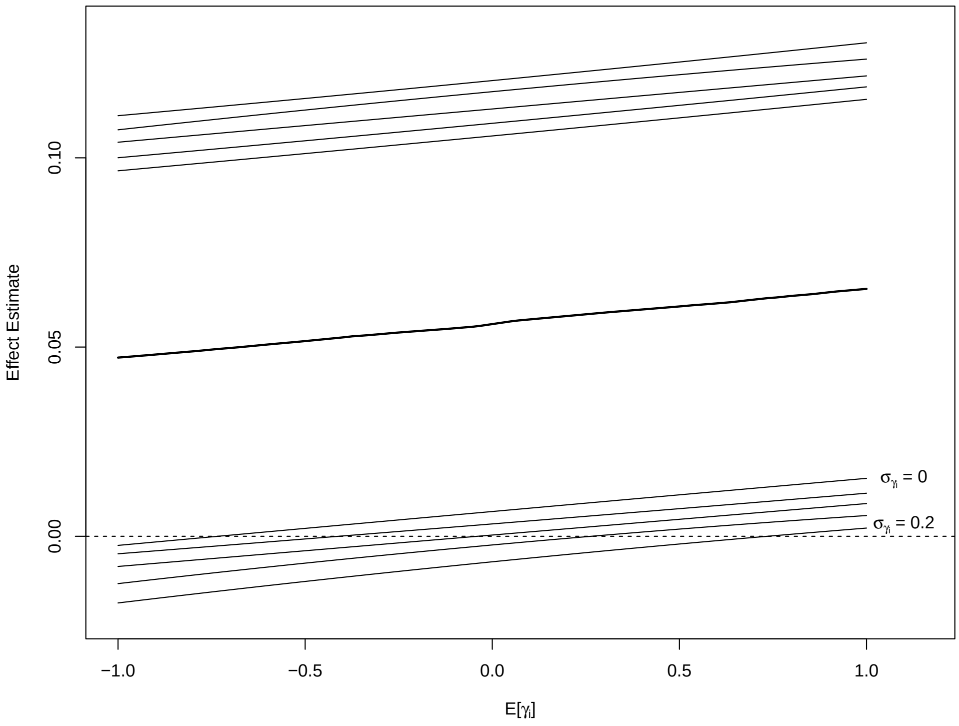

Figure 4: Results from the sensitivity analysis, based on data from Hong, Park and Yang (2023). X-axis depicts the first sensitivity parameter $E[\gamma_i]$, the average direct effect of M_i (share of females in 1966) on Y_i (2012 vote share). Y-axis depicts estimates of effect of interest of D_i (subsidies) on Y_i. Thick solid line represents the mean effect estimate. Thin solid lines represent 95% confidence interval as a function of the second sensitivity parameter σ_{y_i}, effect heterogeneity in the effect of M_i on Y_i. The different values of σ_{y_i} are evenly spaced between the depicted minimum and maximum of σ_{y_i}.

$E[\gamma_i] = 0$, the estimate becomes insignificant for σ_{γ_i} larger than ca. 0.15 (the third outermost confidence interval), while at $E[\gamma_i] = -0.5$, it becomes insignificant for σ_{γ_i} larger than ca. 0.1 (the second outermost confidence interval). So what are plausible values for the sensitivity parameters? Observational analyses of the 2012 election did not find gender effects in the presidential vote Kang (2018). Furthermore, the sensitivity parameters describe the effect of historic, not contemporary sex ratios. Overall, such average effects may therefore be small, and perhaps are more likely to be positive than negative, given evidence on "gender affinity effects" Dolan (2008). Indeed, in the authors' original 2SLS regression, the association between M_i and Y_i is positive but insignificant, although this estimate is very noisy and may suffer from bias. Regarding the variation in causal effects across townships (σ_{γ_i}), the literature on gender effects highlights that these vary significantly from context to context (Goodyear-Grant and Croskill 2011). If we pick a bell-shaped Beta distribution with unit-level causal effects between -0.1 or 0.4, the implied mean effect is $E[\gamma_i] = 0.15$ while $\sigma_{\gamma_i} \approx 0.08$. The resulting confidence interval is $[-0.01, 0.12]$ and the estimate is therefore insignificant.

5 Conclusion

Many researchers use instrumental variables in settings where they try to "control away" a direct effect of the instrument on the outcome by adjusting for post-instrument variables M_i. In this paper, we explained why this strategy only works under restrictive assumptions. Using potential outcomes and causal graphs, 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 and illustrated it using the IV analysis in Hong, Park and Yang (2023). Here, it became clear that researchers need to reason about both mean direct effects of the instrument as well as their variability.

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, proceed with standard estimation routines like 2SLS, where you condition only on X_i.

3. If your assumptions are equivalent to graph a) in Figure 2, implement the diagnostic test by providing evidence that D_i and M_i are independent conditional on Z_i.

4. If the test does not reject the Null, reconsider your assumptions. 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) in 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

Ahmed, Faisal Z. 2012. "The perils of unearned foreign income: Aid, remittances, and government survival." *American Political Science Review* 106(1):146–165.

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.

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.

Boix, Carles. 2011. "Democracy, development, and the international system." *American Political Science Review* 105(4):809–828.

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.

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.

Cinelli, Carlos and Chad Hazlett. 2022. "An omitted variable bias framework for sensitivity analysis of instrumental variables." *Available at SSRN 4217915* .

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

Courtwright, David T. 2008. "Gender imbalances in history: causes, consequences and social adjustment." *Reproductive BioMedicine Online* 16:32–40.

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.

Ding, Peng and Luke W Miratrix. 2015. "To adjust or not to adjust? Sensitivity analysis of M-bias and butterfly-bias." *Journal of Causal Inference* 3(1):41–57.

Dolan, Kathleen. 2008. "Is there a "gender affinity effect" in American politics? Information, affect, and candidate sex in US House elections." *Political Research Quarterly* 61(1):79–89.

Eggers, Andrew C, Guadalupe Tuñón and Allan Dafoe. 2024. "Placebo tests for causal inference." *American Journal of Political Science* 68(3):1106–1121.

Elwert, Felix and Elan Segarra. 2022. Instrumental variables with treatment-induced selection: Exact bias results. In *Probabilistic and Causal Inference: The Works of Judea Pearl.* pp. 575–592.

Felton, Chris and Brandon M Stewart. 2022. "Handle with care: a sociologist's guide to causal inference with instrumental variables." *Sociological Methods & Research* p. 00491241241235900.

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, Miguel R Rueda and Julian Schuessler. 2024. "Post-instrument bias in linear models." *Sociological Methods & Research* 53(4):1829–1845.

Goodyear-Grant, Elizabeth and Julie Croskill. 2011. "Gender affinity effects in vote choice in Westminster systems: Assessing "flexible" voters in Canada." *Politics & Gender* 7(2):223–250.

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

Hong, Ji Yeon, Sunkyoung Park and Hyunjoo Yang. 2023. "In strongman we trust: The political legacy of the new village movement in South Korea." *American Journal of Political Science* 67(4):850–866.

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

Kang, WooJin. 2018. "The past is long-lasting: Park Chung Hee nostalgia and voter choice in the 2012 Korean presidential election." *Journal of Asian and African Studies* 53(2):233–249.

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

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

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

Mellon, Jonathan. 2024. "Rain, rain, go away: 194 potential exclusion-restriction violations for studies using weather as an instrumental variable." *American Journal of Political Science* .

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. 2009. *Causality*. Cambridge university press.

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. 2018. "Elite influence? Religion and the electoral success of the Nazis." *American Journal of Political Science* 62(1):19–36.

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.

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 Paper 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, 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 representations of a nonparametric system of structural equations that describes cause-effect relationships. That is, nodes represent observable or unobservable features of units of interest, 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 precisely determine the value of this variable. 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 in which no variable may have an effect on itself. Finally, we use upper-case letters to denote random variables, and lower-case letters to denote realized or fixed values of these variables.

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.

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

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.

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 must 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.

B.2 From graphs to potential outcomes

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 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 $D_i \rightarrow Y_i \leftarrow U_i$. 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 White and Lu (2011). 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]

C Proof of the proposition

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.*

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

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 Analysis of post-instrument bias using potential outcomes

We here analyze the biases of the unadjusted and adjusted IV estimator in more detail from a potential outcomes perspective. We show that both estimators may be biased and that the bias of the adjusted estimator may be much larger than the bias of the unadjusted estimator, even if the instrument is only weakly associated with the post-instrument variable. We concentrate on expressions for the reduced form for binary Z_i, D_i, M_i. Let $Y_{zdm} = E[Y_i|Z_i = z, D_i = d, M_i = M], M_{zd} = E[M_i|Z_i = z, D_i = d], D_{zm} = E[D_i|Z_i = z, M_i = m], D_z = E[D_i|Z_i = z]$. We also assume consistency: $D_i = d \implies Y_i = Y_i(D_i = d)$. Accordingly, $E[Y_i|Z_i = z, D_i = d, M_i = M] = E[Y_i(D_i = d)|Z_i = z, D_i = d, M_i = M]$. We have

$$E[Y_i|Z_i = 1]-E[Y_i|Z_i = 0] = Y_{111}M_{11}D_1+Y_{101}M_{10}(1-D_1)+Y_{100}(1-M_{10})(1-D_1)+Y_{110}(1-M_{11})D_1- \\ (Y_{011}M_{01}D_0 + Y_{001}M_{00}(1 - D_0) + Y_{000}(1 - M_{00})(1 - D_0) + Y_{010}(1 - M_{01})D_0) \quad (7)$$

and

$$E[Y_i|Z_i = 1, M_i = 1] - E[Y_i|Z_i = 0, M_i = 1] = Y_{111}D_{11} + Y_{101}(1 - D_{11}) - (Y_{011}D_{01} + Y_{001}(1 - D_{01})). \quad (8)$$

Note that all of the average outcomes of Y_i that occur in the adjusted estimator occur also in the expression for the unadjusted estimator. The "free" average outcomes that feature only in the unadjusted estimator are $Y_{100}, Y_{110}, Y_{000}, Y_{010}$. Accordingly, in general, the biases of the two estimators will be unequal. Further simplifications of these expressions do not appear possible: Z_i, D_i, M_i are generally dependent, such that M_{zd} does not simplify. Additionally, if one considered the Wald estimator, no further terms would drop out, as, e.g., $E[D_i|Z_i = 1]-E[D_i|Z_i = 0] = D_{11}M_1+D_{10}(1-M_1)-(D_{01}M_0+D_{00}(1 - M_0))$.

Z_i	D_i	M_i	$E[Y_i(D_i=1)\|Z_i, D_i, M_i]$	$E[Y_i(D_i=0)\|Z_i, D_i, M_i]$
0	0	0		Y_{000}
1	0	0		Y_{100}
0	1	0	Y_{010}	
0	0	1		Y_{001}
1	1	0	Y_{110}	
0	1	1		Y_{011}
1	0	1	Y_{101}	
1	1	1	Y_{111}	

Table A2: Observed potential outcomes for Y_i that feature in both the unadjusted and adjusted IV estimator. Empty cells are unobserved potential outcomes.

Table A2 enumerates average potential outcomes for all eight subgroups defined by Z_i, D_i, M_i. It becomes clear that the estimators are only ever functions of one, but not both, potential outcomes in a subgroup. Accordingly, one could choose the values of the two different potential outcomes in each subgroup to be the same, such that all average causal effects are zero. Nevertheless, both estimators will generally be non-zero insofar as there are differences in average potential outcomes across the subgroups. This shows that both estimators are potentially biased.

The bias of the adjusted estimator can be much larger than the bias of the unadjusted estimator, even if the instrument is only weakly associated with the post-instrument variable. For example, consider values $D_1 = D_{11} = 0.8$, $D_0 = D_{01} = 0.6$, $M_{11} = M_{10} = 0.4$, and $M_{01} = M_{00} = 0.35$. If Z_i is randomized and $Z_i \perp\!\!\!\perp M_i(z)|D_i$ also holds, as in Figure 2 a), then these observational parameters inform about the causal effect of Z_i on D_i and M_i and imply that the first-stage with respect to D_i is 0.2 while it is 0.05 with respect to M_i. Accordingly, the IV would appear reasonably strong and only weakly associated with the post-instrument variable M_i. However, if we then pick $Y_{111} = Y_{010} = 0.8$ while all other average outcomes are zero, assuming a zero average causal effect, the unadjusted estimator has bias -0.056, while the adjusted estimator has bias 0.64, more than ten times as large. This is because of a weak unconditional association of the instrument with potential outcomes, leading to little bias in the unadjusted estimator, but a strong association of M_i with unobserved confounders, leading to a strong association of Z_i with potential outcomes conditional on M_i. Note that these are biases that emerge from the reduced form. They would further increase in absolute magnitude using a Wald estimator that divided by the strength of the instrument with respect to D_i. However, their ratio would not change, as the instrument in this case is constructed to have the same strength regardless of adjustment.

The biases vanish if $Z_i \perp\!\!\!\perp Y_i(D_i = d)$ or $Z_i \perp\!\!\!\perp Y_i(D_i = d)|M_i$ (and suitable monotonicity assumptions) hold. These conditions cannot directly be inferred from Table A2, as the subgroups condition on D_i. For the construction of scenarios where such independencies hold, one would also need to choose suitable values for D_{zm} and M_z.

E 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).

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

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

$$Z_i \perp\!\!\!\perp Y_i(d, m), D_i(z), M_i(z) \text{ for all } z, d, m \tag{9}$$

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

$$P(M_i(Z_i = 1) \geq M_i(Z_i = 0)) = 1 \tag{11}$$

Under these assumptions, we have

$$
\begin{aligned}
&E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] = \\
&\qquad E[\beta_i(D_i(Z_i = 1) - D_i(Z_i = 0))] + E[\gamma_i(M_i(Z_i = 1) - M_i(Z_i = 0))] = \\
&\qquad\quad E[\beta_i|D_i(Z_i = 1) > D_i(Z_i = 0)]P(D_i(Z_i = 1) > D_i(Z_i = 0))+ \\
&\qquad\qquad E[\gamma_i|M_i(Z_i = 1) > M_i(Z_i = 0)]P(M_i(Z_i = 1) > M_i(Z_i = 0)). \quad (12)
\end{aligned}
$$

The first equality follows from model equation 2 and assumption 9. The second equality follows from the monotonicity assumptions 10 and 11.

By the exogeneity assumption 9, $P(D_i(Z_i = 1) > D_i(Z_i = 0))$ and $P(M_i(Z_i = 1) > M_i(Z_i = 0))$ are identified as $E[D_i|Z_i = 1] - E[D_i|Z_i = 0]$ and $E[M_i|Z_i = 1] - E[M_i|Z_i = 0]$. Combining this, we have that

$$
\begin{aligned}
&E[\beta_i|D_i(Z_i = 1) > D_i(Z_i = 0)] = \\
&\frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z = 0]} - \frac{E[\gamma_i|M_i(Z_i = 1) > M_i(Z_i = 0)](E[M_i|Z_i = 1] - E[M_i|Z_i = 0])}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}.
\end{aligned} \quad (13)
$$

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

$$
\frac{E[M_i|Z_i = 1] - E[M_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}
$$

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

E.2 Model 2: Binary Z_i, continuous M_i, D_i

Here, our assumptions in addition to the model in equations 2, 4, and 5 are

$$Z_i \perp\!\!\!\perp (\beta_i, \gamma_i, \alpha_i, \pi_i, \delta_i, \epsilon_{1i}, \epsilon_{2i}, \epsilon_{3i}) \tag{14}$$

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

$$P(\delta_i \geq 0) = 1 \tag{16}$$

$$cov(M_i(0), M_i(1)) \geq 0. \tag{17}$$

Under these assumptions, we have

$$\begin{aligned} E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] = \\ E[\beta_i(\alpha_i + \delta_i \pi_i)] + E[\delta_i \gamma_i]. \end{aligned} \tag{18}$$

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

$$E[D_i|Z_i = 1] - E[D_i|Z_i = 0] = E[\alpha_i + \delta_i \pi_i] \tag{19}$$

and

$$E[M_i|Z_i = 1] - E[M_i|Z_i = 0] = E[\delta_i]. \tag{20}$$

E.2.1 With measured M_i

Rewrite the bias term as

$$E[\delta_i \gamma_i] = cov(\delta_i, \gamma_i) + E[\delta_i]E[\gamma_i]. \tag{21}$$

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

$$cov(\delta_i, \gamma_i) = cor(\delta_i, \gamma_i)\sigma_{\delta_i}\sigma_{\gamma_i}. \tag{22}$$

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

$$\sqrt{var(M_i(1)) + var(M_i(0)) - 2cov(M_i(1), M_i(0))}. \tag{23}$$

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, Green and Lee (2014)) for this quantity using the marginals are not sharp, because the monotonicity does in fact improve the lower bound. (Nutz and Wang 2022) characterize this lower bound under monotonicity. 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 this assumption in section E.5. Using this assumption, an upper bound for equation 23 is

$$\sqrt{var(M_i|Z_i = 1) + var(M_i|Z_i = 0)}. \tag{24}$$

Further using $-1 \leq cor(\delta_i, \gamma_i) \leq 1$, we can bound equation 22 as

$$\begin{aligned} &-\sqrt{(var(M_i|Z_i = 1) + var(M_i|Z_i = 0))}\sigma_{\gamma_i} \\ &\le cov(\delta_i, \gamma_i) \le \\ &\sqrt{(var(M_i|Z_i = 1) + var(M_i|Z_i = 0))}\sigma_{\gamma_i}, \end{aligned} \tag{25}$$

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

$$\begin{aligned} &\frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} - \frac{1}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} \times \\ &\{(E[M_i|Z_i = 1] - E[M_i|Z_i = 0])E[\gamma_i] + \sqrt{var(M_i|Z = 1) + var(M_i|Z = 0)}\sigma_{\gamma_i}\} \\ &\le E\left[\frac{\alpha_i + \delta_i\pi_i}{E[\alpha_i + \delta_i\pi_i]}\beta_i\right] \le \\ &\frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} - \frac{1}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} \times \\ &\{(E[M_i|Z_i = 1] - E[M_i|Z_i = 0])E[\gamma_i] - \sqrt{var(M_i|Z = 1) + var(M_i|Z = 0)}\sigma_{\gamma_i}\}, \end{aligned} \tag{26}$$

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.

E.2.2 With mismeasured M_i

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:

$$M_i^\star = M_i + \eta_i \tag{27}$$

and by assuming $Z_i, M_i \perp\!\!\!\perp \eta_i$ and $E[\eta_i] = 0$. This is "classical" measurement error. We show here 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.

As before, we want to gain information on the bias term (equation 21) from the data. $E[\delta_i]$ remains identified under the measurement model in equation 27 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

$$\begin{aligned} var(M_i(z)) = var(M_i|Z_i = z) = var(M_i^\star - \eta_i|Z = z) = \\ var(M_i^\star|Z = z) + var(\eta_i|Z = z) - 2cov(M_i^\star, \eta_i|Z = z) = \\ var(M_i^\star|Z = z) + var(\eta_i) - 2cov(M_i^\star, \eta_i|Z = z). \end{aligned} \tag{28}$$

Regarding this last term, we have

$$\begin{aligned} cov(M_i^\star, \eta_i|Z = z) = cov(M_i + \eta_i, \eta_i|Z = z) = \\ cov(M_i, \eta_i|Z = z) + var(\eta_i|Z = z) = var(\eta_i). \end{aligned} \tag{29}$$

Accordingly,

$$var(M_i(z)) = var(M_i^\star|Z = z) - var(\eta_i) \leq var(M_i^\star|Z = z). \tag{30}$$

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 30 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 26 remain valid, substituting $M_i^\star$ for M_i.

E.3 Multiple instruments

We extend our model to multiple, possibly continuous instrumental variables. We here analyze the case with two IVs. A generalization to more IVs is straightforward, but notationally cumbersome. There remain two sensitivity parameters as before.

The models for D_i and M_i (suppressing X_i) become

$$D_i = \mu_D + \alpha_{1i}Z_{1i} + \alpha_{2i}Z_{2i} + \pi_i M_i + \epsilon_{2i} \tag{31}$$

$$M_i = \mu_M + \alpha_{1i}Z_{1i} + \alpha_{2i}Z_{2i} + \epsilon_{3i}. \tag{32}$$

We consider a two-stages least squares estimator $\dfrac{cov(\widehat{D}_i, Y)}{var(\widehat{D}_i)}$ based on predicting D_i using the two instruments: $\widehat{D}_i = a + a_1 Z_{1i} + a_2 Z_{2i} = E[D|Z_{1i}, Z_{2i}]$, where $a_j = E[\alpha_{ji} + \delta_{ji}\pi]$. We then have

$$cov(\widehat{D}_i, Y_i) = cov(a + a_1 Z_{1i} + a_2 Z_{2i}, \mu_y + \beta_i D_i + \gamma_i M_i + \epsilon_{1i}). \tag{33}$$

We obtain $cov(Z_{1i}, \beta_i D_i) = var(Z_{1i})E[a_1\beta] + cov(Z_{1i}, Z_{2i})E[a_2\beta]$, and analogous expressions for the other terms. Collecting terms and dividing by $var(\widehat{D}_i)$, the 2SLS estimator evaluates to

$$s_1 E[a_1\beta_i] + s_2 E[a_2\beta_i] + s_1 E[\delta_{1i}\gamma_i] + s_2[\delta_{2i}\gamma_i], \tag{34}$$

where

$$s_1 = \frac{a_1 var(Z_{1i}) + a_2 cov(Z_{1i}, Z_{2i})}{a_1^2 var(Z_{1i}) + a_2^2 var(Z_{2i}) + 2a_1 a_2 cov(Z_{1i}, Z_{2i})}, \tag{35}$$

and

$$s_2 = \frac{a_2 var(Z_{1i}) + a_1 cov(Z_{1i}, Z_{2i})}{a_1^2 var(Z_{1i}) + a_2^2 var(Z_{2i}) + 2a_1 a_2 cov(Z_{1i}, Z_{2i})} \tag{36}$$

measure the relative strength of each of the instruments. This implies weights for the causal effects β_i that are non-negative if $P(\alpha_{ji} + \delta_{ji}\pi > 0) = 1$ for all j as well as $cov(Z_{1i}, Z_{2i}) > 0$, which is the monotonicity assumption we impose. See Mogstad, Torgovitsky and Walters (2021, Proposition 6) for a related result. We also assume $P(\delta_{ji} > 0) = 1$ for all j.

The bias term now is $s_1 E[\delta_{1i}\gamma_i] + s_2[\delta_{2i}\gamma_i]$, which can be rewritten

$$s_1 cov(\delta_{1i}, \gamma_i) + s_2 cov(\delta_{2i}, \gamma_i) + (s_1 E[\delta_{1i}] + s_2 E[\delta_{2i}])E[\gamma_i]. \tag{37}$$

It is straightforward to show that a 2SLS estimator with outcome M_{1i} yields an coefficient on D_i that corresponds to $s_1 E[\delta_{1i}] + s_2[\delta_{2i}]$. This term is multiplied with the first sensitivity parameter, $E[\gamma_i]$, as before.

Rewrite $cov(\delta_{1i}, \gamma_i) = cor(\delta_{1i}, \gamma_i)\sigma_{\delta_{1i}}\sigma_{\gamma i}$. Using similar arguments as before, $\sigma_{\delta_{1i}}$ can be bounded from above using $var(M|Z_{1i} = 1, Z_{2i} = z) + var(M|Z_{1i} = 0, Z_{2i} = z)$, assuming $cov(M_i(Z_{1i} = 1, Z_{2i} = z), M_i(Z_{1i} = 0, Z_{2i} = z) \geq 0$. A similar expression obtains for $\sigma_{\delta_{1i}}$. We use a parametric model to estimate these variance (see Section E.7). $\sigma_{\gamma i}$ remains the sole second sensitivity parameter.

The remaining unknowns in the bias term are the scaling factors s_j. It is straightforward to show that these can be estimated by the coefficient on D_i from 2SLS routines with outcome Z_{ji} and treatment D_i, using both instruments.

E.4 Deriving values for the second sensitivity parameter

We assume the researcher has specified a range of values for $E[\gamma_i]$ as $[a, b]$. If one then assumes that the unit-level causal effects also are within this range, then one can derive values for σ_{γ_i} if we also assume a specific shape for the distribution supported on this range.

We suggest a four-parameter Beta distribution as the class of distributions for the unit-specific causal effects. The distribution has parameters (α, β, a, b). The general formula for its standard deviation is

$$\sqrt{\frac{\alpha\beta(b-a)^2}{(\alpha+\beta)^2(\alpha+\beta+1)}}. \tag{38}$$

For illustrative purposes, Figure A1 shows three examples of different distributions:

1. Little variance / bell curve. $\alpha = \beta = 4$. If unit-causal effect varied between 0 and 1, then this distribution would imply a standard deviation of the causal effects of approx. 0.17.
2. Medium variance / uniform distribution. $\alpha = \beta = 1$. If unit-causal effect varied between 0 and 1, then this distribution would imply a standard deviation of the causal effects of approx. 0.29.
3. High variance. $\alpha = \beta = 0.5$. If unit-causal effect varied between 0 and 1, then this distribution would imply a standard deviation of the causal effects of approx. 0.35.

Our `R` package contains a function to compute the standard deviation given the four parameters. Again, these three specific distributions are just examples to give researchers an idea of how large the sensitivity parameters might be in principle. As discussed in the main text, we suggest to explicitly investigate values of the sensitivity parameter for which main inferences do not hold. Given information from secondary literature on the effect of M_i on Y_i as well as these possible shapes of the distribution of the unit-level causal effects, researchers should then assess whether such an extreme value of the sensitivity parameter appears plausible.

E.5 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:

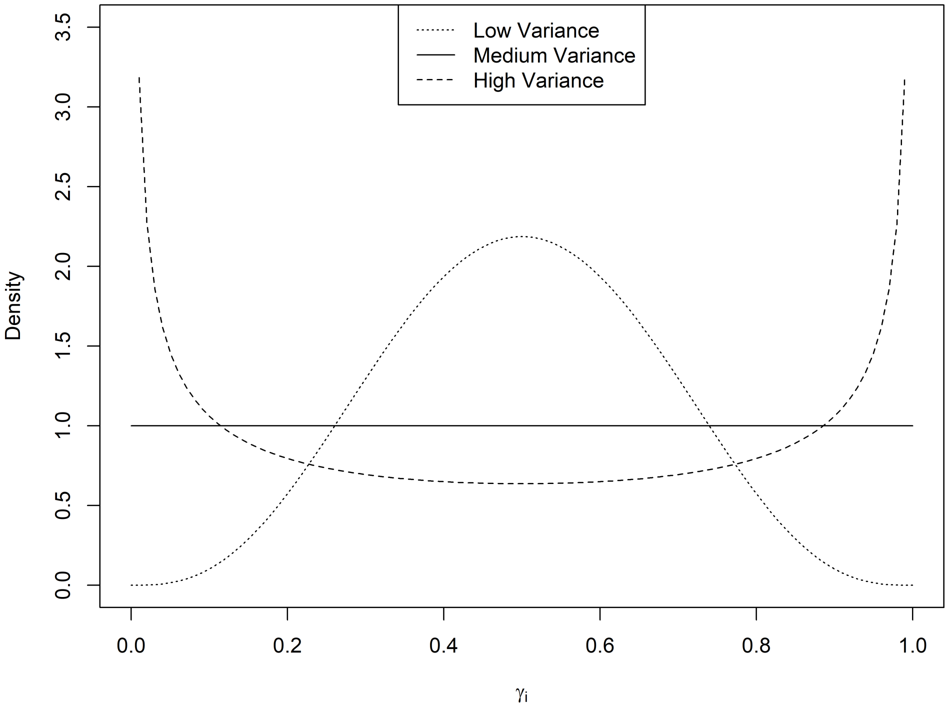

Figure A1: Three examples of possible Beta distributions of unit-level causal effects supported on $[0, 1]$. X-axis depicts unit-level causal effect, y-axis depicts its density.

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

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 A2. 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:

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

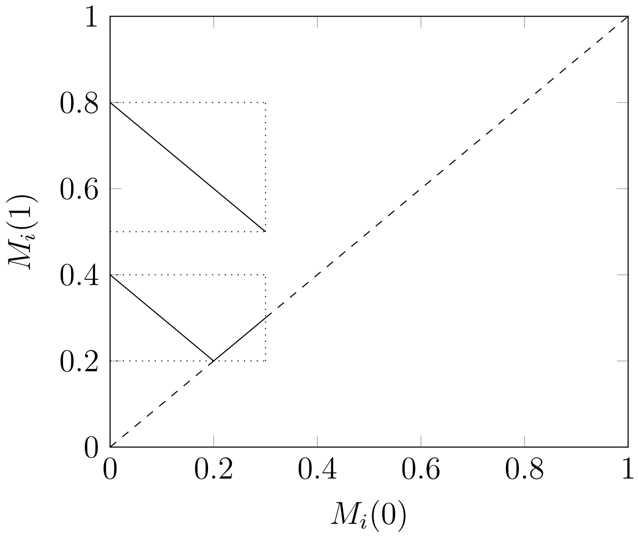

Figure A2: Understanding the relationship between the monotonicity constraint and the covariance between potential outcomes. Solid lines are the domains of two different joint distributions of $M_i(0), M_i(1)$ with negative covariance. Dotted lines indicate the domain of all possible joint distributions 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.

E.6 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

[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.

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

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.

E.7 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 26 can be pursued using two-stage least squares. For the variance terms, we pick corresponding linear conditional variance models. We first estimate auxiliary mean regressions

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

where, in the case of multiple instruments stacked into Z_i, $\zeta_1 isascalarand \zeta_2$, ζ_3 are vectors. We generate residuals $r_i = M_i - \widehat{E}[M_i | Z_i, X_i]$. We then estimate $var(M|Z = z)$ via

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

where ζ_4 is scalar and ζ_5, ζ_6 are vectors. Under this model, we have

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

In the case of one instrument, our estimate for $var(M|Z = 1) + var(M|Z = 0)$ is $2\zeta_4 + \zeta_5$.

We use the nonparametric (paired) bootstrap to estimate the sampling distribution of the resulting estimator for the bounds. 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`.

F Multiple post-instrument covariates

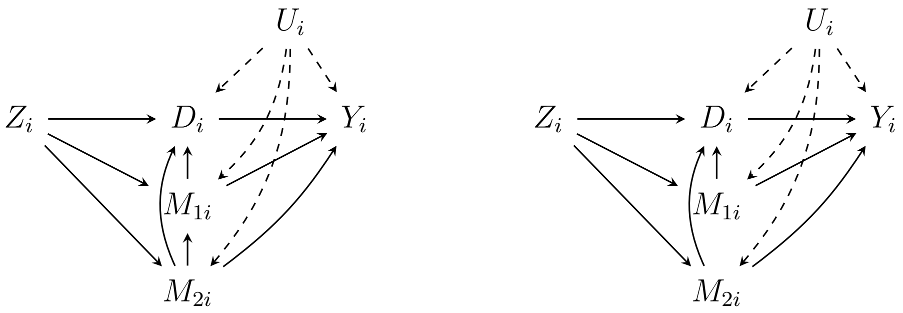

Figure A3: Graphs with $K = 2$ post-instrument covariates. Left graph: Particular causal dependence between post-instrument covariates. Right graph: Assumption of causal independence that is used in the sensitivity analysis.

In this section, we analyze the case of $K > 1$ post-instrument covariates. Figure A3 shows example DAGs where $K = 2$. In the left graph, one allows for "causal dependence" between M_{1i} and M_{i2}, specifically the latter influencing the former. We show that in such cases, a sensitivity analysis becomes practically intractable. The right graph shows an example where one assumes "causal independence" between the the post-instrument covariates. We show that under such an assumption, our proposed sensitivity analysis can easily be generalized.

F.1 Causal dependence of post-instrument covariates

Consider first a system of structural equations with varying coefficients which implies the left graph in Figure A3, leaving X_i implicit:

$$Y_i = \mu_Y + \beta_i D_i + \gamma_{1i} M_{1i} + \gamma_{2i} M_{2i} + \epsilon_{1i}. \tag{39}$$

$$D_i = \mu_D + \alpha_i Z_i + \pi_{1i} M_{1i} + \pi_{2i} M_{2i} + \epsilon_{2i} \tag{40}$$

$$M_{1i} = \mu_{M1} + \delta_{1i} Z_i + \theta_i M_{2i} + \epsilon_{3i}. \tag{41}$$

$$M_{2i} = \mu_{M2} + \delta_{2i} Z_i + \epsilon_{4i}. \tag{42}$$

This is a natural generalization of the model in the main text, with the exception that one commits to a particular causal ordering where M_{2i} influences M_{1i}. θ_i is the individual-level causal effect that corresponds to this influence.

Under a suitable generalization of assumptions 14–17, the bias term then becomes

$$E[\delta_{1i}\gamma_{1i} + \delta_{2i}\gamma_{2i} + \delta_{2i}\theta_i\gamma_{1i}].$$

Here, $E[\delta_{1i}\gamma_{1i} + \delta_{2i}\gamma_{2i}]$ corresponds to the total effect of the instrument through the post-instrument covariates if M_{1i} and M_{2i} were "causally independent", that is, were not influence to each other. As shown in the next sections, the presence of these terms leads to a generalization of sensitivity analysis with one post-instrument covariates such that each post-instrument covariate is associated with two sensitivity parameters as before (a mean effect on Y_i as well as its standard deviation).

Beyond that, under causal dependence, $E[\delta_{2i}\theta_i\gamma_{1i}]$ is an additional effect that corresponds to the path $Z_i \to M_{2i} \to M_{1i} \to Y_i$. Here, one can further write

$$\begin{aligned} E[\delta_{2i}\theta_i\gamma_{1i}] &= cov(\delta_{2i}\theta_i, \gamma_{1i}) + E[\delta_{2i}\theta_i]E[\gamma_{1i}] \\ &= cor(\delta_{2i}\theta_i, \gamma_{1i})\sigma_{\delta_2\theta}\sigma_{\gamma_1} + (cor(\delta_{2i}, \theta_i)\sigma_{\delta_2}\sigma_\theta + E[\delta_{2i}]E[\theta_i])E[\gamma_{1i}]. \end{aligned} \tag{43}$$

In this expression, there are three sensitivity parameters that are introduced by the causal dependence $M_{2i} \rightarrow M_{1i}$ and that cannot be further bounded: $E[\theta_i]$, $\sigma_{\delta_2\theta}$, and σ_θ (the other sensitivity parameters appear also under causal independence). Note that this is a simple case with just two post-instrument covariates. With additional post-instrument covariates, the number of paths connecting Z_i to Y_i and associated sensitivity parameters would further increase. For example, with three causally dependent post-instrument covariates, one may have paths such as $Z_i \rightarrow M_{3i} \rightarrow M_{2i} \rightarrow M_{1i} \rightarrow Y_i$, etc. Therefore, we suggest to focus on the case with causally independent post-instrument covariates.

F.2 Sensitivity analysis under causal independence

The right graph in Figure A3 shows a DAG with $K = 2$ causally independent post-instrument covariates. The assumption of causal independence is not testable. Due to unobserved confounding, one has open paths such as $M_{1i} \leftarrow U_i \rightarrow M_{2i}$ that create a dependence between M_{1i} and M_{2i} even in the absence of direct causal effects.

For the case of K post-instrument covariates, we generalize the structural model straightforwardly as:

$$Y_i = \mu_Y + \beta_i D_i + \sum_{k=1}^{K} \gamma_{ki} M_{ki} + \lambda_{1i}^{'} X_i + \epsilon_{1i}. \tag{44}$$

$$D_i = \mu_D + \alpha_i Z_i + \sum_{k=1}^{K} \pi_{ki} M_{ki} + \lambda_{2i}^{'} X_i + \epsilon_{2i} \tag{45}$$

$$M_{ki} = \mu_{kM} + \delta_{ki} Z_i + \lambda_{3i}^{'} X_i + \epsilon_{(2+k)i}, \text{for } k = 1, ..., K. \tag{46}$$

Leaving the conditioning on X_i implicit as before, assumptions 14–17 become

$$Z_i \perp\!\!\!\perp (\beta_i, \gamma_{ki}, \alpha_i, \pi_{ki}, \delta_{ki}, \epsilon_{ki}), \text{for } k = 1, ..., K \tag{47}$$

$$P(\alpha_i + \delta_{ki}\pi_{ki} \geq 0) = 1, \text{for } k = 1, ..., K \tag{48}$$

$$P(\delta_{ki} \geq 0) = 1, \text{for } k = 1, ..., K \tag{49}$$

$$cov(M_{ki}(0), M_{ki}(1)) \geq 0, \text{for } k = 1, ..., K. \tag{50}$$

It is important to note that the monotonicity assumptions 48 and 49 have to hold for each post-instrument covariate.

Using similar derivations as before, we obtain

$$\begin{aligned}
&\frac{E[Y_i|Z_i=1] - E[Y_i|Z_i=0]}{E[D_i|Z_i=1] - E[D_i|Z_i=0]} - \frac{1}{E[D_i|Z_i=1] - E[D_i|Z_i=0]} \times \\
&\left\{ \sum_{k=1}^{K} \{E[M_{ki}|Z_i=1] - E[M_{ki}|Z_i=0])E[\gamma_{ki}] + \sqrt{var(M_{ki}|Z=1) + var(M_{ki}|Z=0)}\sigma_{\gamma_{ki}}\} \right\} \\
&\leq E\left[\frac{\alpha_i + \sum_{k=1}^{K} \delta_{ki}\pi_{ki}}{E[\alpha_i + \sum_{k=1}^{K} \delta_{ki}\pi_{ki}]} \beta_i \right] \leq \\
&\frac{E[Y_i|Z_i=1] - E[Y_i|Z_i=0]}{E[D_i|Z_i=1] - E[D_i|Z_i=0]} - \frac{1}{E[D_i|Z_i=1] - E[D_i|Z_i=0]} \times \\
&\left\{ \sum_{k=1}^{K} \{E[M_{ki}|Z_i=1] - E[M_{ki}|Z_i=0])E[\gamma_{ki}] - \sqrt{var(M_{ki}|Z=1) + var(M_{ki}|Z=0)}\sigma_{\gamma_{ki}}\} \right\}
\end{aligned} \tag{51}$$

In sum, each post-instrument covariate is associated with two sensitivity parameters, $E[\gamma_{ki}]$ and $\sigma_{\gamma_{ki}}$. The interpretation is as before: $E[\gamma_{ki}]$ is the mean direct effect of M_{ki} on Y_i, holding D_i and all other observed variables constant. $\sigma_{\gamma_{ki}}$ is the standard deviation of this effect across individuals.

It is straightforward to show that classical measurement error, as before, does not change this result.

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 and Donald KK Lee. 2014. "Sharp bounds on the variance in randomized experiments." *The Annals of Statistics* 42(3):850–871.

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.

Mogstad, Magne, Alexander Torgovitsky and Christopher R Walters. 2021. "The causal interpretation of two-stage least squares with multiple instrumental variables." *American Economic Review* 111(11):3663–3698.

Nutz, Marcel and Ruodu Wang. 2022. "The directional optimal transport." *The Annals of Applied Probability* 32(2):1400–1420.

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

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.

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.