

Handle with Care: A Sociologist’s Guide to Causal Inference with Instrumental Variables *

Chris Felton[†] Brandon M. Stewart[‡]

September 5, 2022

Abstract

Instrumental variables (IV) analysis is a powerful, but fragile, tool for drawing causal inferences from observational data. Sociologists have increasingly turned to this strategy in settings where unmeasured confounding between the treatment and outcome is likely. This paper provides an introduction to the assumptions required for IV and consequences of their violations for applications in sociology. We review three methodological problems IV faces: identification bias (asymptotic bias from assumption violations), estimation bias (finite-sample bias that persists even when assumptions hold), and type-M error (exaggeration of effects given statistical significance). In each case, we emphasize how weak instruments exacerbate these problems and make results sensitive to minor violations of assumptions. Our discussion is informed by a new survey of IV papers published in top sociology journals showing that assumptions often go unstated and robust uncertainty measures are rarely used. We provide a practical checklist to show how IV, despite its fragility, can still be useful when handled with care.

1 Introduction

Sociology is full of difficult but important questions. In many cases—for instance, studying the effect of incarceration on future employment—we seek causal evidence where

*For helpful discussions and feedback relevant to this project, we thank Simone Zhang, Yiqing Xu, Patrick Sharkey, Hannah Postel, Ian Lundberg, Angela Li, Katie Donnelly Moran, Dalton Conley, and members of the Stewart Lab. Research reported in this publication was supported by The Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number P2CHD047879.

[†]Ph.D. Candidate, Department of Sociology and Office of Population Research, Princeton University, cfelton@princeton.edu.

[‡]Associate Professor, Department of Sociology and Office of Population Research, Princeton University, brandonstewart.org, bms4@princeton.edu. 149 Wallace Hall, Princeton University, Princeton, NJ 08540.

randomized controlled trials would be impractical or unethical. The most common strategy for identifying causal effects with observational data is *selection on observables*. With selection on observables, we condition on a set of observed confounders (e.g., through regression, weighting, or matching) that help to block non-causal associations between the treatment and the outcome. Under strong assumptions, the remaining correlation is causation.

Selection on observables demands we block *all* confounding between the treatment and outcome. Part of what makes many sociological questions difficult is that we can never measure enough variables to make this no-unmeasured-confounding assumption completely plausible. Following economics, sociologists have increasingly turned to instrumental variables (IV) as an alternative—or complementary—strategy for estimating causal effects. IV exploits the presence of an *instrument*, a random source of variation that affects the treatment but only affects the outcome through its affect on the treatment.¹ IV isolates the variation in treatment caused by the instrument and uses only this variation to estimate the causal effect. The promise of IV is that it allows us to credibly estimate causal effects even when unmeasured confounding plagues the association between the treatment and outcome.

IV is powerful. But it is also more fragile than many researchers realize. IV rests on strong assumptions, and it can be highly sensitive to violations of these assumptions. In this paper, we walk through the assumptions IV requires and explain why the IV estimator can be brittle under even modest violations of the assumptions. To clarify what these assumptions require, we discuss violations in the context of real empirical examples and illustrate these violations graphically when possible. Our hope is that a better understanding of IV’s assumptions will help researchers select more plausible instruments, while a better understanding of its fragility will enable researchers and readers to make appropriate comparisons between IV and selection-on-observables strategies.

IV can still be a useful tool if we handle it with care. To this end, we provide concrete guidance—in the form of a checklist—on how to improve the use of IVs. Of particular importance is *bias analysis*. In practice, IV assumptions will often be violated. Bias analysis allows us to assess how severe the violations would have to be to change the conclusions of our study. In other words, IV is fragile, but bias analysis helps us deal with fragility in a constructive way. In addition to bias analysis, we review robust diagnostics and uncertainty measures for IV that are seldom used in sociology.

Our discussion of IV assumptions and our guidelines for using IV are grounded in a survey of 34 IV papers published in the *American Journal of Sociology* (*AJS*) and the *American Sociology Review* (*ASR*) between 2004 and 2022. Our survey revealed that authors rarely stated or defended the assumptions required for IV. For instance, out of 34 papers, only one stated the “monotonicity” assumption, an important condition we review in Section 3.4. Furthermore, researchers seldom reported key diagnostics that are considered standard in other fields. For instance, only 18 of 34 papers reported a

¹More precisely, an instrument is an *unconfounded* source of variation (Sävje, 2021).

“first-stage F -statistic,” a diagnostic we review in Section 4.2, and no study reported a formal bias analysis, which we describe in Section 4.1. Finally, like reviews in other fields, we found that IV estimates were typically larger in magnitude than selection-on-observables estimates (Jiang, 2017; Lal et al., 2021). This finding is concerning—we often resort to IV because we suspect selection-on-observables estimates are biased *upward* in magnitude—and we consider two potential explanations for this trend in Section 4.3. By collecting guidance from different literatures and organizing it in one place, our checklist can provide researchers with an easy reference to improve the use of IV in sociology.

1.1 Related Work

IV is a popular approach in the social sciences, and there are consequently many resources on the approach. In sociology, Bollen (2012) reviews IV methods primarily in the context of linear structural equation models. In contrast, we discuss IV in the potential outcomes framework, avoiding constant effects assumptions and allowing us to describe identification assumptions in explicitly causal terms more akin to the coverage of Morgan and Winship (2015). Compared with Morgan and Winship (2015), our review places more emphasis on how IV estimates are used and interpreted in practice, including highlights of recently developed diagnostics and inferential tools. We pay special attention to commonly-used instruments and clarify how they may violate critical assumptions.

Reader’s guides of this sort have appeared in numerous fields. Sovey and Green (2011) offer a useful reader’s guide to IV use in political science. In economics, Angrist and Pischke (2008) review identification assumptions for IV and explain the interpretation of IV estimates across a range of settings. Baiocchi, Cheng and Small (2014) give a thorough introduction to using IV in medical research, focusing on instruments that are widely used in medicine and epidemiology. With the exception of Rossi (2014)—who critically surveys IV use in marketing—our review provides a more skeptical perspective on IV than other reviews. Furthermore, we clarify the conditions under which instruments can be used to address “simultaneity bias,” an oft-cited motivation for IV that receives little coverage in IV reviews.

We also discuss weak instruments differently from most reviews. Instrument weakness exacerbates three separate methodological problems, each requiring its own diagnostic tests. But review articles typically mention only two problems and provide diagnostics for only one. Confusingly, methodologists sometimes describe these latter diagnostics as “weak-instrument tests” even though they address only one of the problems aggravated by weakness. To improve both understanding and practice, we provide new names for these distinct problems and suggest diagnostics for all three.

More recently, Young (2019) and Lal et al. (2021) use mass replications of published IV papers to study statistical inference in economics and political science, respectively. Young (2019) demonstrates that published IV results in economics are extremely sensitive to outliers and shows that even robust confidence intervals typically fail to achieve

nominal coverage, particularly when clustered or panel data is used. Lal et al. (2021) find that similar inferential problems plague the political science literature, and further document that IV estimates are often much larger in magnitude than selection-on-observables estimates. We take inspiration from Lal et al. (2021) in comparing the magnitude of IV and selection-on-observables estimates in sociology and providing a checklist for best practices. While these surveys also emphasize the fragility of IV, we primarily focus on research design and identification assumptions rather than statistical inference—concerns that would be hard to surface in mass replications. Additionally, we highlight the problem of *type-M* errors, an issue that has received less attention in the IV setting.

1.2 Structure of the Paper

In Section 2, we introduce the basic logic of instrumental variables and provide several running examples from the sociological literature. Section 3 reviews the assumptions IV methods require, paying close attention to certain subtleties that often go unstated, and review how to estimate treatment effects with IV. Section 4 highlights three methodological challenges IV faces and provide guidelines on how to address each one. In Section 5, we provide a thorough checklist for practitioners using IV as well as readers of papers that use IV.

2 What is an Instrumental Variable, and Why Would We Use One?

If a recently released parolee were to move to a new neighborhood—rather than return home—would they be less likely to reoffend? This is the question asked by Kirk (2009). In observational data, the association between moving and recidivism is almost certainly confounded—for instance, by the resources that allow such a move to occur—and we may lack the information necessary to produce a credible selection-on-observables estimate. If we could randomly assign some parolees to move and others to return home, we could produce a set of data free of confounding, but such an experiment would likely be infeasible.

Kirk (2009) approaches this problem by using the timing of release—before or after Hurricane Katrina—as an instrumental variable (IV). The logic is that many parolees in Louisiana were unable to return home following Katrina. Assuming the timing of the release is random, it is as though those who were released after Katrina were “assigned” to a condition where they were unable to return home in a hypothetical randomized experiment. Importantly, randomness is not enough—we also need the timing of release to affect recidivism *only through* the process of making it difficult to return home. This way, when we examine the association between those induced into being unable to return home and the outcome, we know it is due to being unable to return home (the treatment) and not some other consequence of the timing of the release. If the

necessary assumptions hold—which we detail more explicitly below—the IV strategy does an amazing thing: it provides us the way to identify causal effects even if there are unmeasured factors that cause parolees to both change neighborhoods and reoffend. In practice, though, our estimates can be sensitive to seemingly minor violations of these very strong assumptions. We emphasize that the causal question Kirk (2009) considers is both important and challenging. Thus, even if the identification strategy is imperfect, the analyst may conclude it is the best path forward.

Table 1 summarizes the research design and assumptions employed by Kirk (2009) along with three other IV studies in sociology that we use throughout this paper. We will also make reference throughout to our survey of all 34 papers using IV in *American Journal of Sociology (AJS)* and *American Sociological Review (ASR)* published between 2004 and 2022. Complete details on this survey are in Appendix A.

3 Identification and Estimation of Treatment Effects with IV

We begin by reviewing the assumptions under which IV can provide consistent treatment effect estimates. Our summary of identification assumptions serves three purposes. First, we aim to clarify what exactly the assumptions require so that researchers can select more plausible instruments. Second, we demonstrate how to explain these conditions in a non-technical fashion so that researchers can better convey to readers what assumptions their study relies on. Finally, we illustrate why it is so difficult to find instruments that meet all of our required assumptions. In the Section 4 that follows, we show how to proceed with IV analysis when our assumptions fail to hold exactly.

Before turning to identification assumptions, we briefly review the notation we use throughout the paper (although we describe each assumption in non-mathematical terms as well). Let Z_i represent the instrument—for Kirk (2009) whether parolee i was released before ($Z_i = 0$) or after ($Z_i = 1$) Hurricane Katrina. Let D_i represent the treatment—whether parolee i returned to their home parish ($D_i = 0$) or a different parish ($D_i = 1$). Let Y_i represent the outcome—whether parolee i was re-arrested within one year following release. Kirk (2009) also includes a set of pre-instrument controls, which we call \mathbf{X}_i .

We also employ potential outcomes notation (Rubin, 1974).² Let $Y_i(D_i = 1)$ represent the outcome we would observe for parolee i if—possibly contrary to fact—he had been assigned the treatment. If parolee i had been assigned to the control condition, $Y_i(D_i = 1)$ would be an unobserved, counterfactual outcome. The treatment effect for parolee i is the difference between his potential outcomes under treatment and control: $Y_i(D_i = 1) - Y_i(D_i = 0)$. The fundamental problem of causal inference is that we can only observe one of these potential outcomes (Holland, 1986).

²For accessible introductions to potential outcomes notation, see Morgan and Winship (2015) or Hernán and Robins (2021).

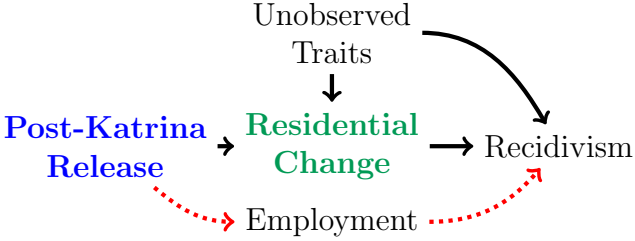
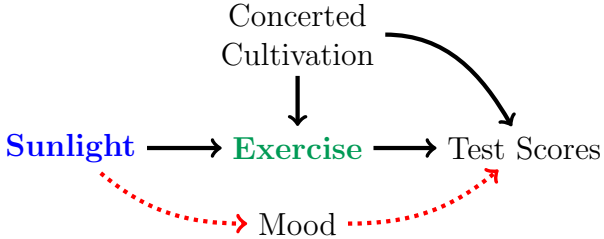
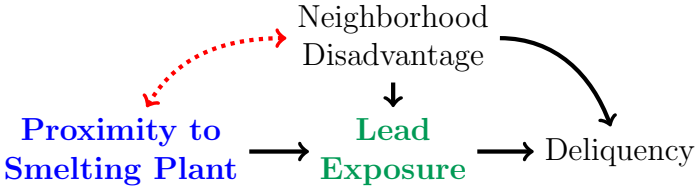
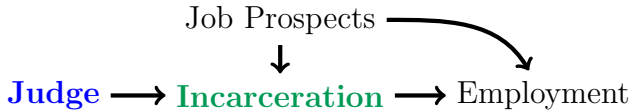
Study	Causal Structure	Identification Assumptions
Kirk (2009)		<p>(1) Post-Katrina release induces moves away from a parolee’s home neighborhood. (2) Post-Katrina release affects recidivism <i>only through</i> residential change and not, e.g., employment or police resources. (3) The timing of release shares no common causes with recidivism. (4) Post-Katrina release <i>never discourages</i> residential change.</p>
Laidley and Conley (2018)		<p>(1) More sunlight causes kids to exercise more. (2) Sunlight affects test scores <i>only through</i> exercise and not, e.g., mood. (3) When person fixed effects are included, within-person sunlight variation shares no common causes with test scores. (4) Sunnier weather <i>never discourages</i> exercise.</p>
Sampson and Winter (2018)		<p>(1) Proximity to a smelting plant increases lead exposure. (2) Proximity affects delinquency <i>only through</i> lead exposure. (3) Proximity shares no common causes—such as neighborhood disadvantage—with delinquency. (4) Moving closer to a smelting plant <i>never</i> causes someone to experience less lead exposure.</p>
Harding et al. (2018)		<p>(1) Judge assignment affects the probability of being incarcerated. (2) Judge assignment affects employment <i>only through</i> incarceration. (3) Judge assignment shares no common causes with future employment. (4) If assigned to a more lenient judge, a defendant will <i>never</i> receive a harsher sentence.</p>

Table 1: Three examples of IV designs from the sociological literature illustrated with Directed Acyclic Graphs (DAGs), where arrows indicate causal effects (Pearl, 1995). The **instruments** are bolded and colored blue, the **treatments** bolded and green. Red, dashed arrows indicate potential violations of identification assumptions. Weather, distance, and judge assignment are commonly used instruments in the social sciences.

Identification Assumption	% of Papers that State Assumption
Relevance	82.35
Unconfoundedness	20.59
Exclusion Restriction	61.76
Monotonicity	2.9
SUTVA	0
Positivity	0

Table 2: The proportion of papers that state each identification assumption in terms of causal effects of the instrument or unmeasured confounders. For instance, we can describe the unconfoundedness assumption as the assumption that the instrument shares no common causes with the treatment or outcome. In contrast, papers that merely state that the instrument must be uncorrelated with an error term are describing the assumption in terms of associations in a statistical model rather than causal effects. We believe stating assumptions in causal terms helps both author and audience reason about them more clearly.

Let $D_i(Z_i = 1)$ represent the potential treatment parolee i would have received had he been randomly assigned to be released post-Katrina rather than pre-Katrina—again, possibly contrary to fact. Also, let $Y_i(D_i = 1, Z_i = 1)$ represent the potential outcome we would observe for person i had he received both the instrument and the treatment.³

In this section, we outline each of the six identification assumptions used in IV analysis and provide examples of their violations. Table 2 outlines the percentage of papers in our survey that state each assumption in causal terms.

3.1 Relevance

In Kirk (2009), the relevance assumption requires that being released after Hurricane Katrina has a causal effect on whether a parolee returned to his home parish or not. Kirk shows that roughly 75% of parolees released pre-Katrina returned to their home parishes, compared with only 50% of parolees released post-Katrina. Such a large and abrupt difference in return rates likely indicates a causal effect of the instrument on the treatment. We state this assumption formally below.

Assumption 1: Relevance

$$E \left[\underbrace{D_i(Z_i = 1)}_{\substack{\text{Treatment status} \\ \text{for unit } i \text{ if she} \\ \text{were assigned} \\ \text{the instrument}}} - \underbrace{D_i(Z_i = 0)}_{\substack{\text{Treatment status} \\ \text{for unit } i \text{ if she} \\ \text{were} \\ \text{not assigned} \\ \text{the instrument}}} \right] \neq 0$$

³We make frequent use of i subscripts to improve clarity, although for many assumptions they are not strictly necessary. We emphasize that X_i denotes the i th draw of X from a population (rather than enumerating particular units within that population). $E[X_i]$ denotes the expectation of the random variable X_i , which can be thought of as the average across infinitely many samples of the random variable X from that population.

More generally, this assumption states that the *instrument* (post-Katrina release) has a non-zero average causal effect on *treatment* uptake (moving neighborhoods). This is an assumption about *causation*—not *association*. The instrument causes the change in the treatment. An instrument that is caused by the treatment, in contrast, cannot be used to identify treatment effects.⁴

It is also possible to estimate treatment effects using a *proxy* instrument—something that has no causal effect on the treatment but shares an unmeasured common cause with the treatment. Even in the proxy case, an instrument that causes the treatment must exist (although we can’t observe it), and key assumptions apply to the true causal instrument. Because proxy-instrument designs require different and more nuanced assumptions, we discuss them separately in Appendix B. Note also that a *relevant* instrument can still be a *weak* instrument, a topic we return to in Section 4.

3.2 Unconfounded Instrument

Consider the work of Sampson and Winter (2018), who are interested in the effect of lead exposure on delinquency. As an instrument, they use a person’s distance from a smelting plant. Smelting plants contaminate soil, exposing nearby residents to lead and making distance a relevant instrument for lead exposure. The unconfoundedness assumption has two parts.⁵ First, the *instrument* (distance to a smelting plant) can share no unmeasured common causes with the *outcome* (delinquency).⁶ Second, the *instrument* (distance to a smelting plant) can share no unmeasured common causes with the *treatment* (lead exposure). We can state the assumptions in potential outcomes notation as follows:

Assumption 2a: Unconfounded Instrument

$$\underbrace{Y_i(d, z)}_{\substack{\text{The distribution} \\ \text{of potential outcomes} \\ \text{under } \text{treatment level } d \\ \text{and } \text{instrument level } z}} \perp\!\!\!\perp \underbrace{Z_i}_{\substack{\text{is independent} \\ \text{of the distribution} \\ \text{of the } \text{instrument}}} \quad \text{for all } d, z \quad \underbrace{\text{across all values}}_{\substack{\text{the } \text{treatment} \\ \text{and } \text{instrument} \\ \text{can take}}}$$

⁴This scenario is ruled out by the unconfoundedness assumption (Section 3.2), so some authors state the more general assumption that the instrument must be associated with the treatment, which covers both causal instruments and proxy instruments.

⁵This assumption goes by many names, including “ignorability” and “exogeneity.” Exogeneity in particular sometimes refers to both the combination of unconfoundedness and the exclusion restriction, which we discuss in Section 3.3 (e.g., Greene (2008), 316 and Wooldridge (2010), 89). While these two assumptions can be collapsed mathematically, we find it is easier to reason about them separately.

⁶More precisely, unconfoundedness requires all *back-door paths* are blocked (Pearl, 2009). If plausible common causes are measured, we can block these paths by conditioning on them (VanderWeele and Shpitser, 2011).

Assumption 2b: Conditionally Unconfounded Instrument

$$\overbrace{Y_i(d, z) \perp\!\!\!\perp Z_i}^{\text{Unconfoundedness}} \underbrace{\mid \mathbf{X}_i}_{\text{Within strata of observed confounders}} \text{ for all } d, z$$

The reason we need to resort to IV in the first place is that we believe the *treatment* is counfounded. As such, the unconfoundedness of the instrument plays a central role in the analysis. Yet, of 34 papers in our survey of IV papers, only seven stated this central assumption in causal terms. Most authors instead describe a broader assumption about the correlation between the instrument and an error term, a common practice in the classical structural equation models setting (see e.g. Bollen, 2012). We avoid this description because it combines both an unconfoundedness assumption and the *exclusion restriction*, a condition we describe in Section 3.3. While these two assumptions can be collapsed mathematically, we find it is easier to reason about them separately.

How plausible is the unconfoundedness assumption in the Sampson and Winter (2018) setting? To answer this question, we refer to a figure from their online supplement reproduced here as Figure 1. The map shows that location of smelting plants is strongly associated with neighborhood poverty composition. Just as neighborhood disadvantage may confound the relationship between the *treatment* and *outcome*, it may confound the relationship between the *instrument* and *outcome*. We depict this potential violation graphically in Figure 2 using red dashed arrows.

Sampson and Winter (2018) acknowledge that distance from smelting plants is likely confounded with delinquency, but they argue that the instrument should be unconfounded conditional on measured confounders such as the percentage of a neighborhood living in poverty. The motivation for IV in this setting relies on the idea that these measured covariates are sufficient for blocking instrument–treatment and instrument–outcome confounding while also being *insufficient* for blocking treatment–outcome confounding. If residents select into neighborhoods on the basis of unmeasured characteristics that are associated with the instrument (e.g., school quality), and these characteristics also affect the outcome, the assumption will be violated.

Distance-based instruments are common: Card (1995) uses distance from a college as an instrument for attending college, and McClellan, McNeil and Newhouse (1994) use differential distance to alternative hospitals as an instrument for intensive heart attack treatment. Distance-based instruments raise concerns about the unconfoundedness assumption, since where someone lives tends to be associated with many other individual- and neighborhood-level characteristics. While instruments may often seem *less* confounded than their corresponding treatments, in Section 4 we explain why IV estimates can still be more biased than selection-on-observables estimates in such settings. In Section 4.1 and Appendix C, we discuss tools for addressing potential violations of unconfoundedness.

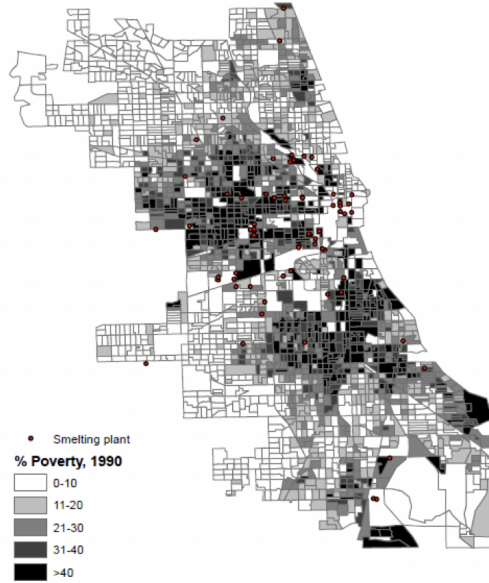


Figure 1: Map depicting the locations of smelting plants relative to neighborhood poverty composition. The figure is taken from Sampson and Winter (2018), Appendix D. The locations are strongly associated with neighborhood poverty composition, raising concerns about the unconfoundedness assumption.

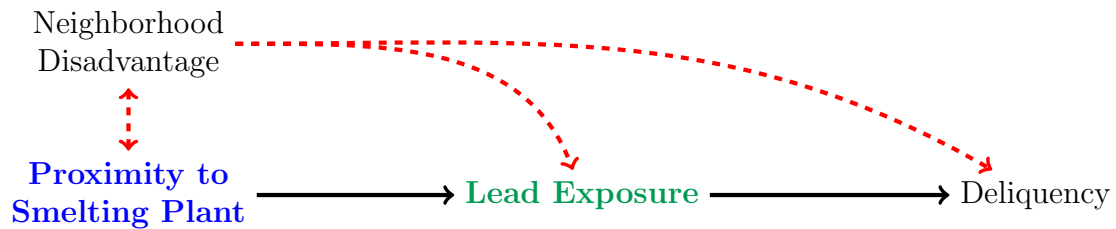


Figure 2: DAG depicting the causal structure in Sampson and Winter (2018), with violations of unconfoundedness drawn using red, dashed arrows. Neighborhood disadvantage generates both instrument–outcome confounding and instrument–treatment confounding through neighborhood disadvantage. The bi-directed arrow between *Neighborhood disadvantage* and *Distance from smelting plant* indicates they are themselves related through some unmeasured confounder.

3.3 The Exclusion Restriction

The exclusion restriction requires that the instrument has no effect on the outcome except through the treatment. For Kirk, the exclusion restriction means that being released after Hurricane Katrina (the **instrument**) has no effect on recidivism (the outcome) *except through residential change* (the **treatment**). This assumption can be written in potential outcomes notation as follows.

Assumption 3: Exclusion Restriction

$$\underbrace{Y_i(d, z) = Y_i(d, z') = Y_i(d)}_{\substack{\text{The outcome for unit } i \text{ we would observe} \\ \text{if she were assigned } \text{treatment level } d \text{ is} \\ \text{the same regardless of the level of the} \\ \text{instrument } (z, z') \text{ assigned}}} \quad \overbrace{\text{for all } z, z', d, i}^{\substack{\text{across all units and all} \\ \text{levels of the } \text{instrument} \\ \text{and } \text{treatment}}}$$

Of the 34 papers we reviewed, 21 describe what the exclusion restriction entails in causal terms. While the exclusion restriction is the most widely stated assumption in papers exploiting IVs, we also believe it may be the most misunderstood by casual readers. We emphasize three subtleties. First, while the exclusion restriction is often described as the instrument having “no direct effect” on the outcome, it is clearer to say that there are no effects *unmediated by the treatment*. The reason is that effects mediated by unmeasured variables other than the treatment still violate the assumption. Second, while we can block exclusion restriction violations by conditioning on post-instrument covariates, doing so may induce violations of unconfoundedness (Glynn, Rueda and Schuessler, 2021). Third, coarsely measured treatments can generate exclusion restriction violations (Marshall, 2016). We expand on each of these points in the three subsections below.

3.3.1 What “No Direct Effect” Means in the Context of the Exclusion Restriction

In the case considered by Kirk (2009), suppose that Hurricane Katrina prevented a hypothetical parolee from getting a job. Suppose further that having a job would prevent him from committing a crime, which in turn reduces his probability of being arrested. This causal chain—from the hurricane (instrument), to employment, to crime, to re-arrest (outcome)—would violate the exclusion restriction. Another potential exclusion restriction might occur through strain on police resources. We depict both in Figure 3.

One case where we should be particularly cautious of exclusion restriction violations through indirect causal chains is when the same instrument is used for multiple different treatments—suggesting that the instrument may affect many different things. Mellon (2021) found that weather has been used as an instrument in at least 111 studies across the social sciences. The fact that weather appears to affect so many different treatments raises concerns about its validity in any given application. Laidley and

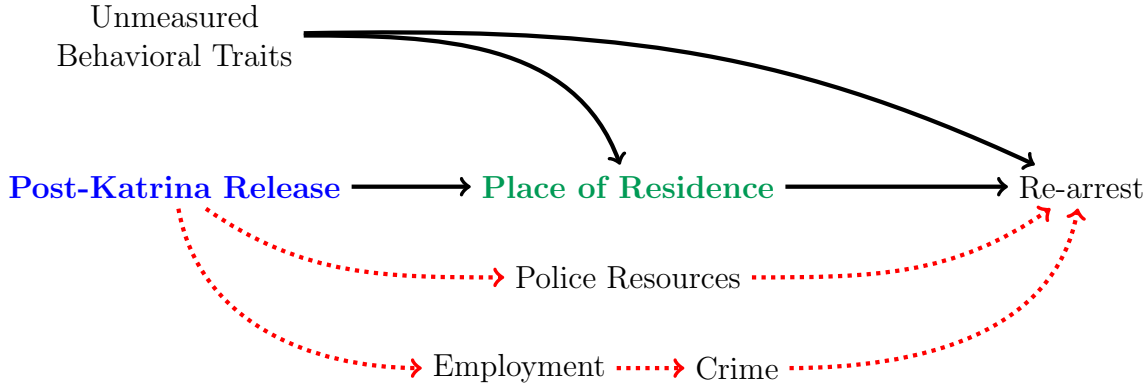


Figure 3: DAG depicting the causal structure in Kirk (2009), with violations of the exclusion restriction drawn using red, dashed arrows. The causal pathways *Post-Katrina release* \rightarrow *Police resources* \rightarrow *Re-arrest* and *Post-Katrina release* \rightarrow *Employment* \rightarrow *Crime* \rightarrow *Re-arrest* each violate the exclusion restriction because they do not operate through the treatment. Note that we can also think of the former causal pathway as differential measurement error rather than a violation of the exclusion restriction, where being re-arrested is a mismeasure of recidivism, and the error in measurement is affected by the instrument. To simplify the DAG, we opted to use re-arrest (whether the parolee was arrested) rather than recidivism (whether the parolee committed a crime) as the outcome.

Conley (2018), for instance, study the effects of outdoor leisure on math test scores using sunlight as an instrument for outdoor activity. As the authors themselves highlight, one potential violation of the exclusion restriction occurs through mood. If sunlight improves mood through pathways other than outdoor activity, and mood itself affects math performance, the exclusion restriction could be violated.⁷ Indeed, Mellon (2021) locates a number of studies showing that weather affects mood and that mood in turn influences a wide range of other outcomes.

Historical instruments—those that occurred long before the treatment—present another complicated case for exclusion restriction assumptions. Consider Rothwell and Massey’s (2010) study using population density in 1910 as an instrument for density zoning in 2000 to assess its effect on segregation in the year 2000. While it is hard to imagine a “completely unmediated” effect of population density in 1910 on segregation in 2000, it is easy to imagine effects *unmediated by the treatment* (density zoning in 2000). Population density affects a lot of things about cities—economic growth, housing prices, wages—and it is difficult to believe that none of these things have any downstream effects on segregation (Ahlfeldt and Pietrostefani, 2019).

⁷Laidley and Conley (2018) argue that effects of weather on mood are likely trivial in magnitude. However, as we emphasize in Section 4.1, even small violations can yield counterintuitively large bias in IV.

3.3.2 Why Blocking Exclusion Restriction Violations Can Induce Unconfoundedness Violations

To block potential violations of the exclusion restriction, Kirk (2009) conditions on the parish-level unemployment rate and other parish-level post-instrument variables. The idea is that while there are paths between hurricane timing and re-arrest, we can block them by identifying those mediators and conditioning on them. Unfortunately, if there are any unmeasured common causes between unemployment (the mediator) and re-arrest, conditioning on the unemployment rate will violate unconfoundedness, the first assumption we discussed.⁸

When we condition on the unemployment rate, we compare *pre*-Katrina parishes suffering from high unemployment with *post*-Katrina parishes suffering from high unemployment in order to estimate the effects of the instrument. But this is not an “apples-to-apples” comparison. Pre-Katrina, high unemployment can be a mark of long-term disadvantage. But in the immediate aftermath of Katrina, high unemployment is largely a function of living near the coast. When we compare pre-Katrina parishes and post-Katrina parishes with the same levels of unemployment, we end up comparing more disadvantaged parishes with less disadvantaged parishes. We have inadvertently induced an association between being released post-Katrina and living in a more advantaged parish. We depict this violation graphically in Figure 4.

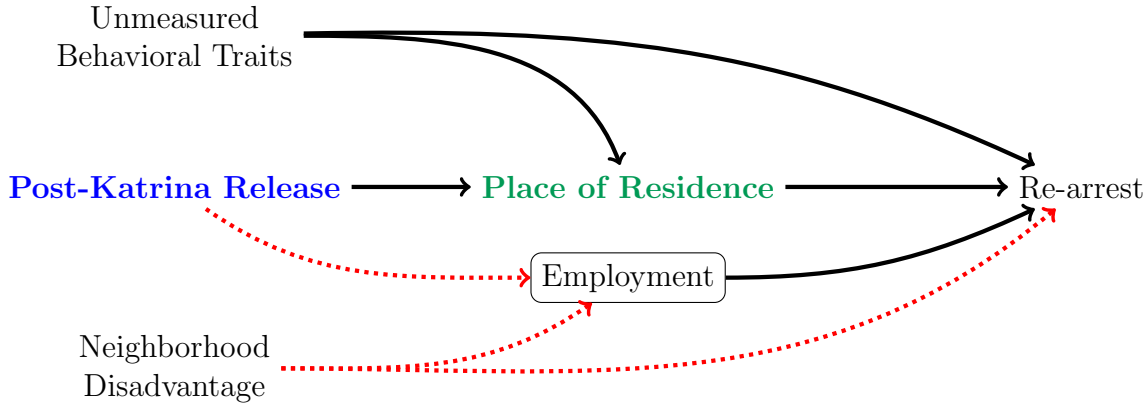


Figure 4: DAG illustrating how conditioning on a post-instrument collider induces an unconfoundedness violation. Conditioning on *Employment* opens up the path *Post-Katrina release* \rightarrow *Employment* \leftarrow *Neighborhood disadvantage* \rightarrow *Re-arrest*, illustrated with red-dashed arrows. The rectangle drawn around *Employment* indicates it has been conditioned on.

More generally, we can say that conditioning on post-instrument variables induces unconfoundedness violations in IV settings. In the language of DAGs, post-instrument variables will often be “colliders” between the instrument and an unmeasured confounder, and conditioning on this collider will induce an association between the instru-

⁸See Elwert and Winship (2014) and Knox, Lowe and Mummolo (2020) for explanations of why conditioning on post-treatment variables induces bias in selection-on-observables settings.

ment and the unmeasured confounder (Elwert and Winship, 2014). Following Glynn, Rueda and Schuessler (2021), we call this *post-instrument bias*. Post-instrument bias parallels post-treatment bias in selection-on-observables settings.

3.3.3 Why Coarse or Coarsened Treatments Induce Exclusion Restriction Violations

Kirk (2009) measures residential change using a parolee’s parish of residence. Parishes measure neighborhoods coarsely. In 2010, there were 204,447 census blocks and 3,471 block groups in Louisiana, but only 64 parishes, so within any parish there is substantial variability in neighborhood conditions. Suppose that Hurricane Katrina induced a parolee to move to a different neighborhood in the same parish, and that moving to this different neighborhood prevented him from reoffending. Because his move will not be captured in our measure of residential change (i.e., he returned to the same parish), the exclusion restriction will be violated: the instrument has an effect on the outcome that is not captured by treatment *as we have measured it*.⁹ More generally, coarsely measured treatments will often bias IV estimates upward in magnitude (Marshall, 2016).¹⁰

3.4 Monotonicity

A common instrument for incarceration is judge leniency. To give a toy example, suppose a courthouse has two judges—Judge Alice and Judge Bob—and that defendants are randomly assigned to one for sentencing. If Bob is more likely to sentence someone to prison than Alice, being assigned to Bob could serve as an instrument for incarceration. Being assigned to Bob increases your probability of incarceration, and it is unconfounded since judges are randomly assigned. We can use variation in incarceration caused by judge leniency to identify the effects of incarceration on a range of outcomes, such as future employment or earnings.

To illustrate the monotonicity assumption, we define four possible “compliance types” from Angrist, Imbens and Rubin (1996) as shown in Table 3. Consider a hypothetical defendant Carol. Carol is a *never-taker* if, regardless of the judge she is assigned to, she is never incarcerated. Carol is an *always-taker* if she is incarcerated regardless of her judge assignment. Carol is a *complier* if she would be incarcerated if assigned to the harsher judge (Bob) but would not be incarcerated if assigned to the more lenient judge (Alice). Finally, Carol is a *defier* if she would be incarcerated if assigned to the more lenient judge (Alice) but would not be incarcerated if assigned

⁹While researchers may disagree on the correct level of geographic aggregation with which to measure “neighborhood” in any given analysis (e.g., Hipp (2007)), our point is that if post-Katrina release causes within-parish moves that in turn affect the outcome, the exclusion restriction will be violated.

¹⁰See Marshall (2016) for a formal discussion of conditions required for the bias to be upward in magnitude. To build intuition for why coarsely measured treatments often bias effect estimates upward in magnitude, consider how this measurement error affects the first-stage and intent-to-treat (ITT) estimates (Section 3.7). The ITT estimate will remain unchanged, but the first-stage regression will underestimate the proportion of compliers, inflating the treatment effect estimate.

Compliance Type	Judge Alice (lenient)	Judge Bob (harsh)
<i>Always-Taker</i>	Incarcerated	Incarcerated
<i>Never-Taker</i>	Not incarcerated	Not incarcerated
<i>Complier</i>	Not incarcerated	Incarcerated
<i>Defier</i>	Incarcerated	Not incarcerated

Table 3: Compliance types in the toy judge example. Monotonicity holds that there can be no defiers in our sample, although we will still recover the LATE under violations of monotonicity so long as defiers have the *exact* same treatment effect as compliers. See Appendix E for compliance type definitions when the instrument or treatment is continuous.

to the harsher judge (Bob). The monotonicity assumption requires that there are *no defiers* in our sample. Because this is an assumption about individual-level effects rather than average effects, it is fundamentally untestable. Of the 34 sociology papers we reviewed, only one stated the monotonicity assumption. We define the assumption formally below.¹¹

Assumption 4: Monotonicity

$$\underbrace{D_i(Z_i = 1)}_{\substack{\text{The treatment status} \\ \text{we would observe for} \\ \text{unit } i \text{ if she} \\ \text{were assigned} \\ \text{the instrument}}} \geq \underbrace{D_i(Z_i = 0)}_{\substack{\text{The treatment status} \\ \text{we would observe for} \\ \text{unit } i \text{ if she were} \\ \text{not assigned} \\ \text{the instrument}}} \quad \text{for every unit} \\ \text{in the population} \quad \text{for all } i$$

How might monotonicity be violated? Suppose that Bob orders harsh sentences for violent offenses but lenient sentences for drug offenses. In contrast, Alice is harsh with drug offenses but lenient with violent offenses. If the courthouse sees more violent offenders than drug offenders, Bob will appear to be more lenient overall. But if our defendant, Carol, is a drug offender, she might be a defier: she has a higher probability of being incarcerated if assigned to Alice rather than Bob. The monotonicity assumption rules out defiers like Carol. If present in our sample, monotonicity will be violated. More generally, the monotonicity assumption holds that an instrument that encourages treatment *can never discourage treatment for any unit*, although it may fail to encourage treatment for many.

When treatment effects differ between compliers and other compliance classes, and monotonicity holds, IV analysis identifies the average treatment effect *just for compliers*—what Imbens and Angrist (1994) term the Local Average Treatment Effect (LATE).¹² When we use judge leniency as an instrument for incarceration, this means

¹¹See Mogstad, Torgovitsky and Walters (2021) for discussion of the “partial monotonicity” assumption required for models with multiple instruments.

¹²This is easiest to think about in the binary instrument, binary treatment case with no covariates. Changes beyond that alter the target estimand (see Appendix E). When all treatment effects are

that we identify the effects of incarceration only for people who were right on the cusp of being incarcerated. If Carol had been found guilty of first-degree murder, for instance, she would be sentenced to prison regardless of her judge. She would be an always-taker, and we would be unable to identify treatment effects for people like her.

Of the 34 IV papers we surveyed, only eight clarified that they could only identify the LATE. As Lundberg, Johnson and Stewart (2021) point out, the population of compliers can be extremely small and may be unrelated to the theoretical motivation for the study. Following them, we urge researchers to not only state the estimand but also make clear why compliers are a population of interest.¹³

The judge instrument, and instruments like it, are common in the social sciences. Harding et al. (2018) uses judges as instruments for incarceration, and de Vaan and Stuart (2019) uses doctors’ differential preferences for prescribing opioids as an instrument for taking opioids. Swanson et al. (2015) call these “preference-based” instruments and provide strong evidence that monotonicity is violated in the case of doctors’ preferences for prescribing certain drugs. One way to address monotonicity violations is by constructing a more complicated set of instruments rather than using a unidimensional measure of preference. For instance, Harding et al. (2018) interact individual judges with indicators for the type of crime committed, leading to a set of dozens of instruments. This approach has its own challenges (as we discuss in Section 4.2).

3.5 SUTVA and Positivity

Causal inference with IV requires two additional assumptions that we discuss only briefly (see Imbens and Rubin, 2015; VanderWeele, 2009, for more detail). The stable-unit-treatment-value assumption (SUTVA) requires no “hidden versions” of the instrument or treatment and rules out interference across units. Formally,

Assumption 5: SUTVA

If $Z_i = z$, then $D_i = D_i(z)$. If $Z_i = z$ and $D_i = d$, then $Y_i = Y_i(d, z)$.

Positivity requires that, in every stratum of measured confounders, at least some units receive the instrument (Aronow and Miller, 2019; Hernán and Robins, 2021). Positivity is only required for confounders that are necessary for (unconfoundedness) of the instrument to hold. Formally,

Assumption 6: Positivity

$$0 < \Pr(Z_i = 1 \mid \mathbf{X}_i) < 1,$$

constant, the need for monotonicity disappears (because the knowledge from any subpopulation applies to all units), but this assumption is implausible. See Cui and Tchetgen Tchetgen (2021) and Hartwig et al. (2021) for more thorough discussions of homogeneity assumptions for IV analysis. Swanson et al. (2018) discuss partial identification approaches that avoid monotonicity assumptions.

¹³See Deaton (2009), Heckman (1997), Heckman and Urzua (2010), and Swanson and Hernán (2018) for critiques of the LATE as a causal contrast of interest. See Angrist and Imbens (1999) and Imbens (2010) for defenses.

where \mathbf{X}_i is a vector of observed confounders for unit i .

3.6 Simultaneity and Over-Identification Tests

Substantial minorities of the surveyed papers mentioned two concerns we largely omit from this paper: simultaneity and over-identification tests.

Of the 34 papers we reviewed, 14 cited “simultaneity” or “reverse causation” as part of their justification for using IV. While IV can be used to identify parameters in so-called “simultaneous equation models,” such models typically capture states of *equilibrium*. The canonical case is the identification of supply and demand functions, where we expect the quantity supplied and quantity demanded to be in equilibrium at a given price (Wright, 1928; Imbens, 2014). To adequately approximate recursive relationships, equilibrium must be achieved very fast relative to the interval at which we measure time averages of the treatment and outcome (Fisher, 1970; Richardson and Robins, 2014). When simultaneity is the concern, we only advise using IV in cases where we have strong theoretical reasons to believe the treatment and outcome achieve equilibrium relatively quickly—a condition we don’t think is frequently satisfied.

Of the 34 papers we reviewed, seven report over-identification tests to support the validity of their instruments, as recommended by Bollen (2012). In the case where we have multiple instruments for a given treatment, over-identification tests assess the validity of a subset of the instruments under two strong conditions. First, we must assume that at least some of the instruments are valid. Given how difficult it is to find even a single valid instrument, this assumption often strains credibility. Second, over-identification tests assume constant treatment effects, but in practice it is likely that different instruments have different LATEs (Wooldridge, 2010; Angrist and Pischke, 2008). In Appendix C, we review more compelling ways to investigate an instrument’s validity.

3.7 Estimation

Having established the assumptions needed to identify the causal effect, we briefly review estimation with IVs. While we can use many estimators for IV, we focus on the most common one: two-stage least-squares (2SLS) regression. As the name suggests, there are two stages to this procedure. First, we regress the treatment on the instrument using a linear regression. This is called the “first-stage” regression. In the “second stage,” we regress the outcome on the fitted values of the treatment from the first-stage regression. Note that performing this procedure manually will produce invalid confidence intervals in the second stage, but standard statistical software for 2SLS will report corrected intervals. The regression of the outcome directly on the instrument (without the treatment) is called the “reduced-form” or “intention-to-treat” (ITT) regression. This ITT effect captures the total effect of the instrument on the outcome. Under the exclusion restriction and unconfoundedness, this effect will operate entirely through the

Term	Definition	Estimand
<i>First-stage regression</i>	Regression of D on Z	Average causal effect of the instrument on the treatment
<i>Second-stage regression</i>	Regression of Y on \hat{D} , the fitted values from the first-stage regression	Average treatment effect for compliers
<i>Reduced-form or intent-to-treat (ITT) regression</i>	Regression of Y on Z , not controlling for D	Total average effect of instrument on the outcome, including effects operating through the treatment

Table 4: Terminology for conducting IV analysis with 2SLS regression.

treatment.¹⁴

4 Weakness Exacerbates Three Methodological Problems

Understanding the identification assumptions for IV can help us select better instruments. But perfectly valid instruments are rare, and even fairly credible instruments can lead us astray if we fail to handle them with care. In this section, we describe three methodological problems that arise with IV—only one of which is regularly discussed in published IV papers—and provide guidance on managing each one. We summarize these problems and approaches for addressing each in Table 5.

What ties these distinct problems together is that they are all exacerbated by *weak* instruments. A weak instrument is one that has only a small effect on the treatment. Weakness is a continuum—the weaker an instrument is, the more it can amplify the methodological problems we describe below.

Weakness complicates comparisons between IV and selection on observables. Both approaches require untestable assumptions, but many scholars view IV assumptions as more credible than the assumptions required for selection on observables. IV, however, is far more sensitive to violations of its assumptions than selection on observables is. When we compare the credibility of the two approaches, we have to consider not just the probability that assumptions are violated but also how severely these violations might bias our estimates. In short, the consequences of credibly small violations of

¹⁴In the case without covariates, the 2SLS treatment coefficient is equivalent to the ITT coefficient on the instrument divided by the first-stage coefficient on the instrument.

IV assumptions may be worse than the consequences of credibly large violations of selection-on-observables assumptions.

Understanding how weakness magnifies these problems can enable researchers to make more prudent choices between IV and selection on observables and empower readers to become more critical consumers of IV studies.

Problem	What is it?	How do I address it?
<i>Identification Bias</i>	An asymptotic bias that can be severe under seemingly trivial violations of unconfoundedness, monotonicity, or the exclusion restriction.	Conduct a bias analysis (see Section 4.1), and run <i>well-powered</i> placebo tests if possible.
<i>Estimation Bias</i>	A finite-sample bias that pulls 2SLS estimates toward the OLS estimator (i.e., the OLS model that uses the same specification apart from the instrumented treatment).	Report the robust partial F -statistic for the instrument in the first-stage. Use Anderson–Rubin confidence bands for the estimated treatment effect. Assess how sensitive results are to the removal of outliers.
<i>Type-M Error</i>	For a given hypothetical effect size, how exaggerated the estimate has to be to achieve statistical significance.	Report the expected type-M error for a hypothetical effect size chosen pre-analysis.

Table 5: Three methodological problems that are exacerbated by weakness. While we are not the first to point out the distinct problems arising from weak instruments, researchers rarely acknowledge identification bias or the type-M error. We avoid the common description of the first-stage F -statistic as a “test” of weakness and emphasize that even an analysis that does not suffer from estimation bias can still suffer from substantial identification bias—and vice versa.

4.1 Identification Bias

As we emphasized above, the identification assumptions for IV are strong, and we suspect they often fail to hold perfectly in the real world. We might nonetheless prefer an imperfect IV analysis over an imperfect selection-on-observables design, suspecting the former to be *less* biased than the latter. For example, a researcher might believe that while both the instrument and the treatment are confounded with the outcome, the instrument is “less” confounded than the treatment. Unfortunately, such comparisons are complicated by weak instruments. Bias from violations of unconfoundedness, monotonicity, and the exclusion restriction are inflated by the inverse of the first-stage

effect. As a result, seemingly “small” violations of unconfoundedness or the exclusion restriction can generate large biases. We call this *identification bias* because it stems from violations of the identification assumptions. We include a stylized example of the surprising severity of identification bias in Figure 5.

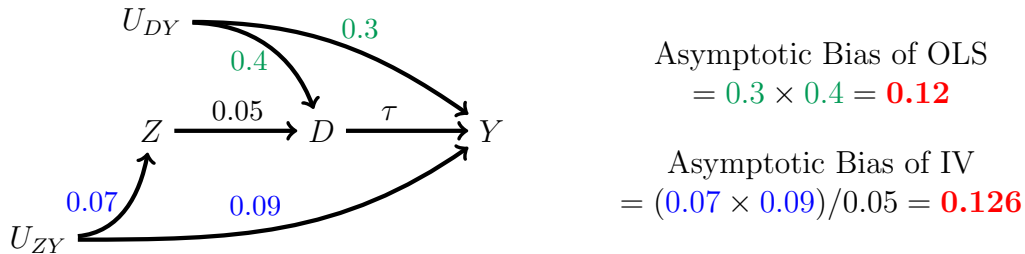


Figure 5: Linear structural equation model with constant effects where both the treatment and instrument are confounded. Each variable is continuous, and the numbers in the LSEM represent hypothetical marginal effects indicated by the corresponding arrows. Y represents the outcome, D the treatment, Z the instrument, U_{DY} an unmeasured treatment–outcome confounder, and U_{ZY} an unmeasured instrument–outcome confounder. OLS refers to an OLS regression of Y on D , while IV refers to a 2SLS regression using Z as an instrument for D . Despite the fact that the effects of U_{DY} (0.3 and 0.4) are more than 4 times larger, respectively, than the effects of U_{ZY} (0.07 and 0.09), the IV estimator exhibits larger (approximate) asymptotic bias. In practice, we typically run more complicated models than this and avoid constant-effects assumptions. This is simply a toy example meant to illustrate a general principle (Pearl, 2013).

Researchers can address identification bias through *bias analysis*.¹⁵ Rather than asking *whether* identification bias is present, this procedure asks *how severe* identification bias would have to be in order to change the conclusions of our study. Of the 34 papers we surveyed, none employed a bias analysis, and 18 (60%) failed to even report the strength of the first-stage effect, which would give the reader a sense of how serious the bias amplification might be. In what follows, we introduce a running example from Conley and Glauber (2006) and illustrate three forms of bias analysis.

4.1.1 Running Example for Bias Analysis

Conley and Glauber (2006) study of the effects of sibship size on private school attendance. The treatment is whether a family has three or more children; the outcome is whether a child attends private school; and the instrument is whether the first two children are the same sex. The logic of the instrument is that parents prefer having at least one boy and one girl, so having two children of the same sex might encourage some parents to have a third child.

Conley and Glauber (2006) focus on effects for second-born boys. In doing so, however, they alter the instrument. By analyzing effects only for second-born sons,

¹⁵In the causal inference literature, this is more commonly called “sensitivity analysis,” but we avoid this term because it is also used to refer to other practices in the applied literature, such as making sure the coefficient of interest remains statistically significant under slightly different model specifications.

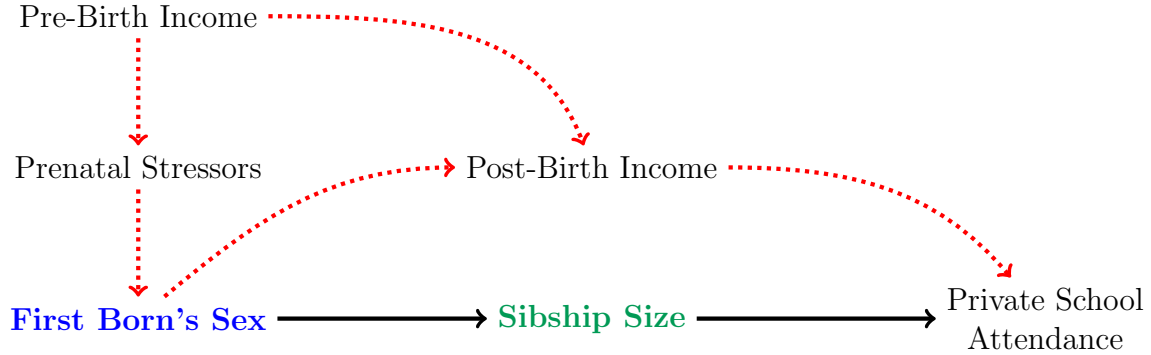


Figure 6: DAG illustrating potential violations of unconfoundedness and the exclusion restriction in Conley and Glauber (2006). The exclusion restriction violation is induced by only examining treatment effects for second-born boys.

the instrument becomes the sex of the first-born child. Using the sex of the first-born child as an instrument likely violates the exclusion restriction. The sex of the first-born affects a wide range of outcomes, including the speed of the transition to marriage, the probability of divorce, and father’s wages and work hours (Lundberg and Rose, 2002, 2003; Dahl and Moretti, 2008). Both marriage and father’s wages plausibly affect household finances, which could in turn influence private school attendance.¹⁶

There is also some evidence that unconfoundedness is violated. External stressors reduce the proportion of male births in a population (Hansen, Møller and Olsen, 1999; Catalano et al., 2005; James, 2009). If such external stressors are more common among materially disadvantaged families, and material disadvantage in turn affects private school attendance, unconfoundedness may be violated. We illustrate these potential violations of the exclusion restriction and unconfoundedness in the Figure 6.

4.1.2 A Simple Bias Analysis for the Exclusion Restriction

One easy-to-implement bias analysis consists of simply reporting a single additional regression—the “intention-to-treat” (ITT) regression. Recall that the ITT estimates the total effect of the instrument on the outcome. If we assume a linear, constant-effects model¹⁷, then the ITT estimate tells us what the direct effect of the instrument would have to be to explain away our estimated treatment effect entirely—that is, to reduce our estimated treatment effect to 0. In Conley and Glauber (2006), the estimated ITT is -0.00368 .¹⁸ This means that a direct effect of -0.00368 —less than *four-tenths of a*

¹⁶Conley and Glauber (2006) only examine effects for households with married couples, which could block some of these exclusion restriction violations, but also risks inducing confounding (see Section 3.3.2).

¹⁷Using a linear, constant-effects model simplifies the bias analysis. If the analysis is unfavorable under the linear, constant-effects setting, we find it unlikely to be favorable under a heterogeneous-effects model.

¹⁸Even when the ITT is not explicitly reported, as in Conley and Glauber (2006), we can calculate it by multiplying together the first-stage estimate (0.08) and the LATE estimate (-0.046).

percentage point—could *entirely* explain away the estimated treatment effect.

The literature suggests that a first-born boy should increase, rather than decrease, the probability that a second-born boy attends private school, so it is unlikely that an exclusion restriction violation alone could explain away the findings in Conley and Glauber (2006). But the analysis illustrates that seemingly trivial violations of identification assumptions can generate counterintuitively large biases in practical IV applications.

4.1.3 Bias Analysis Plots for the Exclusion Restriction

Conley, Hansen and Rossi (2012) and Wang et al. (2018) introduce bias analysis methods that adjust confidence intervals to incorporate hypothetical violations of the exclusion restriction. We present a modified version of these analyses that shows how our estimated treatment effects would change under a range of potential exclusion restriction violations. The procedure works as follows:

1. Specify a hypothetical value θ that represents the direct effect of the instrument Z on the outcome Y .
2. Calculate an adjusted outcome $Y_{\text{adj}} = Y - \theta Z$.
3. Estimate treatment effects with 2SLS using Y_{adj} instead of Y , reporting 95% confidence intervals.
4. Repeat Steps 1–3 for a range of θ values, making sure to include values of θ sufficient to explain away the point estimate entirely.
5. Plot the results.

We include a plot with the results from this sensitivity analysis below.¹⁹ Following Wang et al. (2018), we use Anderson–Rubin confidence intervals, although we do not standardize θ as they do. Our results show that if a first-born son reduced the probability of private school attendance by 0.0015–0.15 percentage points—then 95% confidence intervals on the estimated treatment effect would overlap with 0. We include replication R code for this analysis and plot in Appendix D.

4.1.4 Cinelli and Hazlett’s Bias Analysis

Cinelli and Hazlett (2020) offer a useful bias analysis for both unmeasured confounding and exclusion restriction violations in 2SLS models. One key difference in interpretation is that sensitivity is assessed in terms of predictive power rather than causal

¹⁹Although we were unable to replicate Conley and Glauber’s (2006) results exactly, our estimates are close to theirs. For white, second-born boys, we have a first-stage estimate of 0.077, compared with their estimate of 0.08. For the estimated LATE, we have -0.057 , compared with their estimate of -0.046 . Our estimate of the standard error on the LATE is 0.020, compared with their estimate of 0.022.

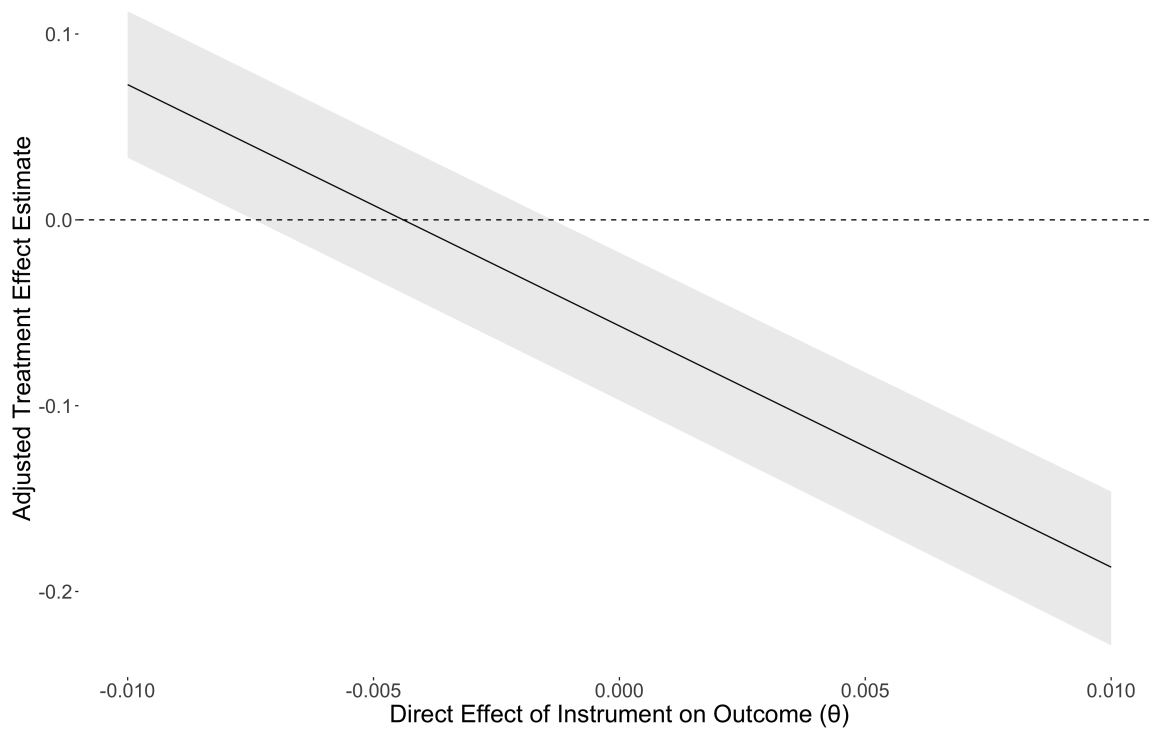


Figure 7: Adjusted treatment effect estimates for each value of θ with 95% Anderson–Rubin confidence intervals. In this case, θ represents the effect of a first-born son on the probability of private school attendance that does not occur through sibship size. An effect of -0.15 percentage points would render the treatment effect statistically insignificant at the $\alpha = 0.05$ level. Replication code for conducting the analysis and generating the plot is included in Appendix D.

effects. This method assesses sensitivity in terms of how much of the residual variation in the instrument or outcome can be *predicted* by an unmeasured confounder. In particular, the “robustness value” indicates how strongly, in terms of R^2 , the unmeasured confounder must predict residual variation in *both* the instrument and the outcome in order to reduce the estimated ITT to 0.²⁰

To perform this analysis, we use the authors’ `sensemakr` package for R to calculate two robustness values. For Conley and Glauber’s (2006) results, we find that if an unmeasured confounder had a partial R^2 of 0.0067 on both the instrument and outcome, it would reduce the effect estimate to 0.²¹ An unmeasured confounder need only predict 0.67% of the residual variation in the instrument and outcome in order to reduce the estimated treatment effect to 0. Additionally, a partial R^2 of 0.002 with both the instrument and outcome would be sufficient to make the treatment effect statistically insignificant. Appendix D includes replication code for this analysis as well as contour plots for visualizing the results.

The bias analyses presented above demonstrate how a weak first stage turns seemingly minor violations of identification assumptions into large biases. Because the first-stage is relatively weak in Conley and Glauber (2006), only small violations of unconfoundedness and the exclusion restriction were necessary to explain away the reported treatment effect. Next, we turn to situations where the first-stage relationship is weak relative to the sampling variability of the first-stage, producing substantial estimation bias.

4.2 Estimation Bias

Another problem aggravated by instrument weakness is what we call the *estimation bias* of 2SLS. The 2SLS estimator is biased toward the OLS regression of the outcome on the treatment (including any covariates used in the 2SLS model).²² To see why, recall that the second stage in 2SLS is an OLS regression that replaces *observed* values of the treatment with *predicted* values of the treatment from the first-stage regression. Because we don’t know the true, population first-stage effect—it must be estimated from our sample—these predicted values reflect some of the randomness in our treatment variable. The weaker the first stage, then, the closer the 2SLS sampling distribution moves toward the probability limit of OLS. This bias is always present—even when the identification assumptions hold perfectly—but it becomes negligible when the instrument

²⁰Researchers can conduct bias analyses for both the first-stage and ITT, but for simplicity we just perform the analysis for the ITT. Note also that Cinelli and Hazlett (2020) propose other sensitivity statistics to report, but we limit our discussion to just one.

²¹Note, however, that the confounder would have to positively predict the instrument while negatively predicting the outcome, or vice versa.

²²Limited information maximum likelihood (LIML) estimation is sometimes recommended as an unbiased alternative to 2SLS. LIML, however, requires multiple instruments and assumes constant effects (Angrist and Pischke, 2008). In the presence of treatment heterogeneity, the LIML estimand is not even guaranteed to lie within the convex hull of LATEs (Kolesár, 2013).

is strong enough and the sample size large enough.²³

How do we know when the instrument is “strong enough?” The conventional rule of thumb is that when the partial F -statistic on the instrument in the first-stage regression is greater than 10, estimation bias should be small. Stock and Yogo (2005) provide a formal justification for this rule along with a comprehensive sets of thresholds indicating different amounts of bias. Of the 34 sociology papers we surveyed, only 18 (53%) reported results from any F -test. Furthermore, we identified only two papers that explicitly reported an F -test robust to heteroskedasticity or clustering.²⁴ For context, 26 papers used either panel data or clustered data, indicating that reporting robust partial F -statistics should be far more common. We recommend sociologists regularly report robust partial F -statistics when using a single instrument and the Olea and Pflueger (2013) effective F -test when using multiple instruments (Andrews, Stock and Sun, 2019).

Estimation bias can be particularly severe when using multiple weak instruments. For instance, Bound, Jaeger and Baker (1995) find strong evidence of estimation bias in a study with over 300,000 observations that uses 180 instruments. Similarly, Harding et al. (2018) report partial F -statistics well below 10 when using judge–covariate interactions as instruments in a sample with over 100,000 observations. Notably, estimates from this specification are very close to the OLS estimates, which is consistent with the direction of estimation bias. Furthermore, the F -statistic thresholds for inference grow rapidly with the number of instruments, so the confidence intervals reported in Harding et al. (2018) may be too narrow (Stock and Yogo, 2005).²⁵

Even when our F -statistics indicate non-trivial estimation bias, we can obtain valid confidence intervals under some assumptions.²⁶ Following Andrews, Stock and Sun (2019) and Young (2019), we recommend researchers report both Anderson–Rubin (AR) and bootstrapped confidence intervals, particular when using clustered or panel data (Anderson and Rubin, 1949).^{27,28} Of the 34 papers we reviewed, none reported AR or

²³More precisely, 2SLS is approximately median-unbiased when our instrument is strong enough.

²⁴It is possible that other authors reported robust F -statistics without explicitly stating the type of F -statistic being reported. Because replication code for papers in our sample is not publicly available, we are unable to explore this possibility.

²⁵In an oft-overlooked table, Stock and Yogo (2005) include a set of critical values for bounding the size (the maximum probability of committing a type-I error) of a 5% significance test. With a single treatment, 30 instruments, and a size of 0.10 (which is already larger than the nominal size of 0.05), we would need $F > 86.17$. Harding et al. (2018) did not report how many instruments they used, so it is unclear what critical value applies.

²⁶Anderson–Rubin confidence intervals assume that the first-stage and ITT estimates are distributed multivariate Normal with known variance (Andrews, Stock and Sun, 2019). Young (2019) finds that variance estimates in published research are quite noisy, and the sensitivity of results to outliers indicates the Normality assumption is often violated.

²⁷Recently, Angrist and Kolesár (2021) argue that in models with a single instrument and single treatment, conventional confidence intervals will have fairly accurate coverage. Crucially, however, they focus on the i.i.d. setting. In contrast, sociologists tend to use panel data in IV settings, where AR confidence sets and the bootstrap may be more reliable.

²⁸Andrews, Stock and Sun (2019) recommend against the bootstrap because it fails to provide valid

bootstrapped confidence sets. Appendix D provides names of software packages in both R and Stata for conducting the diagnostics we recommend.

4.3 Type-M Error

Finally, weak instruments can increase the expected type-M error, where M stands for the *m*agnitude of effect estimates. When treatment effects are small, using high-variance estimators can systematically exaggerate the magnitude of effect sizes that make it into published work (Gelman and Carlin, 2014). The logic is that achieving statistical “significance” with a high-variance estimator requires an unusually large effect estimate, and statistically “*insignificant*” results often go unreported. As a result, published findings may systematically overestimate the magnitude of causal effects.

Reviews of published IV findings are consistent with systematic exaggeration of effect sizes. In political science, Lal et al. (2021) find that IV estimates are almost always larger in magnitude than OLS estimates, despite the fact that we often use IV because we are worried about OLS estimates being biased away from 0. Alarming, they find that in 32.8% of papers, the IV estimates are at least *five times larger* than OLS estimates. While the authors primarily attribute this finding to assumption violations, it is also consistent with widespread type-M errors. Similarly, in a review of finance studies using instrumental variables, Jiang (2017) finds that IV estimates are, on average, 9 times larger in magnitude than their corresponding OLS estimates in cases where we would expect the IV estimate to be smaller in magnitude.

Of the 17 papers in our sample that report both OLS and 2SLS results, the 2SLS estimates were larger in 15 (88.2%).²⁹ In 11 (64.7%) of the papers, 2SLS estimates were more than twice as large, and in five (29.4%), they were more than five times larger. We emphasize that this is not evidence of any wrongdoing on the part of any researchers. Rather, it is evidence that publishing only statistically “significant” findings can lead to systematically exaggerated effect sizes when using high-variance estimators like 2SLS.

Type-M error is difficult to address because it involves a complex interplay between the variance of the estimator and the practices of scientific publishing. Following Gelman and Carlin (2014), we can calculate expected type-M error for IV estimates if we are willing to specify a hypothetical “true” effect size based on prior research. We also recommend plotting the estimates from IV and selection on observables, with confidence intervals, side by side. Such figures convey how precise our IV estimates are relative to our selection-on-observables estimates. Finally, researchers reporting IV estimates considerably larger than selection-on-observables estimates should make a strong case for why selection-on-observables estimates would be biased toward zero, since the typical

coverage in weak-IV settings. Young (2019), in contrast, argues that the bootstrap will nonetheless work better than alternatives (like AR confidence sets) in high-leverage settings, where the assumptions required for AR confidence sets may fail.

²⁹A number of papers only report 2SLS models and others use alternative IV estimators for binary outcomes, such as Two-Stage Residual Inclusion (2SRI) models. Without publicly available replication code, we are unable to compare OLS and 2SLS estimates for these papers.

motivation for IV is that selection on observables may be biased away from zero.

4.4 Weakness Complicates the Comparison of IV and Selection on Observables

Most causal work in sociology implicitly follows a selection-on-observables design (whether via regression, matching, weighting or some other estimation strategy). Authors turn to IV because the identification assumptions often seem more plausible than assuming the treatment is (conditionally) unconfounded. But comparisons between these approaches are more complicated than they appear. As we have seen, weak instruments create challenges for identification, for estimation, and for interpreting reported results in the literature. When instruments are even modestly weak, we must handle IV estimates with care because they can be substantially more sensitive to assumptions of violations than selection-on-observables estimates.

5 A Checklist for Conducting and Reporting Instrumental Variables Analysis

Taking inspiration from Sovey and Green (2011) and Lal et al. (2021), we have compiled a checklist summarizing our recommendations to aid researchers in conducting and reporting IV analysis. Like Sovey and Green (2011), we emphasize the importance of clarifying the estimand and defending assumptions. We particularly underscore the value of communicating those assumptions to readers in non-technical language rooted in the substantive problem at hand. Like Lal et al. (2021), we encourage using the bootstrap and AR confidence intervals for inference. They also encourage routine use of bias analysis, along with placebo tests to set the bias analysis parameters. While this is highly effective when a placebo sub-population is available, we expect such populations to be difficult to locate in many sociological settings (but see Appendix C for more on placebo tests and balance checks). For this reason, we place more emphasis on simpler forms of bias analysis.

Our checklist focuses on identification of the LATE, although most items still apply if the researcher is willing to invoke constant effects assumptions to target the ATE. We show an abridged version of the checklist in Figure 8.

1. State the Theoretical Estimand

1a. State the theoretical estimand: a Local Average Treatment Effect (Section 3) or weighted average of LATEs (Appendix E). Following Lundberg, Johnson and Stewart (2021), we urge researchers to explicitly state what they term the *theoretical estimand* and clarify how it connects to the argument of the paper. Without imposing strong restrictions on treatment effect heterogeneity, IV analysis identifies the average treatment effect only for *compliers*—those induced into treatment by the instrument. This

can be phrased substantively—for instance, Kirk (2009) clearly states that he is estimating a treatment effect only “for parolees who would not have moved had it not been for Hurricane Katrina” (495). We also recommend discussing what can and cannot be learned from such a subpopulation. When non-dichotomous treatments or instruments are used, or when covariates are included in the model, 2SLS will identify a weighted average of LATEs, a nuance we discuss in Appendix E.

2. Explain Assumptions

2a. State the identification assumptions: unconfoundedness, exclusion restriction, monotonicity, instrument relevance, SUTVA, and positivity (Section 3). Researchers should state these assumptions in substantive terms for the reader. Consider the unconfoundedness assumption in the context of Kirk (2009). This assumption requires that being released after Hurricane Katrina shares no unmeasured common causes with being re-arrested. Compare this explanation with the more opaque statement that the instrument must be uncorrelated with the error term in some regression model, a common description in the applied literature.

2b. Discuss plausible violations of any identification assumptions. In practice, it is rare that no plausible violations exist. Indeed, even the Vietnam draft lottery instrument, which helped popularize IV methods in economics and launch the “credibility revolution,” likely violates the exclusion restriction for many outcomes. The instrument is whether a person received a high or low lottery number for the Vietnam draft, and the treatment is whether the person served in the military. But the lottery numbers caused some people to change their educational plans in order to avoid the draft, and education affects a wide range of outcomes (Angrist, Imbens and Rubin, 1996). Help the reader understand that the instrument can have no “direct” effect on the outcome in the sense that it can cause the outcome *only through the treatment* (Section 3.3.1). Weather-based instruments, large-scale disasters, and historical instruments may need special care in this regard. Remember that while conditioning on post-instrument covariates may block exclusion restriction violations, they can induce violations of unconfoundedness (Section 3.3.2). Finally, note that coarsely measured treatments often violate the exclusion restriction (Section 3.3.3).

2c. If using a proxy instrument, be sure to state the causal instrument (Appendix B). The estimand and identification assumptions are defined with respect to the causal instrument—not the proxy instrument. A proxy instrument is one that does not cause the treatment, but rather shares an unmeasured common cause with the treatment. The unmeasured common cause is the causal instrument. Crucially, identification assumptions must hold with respect to the causal instrument. Furthermore, compliers are those who are induced into treatment by the causal instrument—not the proxy instrument, which exerts no causal influence on the treatment. We discuss examples in Appendix B.

3. Report Results with Weak-Instrument-Robust Confidence Intervals

3a. Report the estimated first-stage effect, “reduced-form” or ITT effect, and treatment effect estimates with 95% confidence intervals (Section 4). The estimated first-stage effect represents how strongly the instrument affects the treatment. The smaller it is, the greater the potential for both identification bias and estimation bias. Under certain assumptions, the ITT effect estimate represents how strong the direct effect of the instrument would have to be to explain away the effect estimate (Section 4.1).

3b. Report Anderson–Rubin confidence sets and bootstrapped confidence intervals for the treatment effect (Section 4.2). In the case with a single instrument and treatment, Anderson–Rubin confidence sets will be robust to weak instruments and heteroskedasticity under assumptions described in Andrews, Stock and Sun (2019). Young (2019) finds that bootstrapped confidence intervals perform better in real-world settings because these assumptions are likely violated.³⁰ In the multi-instrument setting, researchers can use Moreira’s (2003) conditional likelihood ratio test, although this is not robust to heteroskedasticity.

4. Assess Bias and Type-M Error

4a. Conduct bias analysis for violations of key assumptions (Section 4.1). When using IV in observational settings, some of our identification assumptions will likely be violated. Bias analysis allows us to assess *how badly* our assumptions need to be violated in order to change the conclusions of our study. We describe several different bias analyses Section 4.1, and provide example code for conducting these analyses in Appendix D.

4b. Report the results of placebo tests and scaled balance checks if possible. Placebo tests can sometimes provide evidence in support of the exclusion restriction and unconfoundedness assumption. Balance checks, when scaled by the inverse of the first-stage, can provide evidence in favor of unconfoundedness. We describe these procedures and their limitations in Appendix C.

4c. Report a weak-instrument diagnostic that is robust to heteroskedasticity and clustering, namely, the Oleg and Pflueger effective F -statistic (Section 4.2). In the case with a single instrument and treatment, this quantity is equivalent to the Kleibergen–Paap F -statistic. In general, an effective F -statistic > 10 will indicate small relative bias of 2SLS, but with multiple instruments, the F -statistic should be much higher to ensure accurate confidence interval coverage.

4d. Report the expected type-M error (Section 4.3). The expected type-M error quantifies how exaggerated an effect estimate would have to be in order to achieve statistical significance. This quantity depends on a hypothetical effect size, which can be provided by either the existing literature or, if the existing literature is uninformative, the OLS estimate.

³⁰In particular, he assesses performance in settings with clustered errors and high-leverage observations.

Instrumental Variables Checklist

1. State the Theoretical Estimand

- a. State the estimand: a Local Average Treatment Effect (Section 3.4) or weighted average of LATEs (Appendix E). Explain how compliers may differ from non-compliers and why we should be interested in the complier average treatment effect.

2. Explain Assumptions

- a. State the identification assumptions in clear, non-technical language. unconfoundedness, exclusion restriction, monotonicity, instrument relevance, SUTVA, and positivity (Section 3).
- b. Discuss plausible violations of any identification assumptions. It is unlikely that each and every assumption is beyond reproach.
- c. If using a *proxy* instrument, be sure to state the *causal* instrument (Appendix B). The estimand and identification assumptions are defined with respect to the causal instrument—not the proxy instrument.

3. Report Results with Weak-Instrument-Robust Confidence Intervals

- a. Report the estimated first-stage effect, “reduced-form” or ITT effect, and treatment effect with 95% confidence intervals (Section 4).
- b. Report robust, Anderson–Rubin, and bootstrapped confidence intervals for the treatment effect (Section 4.2).

4. Assess Bias and Expected Type-M Error

- a. Conduct bias analysis for violations of unconfoundedness or the exclusion restriction (Section 4.1).
- b. Report the results of placebo tests and scaled balance checks if possible (Appendix C).
- c. Report estimation-bias diagnostics that are robust to heteroskedasticity and clustering, like the Olea and Pflueger effective F -statistic (Section 4.2).
- d. Compare the OLS and 2SLS estimates and report the expected Type-M error (Section 4.3).

Figure 8: A summary of the checklist for conducting instrumental variables analysis described in this section with references to the relevant portions of the text.

6 Conclusion

In this paper, we provided an introduction to assumptions required for IV methods and outlined various tools for improving empirical practice. We documented ways to improve the communication of these assumptions to readers. We also identified under-utilized diagnostics and approaches to constructing confidence intervals. We compiled everything into a simple checklist.

Researchers typically employ IV to make causal claims when they suspect that a selection-on-observables strategy will be biased due to unmeasured confounding. Our discussion highlighted potential sources of bias in IV, focusing on sensitivity to even minor violations in the weak-instrument setting. We hope this helps analysts and researchers improve their intuitions about which estimates may be more biased. More generally, we argue that we should spend less time thinking about *whether* our causal estimates are biased and more time reasoning about *how severe* the bias might be. Bias analysis, balance checks, and placebo tests are some of the tools we can use to help assess this bias—for both IV and selection on observables.

In practice, researchers often use multiple identification strategies and consider the combination of results together. Triangulating different evidence bases can be a powerful way to build support for a claim using individually imperfect pieces of evidence (Cochran and Chambers, 1965; Martin, 2017; Karmakar, French and Small, 2019). We have focused on interpretation of IV in isolation, but in our sample of papers, we found that IV results were often presented alongside selection-on-observables results—although, concerningly, IV typically produced larger estimates. Thinking about common combinations of results is an important area for future work.

Sociologists have understandably been attracted to IV—it is a powerful strategy that can accomplish the miraculous task of causal identification in the presence of unmeasured confounding. We simply emphasize here that it is also a fragile technique that we must handle with care.

References

- Abadie, Alberto. 2003. “Semiparametric instrumental variable estimation of treatment response models.” *Journal of econometrics* 113(2):231–263.
- Ahlfeldt, Gabriel M and Elisabetta Pietrostefani. 2019. “The economic effects of density: A synthesis.” *Journal of Urban Economics* 111:93–107.
- Anderson, Theodore W and Herman Rubin. 1949. “Estimation of the parameters of a single equation in a complete system of stochastic equations.” *The Annals of mathematical statistics* 20(1):46–63.
- Andrews, Isaiah, James H Stock and Liyang Sun. 2019. “Weak instruments in instrumental variables regression: Theory and practice.” *Annual Review of Economics* 11:727–753.

- Angelucci, Manuela. 2012. "US border enforcement and the net flow of Mexican illegal migration." *Economic Development and Cultural Change* 60(2):311–357.
- Angrist, Joshua D and Alan B Krueger. 1999. Empirical strategies in labor economics. In *Handbook of labor economics*. Vol. 3 Elsevier pp. 1277–1366.
- 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 and Guido W Imbens. 1999. "Comment on James J. Heckman," Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations". *Journal of Human Resources* pp. 823–827.
- 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 D and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics*. Princeton university press.
- Angrist, Joshua D, Kathryn Graddy and Guido W Imbens. 2000. "The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish." *The Review of Economic Studies* 67(3):499–527.
- Angrist, Joshua and Michal Kolesár. 2021. "One Instrument to Rule Them All: The Bias and Coverage of Just-ID IV." *arXiv preprint arXiv:2110.10556* .
- Aronow, Peter M and Benjamin T Miller. 2019. *Foundations of agnostic statistics*. Cambridge University Press.
- Aronow, Peter M and Cyrus Samii. 2016. "Does regression produce representative estimates of causal effects?" *American Journal of Political Science* 60(1):250–267.
- Baiocchi, Michael, Jing Cheng and Dylan S Small. 2014. "Instrumental variable methods for causal inference." *Statistics in medicine* 33(13):2297–2340.
- Bollen, Kenneth A. 2012. "Instrumental variables in sociology and the social sciences." *Annual Review of Sociology* 38:37–72.
- 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.

- Card, David. 1995. Using Geographic Variation in College Proximity to Estimate the Return to Schooling. In *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*, ed. EK Christofides, LN abd Grant and R Swidinsky. Toronto: University of Toronto Press.
- Catalano, Ralph, Tim Bruckner, Jeff Gould, Brenda Eskenazi and Elizabeth Anderson. 2005. "Sex ratios in California following the terrorist attacks of September 11, 2001." *Human Reproduction* 20(5):1221–1227.
- Chalak, Karim. 2017. "Instrumental variables methods with heterogeneity and mis-measured instruments." *Econometric Theory* 33(1):69–104.
- Cinelli, Carlos and Chad Hazlett. 2020. "An omitted variable bias framework for sensitivity analysis of instrumental variables." *Work. Pap* .
- Cochran, William G and S Paul Chambers. 1965. "The planning of observational studies of human populations." *Journal of the Royal Statistical Society. Series A (General)* 128(2):234–266.
- Conley, Dalton and Rebecca Glauber. 2006. "Parental educational investment and children's academic risk estimates of the impact of sibship size and birth order from exogenous variation in fertility." *Journal of human resources* 41(4):722–737.
- Conley, Timothy G, Christian B Hansen and Peter E Rossi. 2012. "Plausibly exogenous." *Review of Economics and Statistics* 94(1):260–272.
- Cornelissen, Thomas, Christian Dustmann, Anna Raute and Uta Schönberg. 2016. "From LATE to MTE: Alternative methods for the evaluation of policy interventions." *Labour Economics* 41:47–60.
- Cui, Yifan and Eric Tchetgen Tchetgen. 2021. "A semiparametric instrumental variable approach to optimal treatment regimes under endogeneity." *Journal of the American Statistical Association* 116(533):162–173.
- Currie, Janet, Jonas Jin and Molly Schnell. 2019. "US Employment and Opioids: Is There a Connection?" *Health and Labor Markets* 47:253–280.
- Dahl, Gordon B and Enrico Moretti. 2008. "The demand for sons." *The review of economic studies* 75(4):1085–1120.
- De Chaisemartin, Clément and Xavier d'Haultfoeuille. 2020. "Two-way fixed effects estimators with heterogeneous treatment effects." *American Economic Review* 110(9):2964–96.
- de Vaan, Mathijs and Toby Stuart. 2019. "Does intra-household contagion cause an increase in prescription opioid use?" *American Sociological Review* 84(4):577–608.

- Deaton, Angus. 2009. Instruments of Development: Randomisation in the Tropics, and the Search for the Elusive Keys to Economic Development. In *Proceedings of the British Academy*. Vol. 162 pp. 123–160.
- Elwert, Felix and Christopher Winship. 2014. “Endogenous selection bias: The problem of conditioning on a collider variable.” *Annual review of sociology* 40:31–53.
- Evdokimov, Kirill S and Michal Kolesár. 2019. Inference in Instrumental Variables Analysis with Heterogeneous Treatment Effects. Working paper.
- Fisher, Franklin M. 1970. “A correspondence principle for simultaneous equation models.” *Econometrica: Journal of the Econometric Society* pp. 73–92.
- Gelman, Andrew and John Carlin. 2014. “Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors.” *Perspectives on Psychological Science* 9(6):641–651.
- Glynn, Adam N, Miguel R Rueda and Julian Schuessler. 2021. Post-Instrument Bias in Linear Models. Working paper.
- Greene, William H. 2008. *Econometric analysis*. 6 ed. Pearson Education.
- Hansen, Dorthe, Henrik Møller and Jørn Olsen. 1999. “Severe periconceptional life events and the sex ratio in offspring: follow up study based on five national registers.” *Bmj* 319(7209):548–549.
- Harding, David J, Jeffrey D Morenoff, Anh P Nguyen and Shawn D Bushway. 2018. “Imprisonment and labor market outcomes: Evidence from a natural experiment.” *American Journal of Sociology* 124(1):49–110.
- Hartwig, Fernando Pires, Linbo Wang, George Davey Smith and Neil Martin Davies. 2021. “Homogeneity in the instrument-treatment association is not sufficient for the Wald estimand to equal the average causal effect when the exposure is continuous.” *arXiv preprint arXiv:2107.01070* .
- Heckman, James. 1997. “Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations.” *Journal of human resources* pp. 441–462.
- Heckman, James J and Edward Vytlacil. 2001. “Policy-relevant treatment effects.” *American Economic Review* 91(2):107–111.
- Heckman, James J and Edward Vytlacil. 2005. “Structural equations, treatment effects, and econometric policy evaluation 1.” *Econometrica* 73(3):669–738.
- Heckman, James J and Sergio Urzua. 2010. “Comparing IV with structural models: What simple IV can and cannot identify.” *Journal of Econometrics* 156(1):27–37.

- Heckman, James J, Sergio Urzua and Edward Vytlacil. 2006. "Understanding instrumental variables in models with essential heterogeneity." *The review of economics and statistics* 88(3):389–432.
- Hernán, Miguel A and James M Robins. 2006. "Instruments for causal inference: an epidemiologist's dream?" *Epidemiology* pp. 360–372.
- Hernán, Miguel A and James M Robins. 2021. "Causal inference."
- Hipp, John R. 2007. "Block, tract, and levels of aggregation: Neighborhood structure and crime and disorder as a case in point." *American Sociological Review* 72(5):659–680.
- Ho, Daniel E, Kosuke Imai, Gary King and Elizabeth A Stuart. 2007. "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference." *Political analysis* 15(3):199–236.
- Holland, Paul W. 1986. "Statistics and causal inference." *Journal of the American statistical Association* 81(396):945–960.
- Imai, Kosuke and In Song Kim. 2021. "On the use of two-way fixed effects regression models for causal inference with panel data." *Political Analysis* 29(3):405–415.
- Imbens, Guido W. 2010. "Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic literature* 48(2):399–423.
- Imbens, Guido W. 2014. "Instrumental Variables: An Econometrician's Perspective." *Statistical Science* 29(3):323–358.
- Imbens, Guido W and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.
- Imbens, Guido W and Joshua D Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica: Journal of the Econometric Society* pp. 467–475.
- Jackson, John W and Sonja A Swanson. 2015. "Toward a clearer portrayal of confounding bias in instrumental variable applications." *Epidemiology (Cambridge, Mass.)* 26(4):498.
- James, William H. 2009. "The variations of human sex ratio at birth during and after wars, and their potential explanations." *Journal of Theoretical Biology* 257(1):116–123.
- Jiang, Wei. 2017. "Have instrumental variables brought us closer to the truth." *The Review of Corporate Finance Studies* 6(2):127–140.

- Karmakar, B, B French and DS Small. 2019. “Integrating the evidence from evidence factors in observational studies.” *Biometrika* 106(2):353–367.
- Kennedy, Edward H, Scott Lorch and Dylan S Small. 2019. “Robust causal inference with continuous instruments using the local instrumental variable curve.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81(1):121–143.
- Kirk, David S. 2009. “A natural experiment on residential change and recidivism: Lessons from Hurricane Katrina.” *American Sociological Review* 74(3):484–505.
- Knox, Dean, Will Lowe and Jonathan Mummolo. 2020. “Administrative records mask racially biased policing.” *American Political Science Review* 114(3):619–637.
- Kolesár, Michal. 2013. “Estimation in an instrumental variables model with treatment effect heterogeneity.” *Unpublished Manuscript* .
- Laidley, Thomas and Dalton Conley. 2018. “The effects of active and passive leisure on cognition in children: Evidence from exogenous variation in weather.” *Social Forces* 97(1):129–156.
- Laitin, David D and Rajesh Ramachandran. 2016. “Language policy and human development.” *American Political Science Review* 110(3):457–480.
- Lal, Apoorva, Mackenzie William Lockhart, Yiqing Xu and Ziwen Zu. 2021. How Much Should We Trust Instrumental Variable Estimates in Political Science? Practical Advice based on Over 60 Replicated Studies. Working paper.
- Levitt, Steven D. 2002. “Using electoral cycles in police hiring to estimate the effects of police on crime: Reply.” *American Economic Review* 92(4):1244–1250.
- Lundberg, Ian, Rebecca Johnson and Brandon M Stewart. 2021. “What is your estimand? Defining the target quantity connects statistical evidence to theory.” *American Sociological Review* 86(3):532–565.
- Lundberg, Shelly and Elaina Rose. 2002. “The effects of sons and daughters on men’s labor supply and wages.” *Review of Economics and Statistics* 84(2):251–268.
- Lundberg, Shelly and Elaina Rose. 2003. “Child gender and the transition to marriage.” *Demography* 40(2):333–349.
- Marshall, John. 2016. “Coarsening bias: How coarse treatment measurement upwardly biases instrumental variable estimates.” *Political Analysis* 24(2):157–171.
- Martin, John Levi. 2017. *Thinking through methods: A social science primer*. University of Chicago Press.
- Massey, Douglas S, Karen A Pren and Jorge Durand. 2016. “Why border enforcement backfired.” *American journal of sociology* 121(5):1557–1600.

- McClellan, Mark, Barbara J McNeil and Joseph P Newhouse. 1994. “Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality?: analysis using instrumental variables.” *Jama* 272(11):859–866.
- Mellon, Jonathan. 2021. Rain, Rain, Go Away: 176 potential exclusion-restriction violations for studies using weather as an instrumental variable. Working paper.
- 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):3663–98.
- Morgan, Stephen L and Christopher Winship. 2015. *Counterfactuals and causal inference*. Cambridge University Press.
- Nunn, Nathan and Leonard Wantchekon. 2011. “The slave trade and the origins of mistrust in Africa.” *American Economic Review* 101(7):3221–52.
- Ogburn, Elizabeth L, Andrea Rotnitzky and James M Robins. 2015. “Doubly robust estimation of the local average treatment effect curve.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77(2):373–396.
- Olea, José Luis Montiel and Carolin Pflueger. 2013. “A robust test for weak instruments.” *Journal of Business & Economic Statistics* 31(3):358–369.
- Pearl, Judea. 1995. “Causal diagrams for empirical research.” *Biometrika* 82(4):669–688.
- Pearl, Judea. 2009. *Causality*. Cambridge university press.
- Pearl, Judea. 2013. “Linear models: A useful “microscope” for causal analysis.” *Journal of Causal Inference* 1(1):155–170.
- Richardson, Thomas S and James M Robins. 2014. “ACE bounds; SEMs with equilibrium conditions.” *Statistical Science* 29(3):363–366.
- Rossi, Peter E. 2014. “Even the rich can make themselves poor: A critical examination of IV methods in marketing applications.” *Marketing Science* 33(5):655–672.
- Rothwell, Jonathan T and Douglas S Massey. 2010. “Density zoning and class segregation in US metropolitan areas.” *Social science quarterly* 91(5):1123–1143.
- Rubin, Donald B. 1974. “Estimating causal effects of treatments in randomized and nonrandomized studies.” *Journal of educational Psychology* 66(5):688.
- Sampson, Robert J. and Alix S. Winter. 2018. “Poisoned development: Assessing childhood lead exposure as a cause of crime in a birth cohort followed through adolescence.” *Criminology* 56(2):269–301.

- Sävje, Fredrik. 2021. “Randomization does not imply unconfoundedness.” *arXiv preprint arXiv:2107.14197*.
- Sharkey, Patrick, Gerard Torrats-Espinosa and Delaram Takyar. 2017. “Community and the crime decline: The causal effect of local nonprofits on violent crime.” *American Sociological Review* 82(6):1214–1240.
- Słoczyński, Tymon. 2020. “When Should We (Not) Interpret Linear IV Estimands as LATE?” *arXiv preprint arXiv:2011.06695*.
- Słoczyński, Tymon, S Derya Uysal and Jeffrey M Wooldridge. 2022. “Abadie’s Kappa and Weighting Estimators of the Local Average Treatment Effect.” *arXiv preprint arXiv:2204.07672*.
- 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.
- Stock, James H and Motohiro Yogo. 2005. “Testing for Weak Instruments in Linear IV Regression.” *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg* p. 80.
- Swanson, Sonja A, Matthew Miller, James M Robins and Miguel A Hernán. 2015. “Definition and evaluation of the monotonicity condition for preference-based instruments.” *Epidemiology (Cambridge, Mass.)* 26(3):414.
- Swanson, Sonja A and Miguel A Hernán. 2018. “The challenging interpretation of instrumental variable estimates under monotonicity.” *International journal of epidemiology* 47(4):1289–1297.
- Swanson, Sonja A, Miguel A Hernán, Matthew Miller, James M Robins and Thomas S Richardson. 2018. “Partial identification of the average treatment effect using instrumental variables: review of methods for binary instruments, treatments, and outcomes.” *Journal of the American Statistical Association* 113(522):933–947.
- Tan, Zhiqiang. 2006. “Regression and weighting methods for causal inference using instrumental variables.” *Journal of the American Statistical Association* 101(476):1607–1618.
- Van Kippersluis, Hans and Cornelius A Rietveld. 2018. “Beyond plausibly exogenous.” *The Econometrics Journal* 21(3):316–331.
- VanderWeele, Tyler J. 2009. “Concerning the consistency assumption in causal inference.” *Epidemiology* 20(6):880–883.
- VanderWeele, Tyler J and Ilya Shpitser. 2011. “A new criterion for confounder selection.” *Biometrics* 67(4):1406–1413.

- Wang, Xuran, Yang Jiang, Nancy R Zhang and Dylan S Small. 2018. “Sensitivity analysis and power for instrumental variable studies.” *Biometrics* 74(4):1150–1160.
- Wilmers, Nathan. 2018. “Wage stagnation and buyer power: How buyer-supplier relations affect US workers’ wages, 1978 to 2014.” *American Sociological Review* 83(2):213–242.
- Wooldridge, Jeffrey M. 2010. *Econometric analysis of cross section and panel data*. MIT press.
- Wright, Philip G. 1928. *Tariff on animal and vegetable oils*. Macmillan Company, New York.
- Young, Alwyn. 2019. Consistency without inference: Instrumental variables in practical application. Working paper.
- Zeng, Shuxi, Fan Li and Peng Ding. 2020. “Is being an only child harmful to psychological health?: evidence from an instrumental variable analysis of China’s one-child policy.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 183(4):1615–1635.

A Published IV Papers in Sociology

We initially collected a sample of papers in 2019, limited our search to the prior 15 years. We searched for words “instrument,” “instruments,” and “instrumental” in the *American Sociological Review* and the *American Journal of Sociology*. In May of 2022, we expanded the sample to include more recent papers.

We included only papers that reported an IV analysis in the main text or appendix of the paper. We excluded papers reporting Heckman selection models or dynamic GMM models, which require different assumptions and target different estimands. We also excluded papers using instrumental variables to correct for measurement error rather than for causal inference. We excluded one paper that stated their results were robust to an instrumental variables specification but did not report the actual specification. Our final sample consisted of 34 papers. Table A.1 and Table A.2 describe features of these papers.

When coding whether or not authors stated assumptions, we assessed whether they were stated in *causal* terms. A paper that merely stated the instrument must be uncorrelated with an error term, but didn’t clarify this assumption in causal terms, was coded as having not stated the unconfoundedness or exclusion restriction assumptions. As noted in the main text, the assumption that the instrument is uncorrelated with the error term in the “structural” model (i.e., the regression of the outcome on the uninstrumented treatment) captures both the exclusion restriction and unconfoundedness assumption, and we find it helpful to separate them.

For the relevance assumption, many authors described the assumption itself in terms of correlation but proceeded to explain the correlation in terms of the causal effect of the instrument on the treatment. These papers were coded as having stated the assumption in causal terms. If a paper used a proxy instrument, and the authors described the correlation between the proxy and treatment in terms of an unmeasured common cause, the paper was also coded as having stated the assumption in causal terms. One paper explained the correlation between the instrument and treatment in terms of the treatment’s effect on the instrument. This paper was coded as having not stated the relevance assumption, because this “instrument” is simply not an instrument.

To calculate $|\widehat{2SLS}/\widehat{OLS}|$ for each paper, we used the following procedure. First, we located matching 2SLS and OLS regressions—i.e., a 2SLS and OLS regression that used the same sample and covariates. For each matched pair, we took the absolute value of each estimated treatment effect and then calculated the ratio. When a paper reported multiple pairs of matching regressions, we took the average ratio across all regression pairs.

Authors	Year	Journal	States Monotonicity	States Uncon-foundedness	States Exclusion Restriction	States the Estimand (LATE)	Includes Post-Instrument Covariates
Braun	2022	<i>ASR</i>	✗	✗	✓	✗	Definitely
Lin & Hung	2022	<i>AJS</i>	✗	✗	✓	✗	No
Muller & Schrage	2021	<i>AJS</i>	✗	✓	✓	✗	No
VanHeuvelen	2020	<i>AJS</i>	✗	✗	✗	✗	Potentially
de Vaan & Stuart	2019	<i>ASR</i>	✗	✓	✓	✓	No
Light & Thomas	2019	<i>ASR</i>	✗	✗	✓	✓	Definitely
Wang et al.	2019	<i>ASR</i>	✗	✗	✓	✓	No
Muller	2018	<i>AJS</i>	✗	✗	✗	✗	Potentially
Wilmers	2018	<i>ASR</i>	✗	✓	✗	✓	Potentially
Harding et al.	2018	<i>AJS</i>	✓	✓	✓	✓	Definitely
Pernell et al.	2017	<i>ASR</i>	✗	✗	✓	✗	No
Sharkey et al.	2017	<i>ASR</i>	✗	✓	✓	✓	Potentially
Samila & Sorenson	2017	<i>ASR</i>	✗	✗	✗	✗	Potentially
Massey et al.	2016	<i>AJS</i>	✗	✗	✗	✗	No
Hagan et al.	2016	<i>ASR</i>	✗	✗	✓	✗	Potentially
Cole	2015	<i>ASR</i>	✗	✗	✗	✗	Definitely
Gangl & Ziefle	2015	<i>AJS</i>	✗	✗	✓	✓	No
Polavieja	2015	<i>ASR</i>	✗	✓	✓	✗	Definitely
Heaney & Rojas	2014	<i>AJS</i>	✗	✗	✓	✗	Definitely
Pais	2013	<i>AJS</i>	✗	✗	✓	✗	Potentially
Akchurin & Lee	2013	<i>ASR</i>	✗	✗	✗	✗	Potentially
Cole & Ramirez	2013	<i>ASR</i>	✗	✗	✗	✗	Definitely
Lyons et al.	2013	<i>ASR</i>	✗	✗	✗	✗	Definitely
Muller	2012	<i>AJS</i>	✗	✗	✓	✗	Potentially
Cole	2012	<i>AJS</i>	✗	✗	✗	✗	Definitely
Vasi & King	2012	<i>ASR</i>	✗	✗	✗	✗	Potentially
Rugh & Massey	2010	<i>ASR</i>	✗	✗	✓	✗	Potentially
Luke	2010	<i>AJS</i>	✗	✗	✓	✗	Potentially
Bandelj	2009	<i>ASR</i>	✗	✗	✗	✗	Definitely
Kirk	2009	<i>ASR</i>	✗	✗	✓	✓	Definitely
Lizardo	2006	<i>ASR</i>	✗	✗	✓	✗	Definitely
Ingram et al.	2005	<i>AJS</i>	✗	✗	✗	✗	Potentially
Jong-sung & Khagram	2005	<i>ASR</i>	✗	✗	✓	✗	Potentially
Burris	2004	<i>ASR</i>	✗	✓	✓	✗	Potentially

Table A.1: Table contains all *ASR* and *AJS* papers published between 2004–2022 meeting criteria listed above. Stating an assumption in substantive terms requires specifying what causal relations are not permitted (see Section 3).

Authors	Year	Journal	Reports First-Stage Effect Size	Reports Any First-Stage F -statistic	$\left \frac{\widehat{2SLS}}{\widehat{OLS}} \right $
Braun	2022	<i>ASR</i>	✓	✓	NA
Lin and Hung	2022	<i>AJS</i>	✓	✗	NA
Muller and Schrage	2021	<i>AJS</i>	✓	✓	NA
VanHeuvelen	2020	<i>AJS</i>	✗	✗	3.55
de Vaan & Stuart	2019	<i>ASR</i>	✓	✗	4.36
Light & Thomas	2019	<i>ASR</i>	✓	✓	3.64
Wang et al.	2019	<i>ASR</i>	✓	✗	NA
Muller	2018	<i>AJS</i>	✓	✓	NA
Wilmers	2018	<i>ASR</i>	✓	✓	3.66
Harding et al.	2018	<i>AJS</i>	✗	✓	10.83
Pernell et al.	2017	<i>ASR</i>	✗	✗	NA
Sharkey et al.	2017	<i>ASR</i>	✓	✓	1.56
Samila & Sorenson	2017	<i>ASR</i>	✓	✓	6.77
Massey et al.	2016	<i>AJS</i>	✗	✗	NA
Hagan et al.	2016	<i>ASR</i>	✗	✗	NA
Cole	2015	<i>ASR</i>	✗	✓	NA
Gangl & Ziefle	2015	<i>AJS</i>	✓	✓	10.56
Polavieja	2015	<i>ASR</i>	✓	✓	NA
Heaney & Rojas	2014	<i>AJS</i>	✗	✓	NA
Pais	2013	<i>AJS</i>	✗	✓	1.91
Akchurin & Lee	2013	<i>ASR</i>	✗	✗	NA
Cole & Ramirez	2013	<i>ASR</i>	✗	✗	NA
Lyons et al.	2013	<i>ASR</i>	✗	✗	NA
Muller	2012	<i>AJS</i>	✓	✓	0.93
Cole	2012	<i>AJS</i>	✗	✗	NA
Vasi & King	2012	<i>ASR</i>	✗	✓	7.90
Rugh & Massey	2010	<i>ASR</i>	✗	✓	1.30
Luke	2010	<i>AJS</i>	✓	✗	5.11
Bandelj	2009	<i>ASR</i>	✗	✗	4.64
Kirk	2009	<i>ASR</i>	✓	✓	NA
Lizardo	2006	<i>ASR</i>	✗	✗	NA
Ingram et al.	2005	<i>AJS</i>	✗	✗	0.75
Jong-sung & Khagram	2005	<i>ASR</i>	✓	✓	3.13
Burris	2004	<i>ASR</i>	✗	✗	1.13

Table A.2: Table contains all *ASR* and *AJS* papers published between 2004–2022 meeting the criteria above. Most papers did not specify whether the reported F -test was robust to heteroskedasticity. The final column only considers coefficients on the treatment, excluding covariate–treatment interaction terms. When multiple OLS and 2SLS specifications were reported, ratios were calculated for each OLS–2SLS pair and then averaged. NAs occur either because authors reported a 2SLS model but no OLS model, or because authors used, e.g., an IV probit model rather than 2SLS.

B Proxy Instruments

Sometimes, researchers use instruments that have no causal effect on the treatment but instead share a common unmeasured cause with the treatment. We call these proxy, surrogate, or mismeasured instruments (Hernán and Robins, 2006; Chalak, 2017). While we can use proxy instruments to estimate treatment effects, identification assumptions must hold with respect to the underlying causal instrument, a nuance not well understood in the applied literature. Furthermore, compliers are defined with respect to the causal instrument.

This section proceeds as follows. First, we introduce several example proxy instruments. Next, we review the assumptions and causal estimand. We conclude by summing up the challenges of proxy instruments.

B.1 Example Proxy Instruments

B.1.1 Sharkey, Torratts-Espinosa, and Takyar (2017): Do Community Nonprofits Reduce Crime?

Sharkey, Torratts-Espinosa and Takyar (2017) are interested in estimating the effect of nonprofit organizations focused on reducing violence and crime (so-called “community nonprofits”) on crime rates in American cities. In one of their analyses to address concerns about unmeasured confounding, they use the change in the number of nonprofits focusing on the arts, medical research, and environmental protection as an instrument for the change in the number of community nonprofits. The authors are clear that the instrument and treatment are related through unmeasured confounding: “[Changes] in the prevalence of nonprofits that have nothing to do with crime and violence are associated with changes in the prevalence of nonprofits related to crime and violence through common mechanisms of funding availability” (1225). They also note that their instrument was inspired by Levitt (2002), who similarly argues that his instrument and treatment are related through unmeasured common causes.

The change in the number of “other” nonprofits represents our *proxy* or *surrogate* instrument, while the unmeasured common funding availability is our *causal* instrument. Note that some funding might be specifically reserved for community nonprofits, and other funding might be specifically reserved for other nonprofits. The causal instrument does not include the funding availability specific to one type of nonprofit; it includes only the funding availability common to *both* categories of nonprofits. In the rest of this section, we simply refer to the causal instrument as “funding availability,” but we emphasize that we are specifically describing funding availability common to both types of nonprofits. We depict this causal structure in the DAG in Figure B.1, omitting control variables and potential violations of identification assumptions.

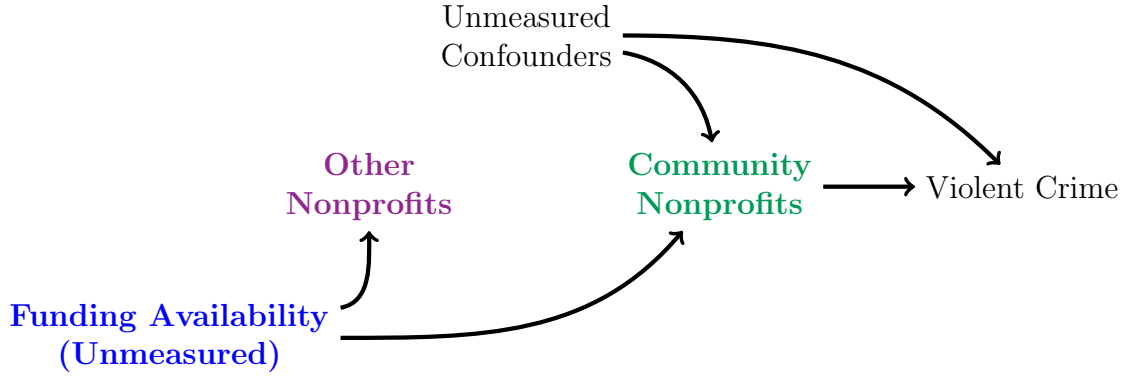


Figure B.1: DAG depicting the causal structure in Sharkey, Torrats-Espinosa and Takyar (2017), omitting potential violations of identification assumptions. *Other Nonprofits* is the **proxy instrument**, and *Funding Availability* is the **causal instrument**.

B.1.2 Massey, Pren, and Durand (2016): Does Border Enforcement Reduce Undocumented Migration to the U.S.?

Massey, Pren and Durand (2016) seek to estimate the causal effect of increasing the border patrol budget on undocumented migration. Following Angelucci (2012), they instrument the annual Border Patrol budget with the annual budget for the Drug Enforcement Administration (DEA). According to Angelucci, who uses the DEA budget to instrument border line watch hours, “[changes] in US taste for drugs (e.g., a tougher “War on Drugs”) simultaneously increase the DEA budget and border line watch hours” (328). The logic of using the DEA budget as an instrument is that it shares an unmeasured common cause—US taste for drugs—with the treatment. We depict this causal structure in the DAG in Figure B.2, omitting control variables and potential violations of identification assumptions.

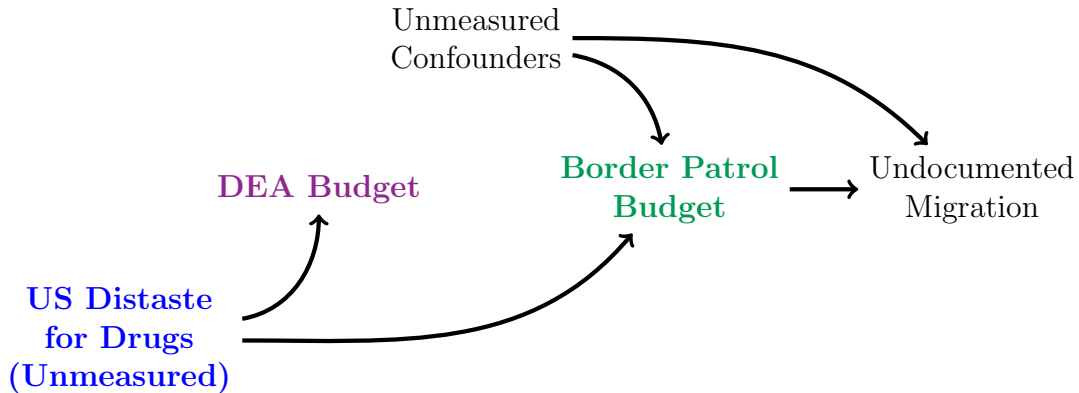


Figure B.2: DAG depicting the causal structure in Massey, Pren and Durand (2016), omitting potential violations of identification assumptions. *DEA Budget* is the proxy instrument, and *US Distaste for Drugs* is the causal instrument.

B.1.3 Currie, Jin, and Schnell (2019): Does the Opioid Prescription Rate Reduce Employment?

Currie, Jin and Schnell (2019) are interested in estimating the effect of the local opioid prescription rate among working-age people on the local employment-to-population ratio. The authors employ the opioid prescription rate among elderly people as an instrument. The logic behind the instrument is that “doctors are more likely to prescribe opioids to everyone in some places than in others (Schnell, 2017), so that places where elderly people are more likely to get prescriptions are places where working-age people are also more likely to get them” (13). In other words, there appears to be some underlying place-specific propensity to prescribe opioids that affect both the elderly prescription rate and the working-age prescription rate. We depict this causal structure in the DAG in Figure B.3, omitting control variables and potential violations of identification assumptions.

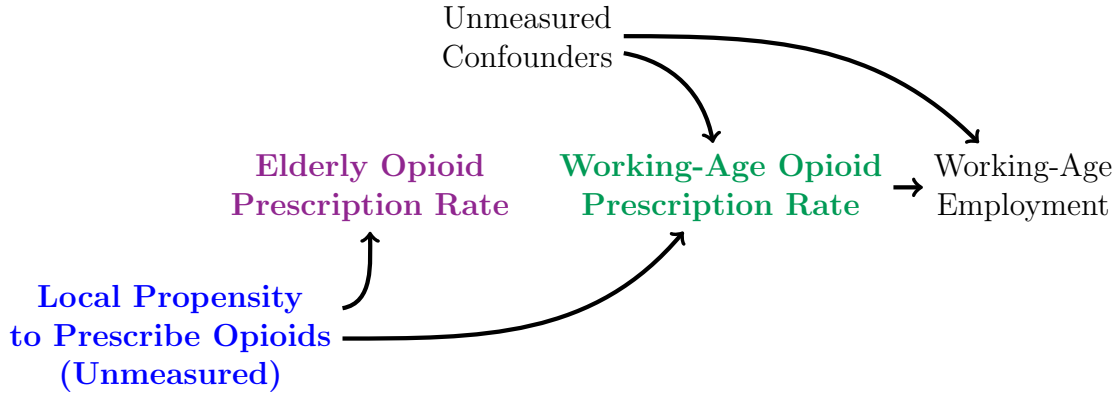


Figure B.3: DAG depicting the causal structure in Currie, Jin and Schnell (2019), omitting potential violations of identification assumptions. *Elderly Opioid Prescription Rate* is the proxy instrument, and *Local Propensity to Prescribe Opioids* is the causal instrument.

B.2 Causal Estimand and Identification Assumptions

In the simple case with a binary causal instrument, binary treatment, and no covariates, we can identify the Local Average Treatment Effect (LATE).³¹ Crucially, however, compliance is defined with respect to the *causal* instrument, not the proxy instrument. That is, compliers are those whose treatment status is affected by the causal instrument. In Sharkey, Torrats-Espinosa and Takyar (2017), for example, compliers are cities in which the number of community nonprofits is affected by funding availability.

To identify the treatment effect, all of our standard IV assumptions must hold with respect to the *causal* instrument. Consider again Sharkey, Torrats-Espinosa and Takyar

³¹When our causal instrument or treatment is multi-valued or continuous, 2SLS instead recovers a weighted average of LATEs. Chalak (2017) provides conditions that ensure weights will be positive.

(2017), where the causal instrument is funding availability. In this setting, the exclusion restriction holds that funding mechanisms have no effect on crime rates except through community nonprofits, and unconfoundedness holds that funding mechanisms share no common causes with crime rates.³² We depict potential violations of these assumptions in Figure B.4.

Second, we assume that the proxy instrument is unassociated with the treatment and outcome *conditional* on the unobserved causal instrument.³³ In other words, if we were somehow able to measure the causal instrument, the proxy should provide us with no information about the treatment or outcome beyond what the causal instrument provides us. More formally, the proxy instrument should be conditionally independent of both the treatment and outcome given the causal instrument: $Z^* \perp\!\!\!\perp D, Y \mid Z$, where Z is the unmeasured causal instrument and Z^* is the observed proxy instrument. The conditional independence assumption requires that our proxy instrument (i) is unconfounded with the outcome, (ii) has no effects on the treatment or the outcome, and (iii) is not itself affected by the treatment and outcome. We depict potential violations of this assumption in Figure B.5.

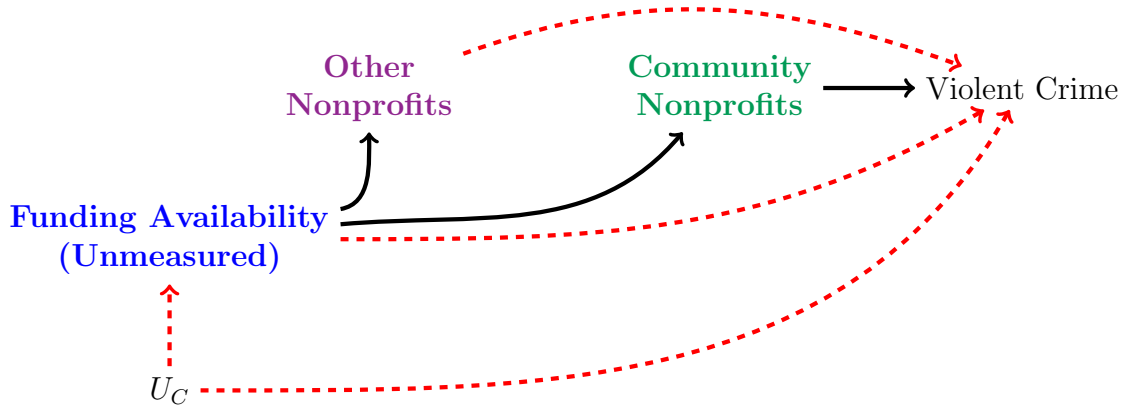


Figure B.4: DAG depicting the causal structure in Sharkey, Torrats-Espinosa and Takyar (2017). Red, dashed arrows encode potential violations of the exclusion restriction and unconfoundedness for the causal instrument. *Funding Availability* \rightarrow *Other Nonprofits* \rightarrow *Violent Crime* and *Funding Availability* \rightarrow *Violent Crime* depict exclusion restriction violations, and *Funding Availability* $\leftarrow U_C \rightarrow$ *Violent Crime* represents a violation of unconfoundedness.

³²In one specification, the authors include city and year fixed effects, which weakens the unconfoundedness assumption but introduces further restrictions on effect heterogeneity (De Chaisemartin and d'Haultfoeuille, 2020; Imai and Kim, 2021).

³³If we think of the proxy instrument as a mismeasured instrument, this assumption can be viewed as restricting the measurement error to be non-differential.

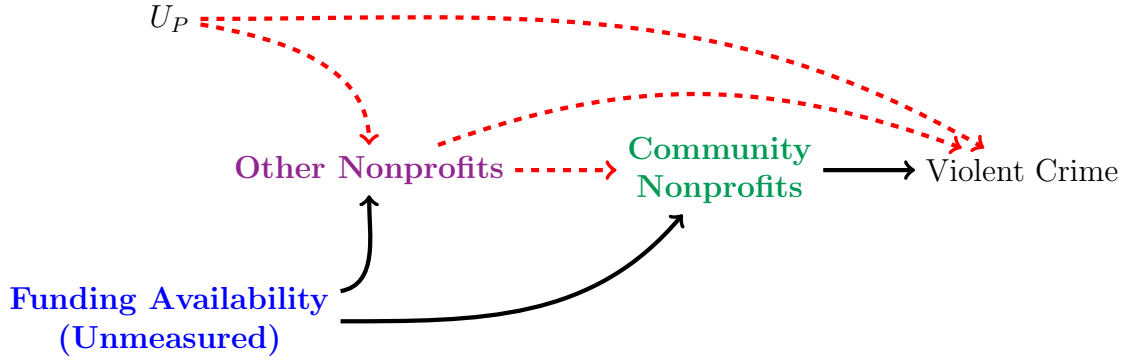


Figure B.5: DAG depicting the causal structure in Sharkey, Torratts-Espinosa and Takyar (2017). Red, dashed arrows indicate violations of the conditional independence assumption. This assumption requires that the proxy instrument gives us no information about the treatment or outcome conditional on the causal instrument.

B.3 The Challenges of Proxy Instruments

Proxy instruments require more complex assumptions to identify the treatment effect. The unconfoundedness and exclusion restriction assumptions are made with respect to the underlying causal instrument rather than the proxy instrument. This fact makes it essential to clearly define the causal instrument and study its causes and consequences—that is, to understand the relevant DAG. Proxy instruments thus create a natural tension: we often resort to proxy instruments precisely because the causal instrument is not well understood, *but we have to understand the causal instrument for identification to be credible*.

Sharkey, Torratts-Espinosa and Takyar (2017) are clear that their measured instrument is a proxy, but part of why they can't measure the causal instrument (funding availability) is that we don't know enough about the common funding sources for community nonprofits and other nonprofits. We would have to study not all types of funding for nonprofits, but specifically the funding availability that is common to both community nonprofits and other nonprofits to be able to evaluate the relevant identification assumptions.

A case where we might be particularly concerned is the setting where the proxy instrument and the treatment are *very* highly associated. In Massey, Pren and Durand (2016), the authors report an R^2 of 0.97 from a regression of the treatment (logged Border Patrol budget) on the proxy instrument (DEA budget) with no additional covariates. Similarly, Currie, Jin and Schnell (2019) report first-stage correlation coefficients ranging from 0.721 to 0.934 depending on age and gender. If these studies both use valid instruments, it would suggest that the *vast* majority of the variation in the two treatments is unconfounded and explained by a single underlying causal instrument. This implication sits uneasily with the view that treatments in sociology are confounded and require research designs that move beyond selection on observables.

C The Limits of Testing Identification Assumptions

We can never know with certainty whether our identification assumptions hold. Nonetheless, a number of tests can boost our confidence in various assumptions by evaluating whether the data is consistent with our expectations. We review some of these tests, discuss how to properly conduct them, and clarify their limitations.

C.1 Placebo Tests for IV

Placebo tests can help us assess both the unconfoundedness and exclusion restriction assumptions. Researchers have at least three types of placebo tests at their disposal: the first using a placebo population, the second using a placebo instrument, and the third using a placebo outcome. We consider each in turn.

The first involves locating a “placebo” population in which the instrument has no effect on the treatment. This condition can be assessed empirically by examining the instrument–treatment association in the placebo population. Under unconfoundedness and the exclusion restriction, the instrument should be unassociated with the outcome in the placebo population. An association provides evidence of confounding or an exclusion restriction violation. This test thus jointly tests for the presence of bias from unmeasured confounding and bias from exclusion restriction violations.

Crucially, this test is most informative when the bias in the placebo population is equal to the bias in the target population. For instance, if theory predicts that the exclusion restriction violation in the placebo population will be much weaker than the exclusion restriction violation in the target population, the placebo test will do little to assuage our concerns about the latter. Examples of this type of placebo test can be found in Nunn and Wantchekon (2011) and Angrist and Pischke (2008). We depict the causal structure for this type of placebo test in Figure C.1.

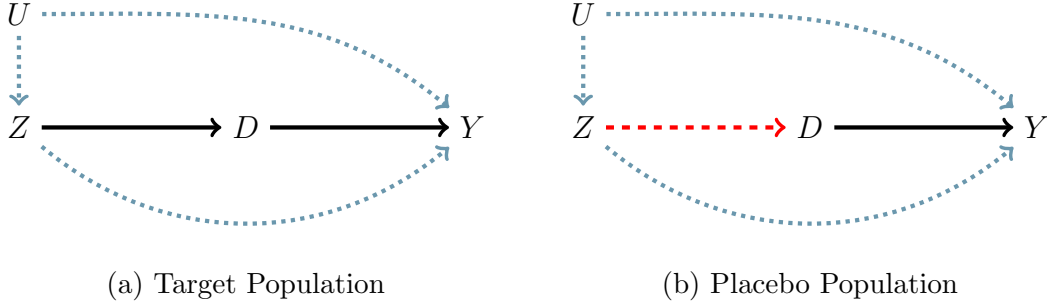


Figure C.1: DAGs depicting the causal structures for the population placebo test. In the placebo population, Z should have no effect on D —the dashed, red edge $Z \rightarrow D$ in Figure C.1b should not exist. The test will be most informative when the bias in the target population is equal to the bias in the placebo population. By examining the association between Z and Y in the placebo population, we jointly test for the presence of all three dotted, blue-gray arrows in the placebo population. If we think the magnitude of bias in the two populations should be equal, the lack of an association between Z and Y in the placebo population will make us more confident that the identification assumptions hold in the target population.

It is worth clarifying an important detail about this test. The analyst should test for an association between the instrument and outcome in the placebo population, possibly conditional on a set of measured confounders. The analyst should *not* run a 2SLS model in the placebo population. Because the instrument has no effect on the treatment, estimation bias will be severe and point estimates could be extremely noisy.

Alternatively, we could identify a placebo instrument, Z^* , that has no effect on the treatment. This assumption can be assessed empirically. If we think this placebo instrument is subject to the same unmeasured confounding and exclusion restriction violations as the actual instrument, examining the Z^* – Y relationship can inform us about the validity of these assumptions. Like the prior test, this test is most informative when the biases are similar in magnitude.

Again, the analyst should *not* run a 2SLS model using the placebo instrument. If the instrument has no effect on the treatment, estimation bias will be severe and point estimates will be extremely noisy. An example of this placebo test can be found in Wilmers (2018).³⁴ We illustrate the causal structure of this test in Figure C.2.

³⁴Unfortunately, Wilmers (2018) makes the mistake of running a 2SLS model using the placebo instrument. The point estimates are extremely noisy, with confidence intervals roughly four times longer than those for the regular IV specification (Table 3 in Wilmers (2018)). It is also worth noting that the first-stage F -statistic in the placebo 2SLS is 49.36, indicating the placebo instrument may in fact affect the treatment.

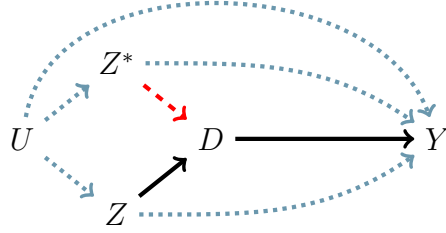


Figure C.2: DAG depicting the causal structure of the placebo-instrument test, where Z represents the actual instrument and Z^* represents the pseudo-instrument. The test requires that the red, dashed edge $Z^* \rightarrow D$ does not exist. The test will be most informative when the magnitude of bias for the placebo instrument is equal to the magnitude of bias for the actual instrument. By testing for an association between Z^* and Y , we jointly test for the dotted, blue-gray arrows $U \rightarrow Y$, $U \rightarrow Z^*$, $U \rightarrow Z$, $Z \rightarrow Y$, and $Z^* \rightarrow Y$.

Finally, we can locate a placebo outcome that cannot be plausibly caused by the treatment. As always, the test will be most informative when the bias from confounding or exclusion restriction violations for the placebo outcome are equal to the bias for the actual outcome. If the instrument shares no association with the placebo outcome, that provides some evidence that the identification assumptions may be valid. Laitin and Ramachandran (2016) provides an example of such a placebo test.

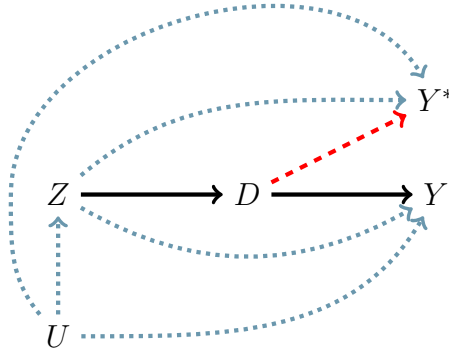


Figure C.3: DAG depicting the causal structure of the placebo outcome test. The test requires the dashed, red edge $D \rightarrow Y^*$ does not exist—the treatment has no effect on the placebo outcome. The test will be most informative when the bias produced by the arrows $Z \rightarrow Y$ and $U \rightarrow Y$ is equal to the bias produced by the arrows $Z \rightarrow Y^*$ and $U \rightarrow Y^*$. By examining the association between Z and Y^* , we jointly test for the presence of the blue, dotted arrows $U \rightarrow Z$, $U \rightarrow Y$, $U \rightarrow Y^*$, $Z \rightarrow Y$, and $Z \rightarrow Y^*$.

Before discussing what we see as the core challenges facing placebo tests with IV,

we highlight one tempting but ultimately misleading “test” of an instrument’s validity. Researchers might be inclined to test for an association between the instrument and the outcome conditional on the treatment. The thinking behind such a test is that any remaining association between the instrument and outcome will be due to unconfoundedness or exclusion restriction violations. But this is untrue. The treatment is a *collider* between the instrument and the unmeasured treatment–outcome confounders (that is, the treatment is a consequence of both the instrument and confounders). Conditioning on the treatment will therefore induce an association between the instrument and these confounders (Elwert and Winship, 2014).

C.1.1 Challenges of Placebo Tests

Placebo tests can be challenging for two main reasons. First, it can be difficult to find suitable placebo populations, instruments, or outcomes. For our placebo test to be credible, we must have strong reasons for expecting that placebo instruments and outcomes are subject to the same confounding (or exclusion restriction violations) as the primary instruments and outcomes.

The second shortcoming of placebo tests is that it is sometimes unclear whether there is “no association” between the instrument and outcome. Common practice is to consider the placebo test “passed” if the association between the instrument and outcome is statistically insignificant at the $\alpha = 0.05$ level. But because IV estimates are so sensitive to violations of unconfoundedness and the exclusion restriction, placebo tests might be underpowered for detecting associations that would be sufficient to explain away a study’s findings.

One way to tackle this challenge is to use bias analysis and placebo tests in tandem, as in Van Kippersluis and Rietveld (2018) and Lal et al. (2021). This approach involves using the resulting association between the instrument and outcome in a placebo population as the input to a bias analysis. We also recommend that researchers plot the primary treatment effect estimates and placebo effect estimates side-by-side with 95% confidence intervals to more clearly convey uncertainty about the results of the placebo test.

C.2 Why Balance Checks Need to Be Scaled

One way to assess the unconfoundedness assumption is to compare units that receive the instrument with units that do not across potential instrument–outcome confounders. If the two sets of units show no differences, we can be more confident in the unconfoundedness assumption.

As Jackson and Swanson (2015) note, however, comparing the raw differences in covariate values can be misleading, since the bias is scaled by strength of the first-stage effect. Jackson and Swanson (2015) recommend dividing the differences in covariate values by the estimated first-stage effect to more clearly convey the potential for confounding bias.

Finally, we recommend against reporting the statistical significance of the scaled differences. Statistically insignificant differences are typically interpreted as support for the null hypothesis of no difference. But it is often more plausible that we simply lack the statistical power to “detect” the difference. Furthermore, as Ho et al. (2007) emphasize, reducing imbalance (by conditioning on confounders) will often reduce the variance of our estimator, regardless of whether the in-sample imbalances reflect a systematic difference in the target population.

D Software and Replication Code

D.1 Software Packages

Table D.1 provides the names of software packages and commands that can be used to carry out procedures recommended in the main text.

Procedure	R	Stata
Robust F -test	<code>waldtest</code> function in <code>lmtest</code> package	<code>weakivtest</code>
Olea and Pflueger effective F -test		<code>weakivtest</code> command
Heteroskedastic-robust SEs	<code>robust.se</code> command in <code>ivpack</code> package	<code>robust</code> option in <code>ivreg2</code> command
Cluster-robust SEs	<code>cluster.robust.se</code> command in <code>ivpack</code> package	<code>cluster</code> option in <code>ivreg2</code> command
Anderson–Rubin CIs (homoskedastic)	<code>anderson.rubin.ci</code> command in <code>ivpack</code> package	<code>ar</code> option in <code>condivreg</code> package
Anderson–Rubin CIs (heteroskedastic)		<code>weakiv</code> package
Bootstrapped CIs	<code>bootstrap</code> package	<code>bootstrap</code> command
Expected Type-M Error	<code>retrodesign</code> package	<code>retrodesign</code> package
Cinelli and Hazlett Sensitivity Analysis	<code>sensemakr</code> package	<code>sensemakr</code> package
Sensitivity Plot	See replication code at https://github.com/cmfelton/iv_checklist	
OLS vs. 2SLS Plot	See replication code at https://github.com/cmfelton/iv_checklist	

Table D.1: Software required for various IV tools discussed in the paper.

D.2 Checklist R Code for Conley and Glauber (2006)

For example R code showing how to employ the packages and functions listed above, visit https://github.com/cmfelton/iv_checklist.

E 2SLS Estimands

In the setting with a binary instrument, binary treatment, and no additional covariates, the 2SLS estimand can be interpreted as a Local Average Treatment Effect (LATE) under the identification assumptions discussed in the text. However, when (i) either the instrument or treatment are not binary or (ii) additional covariates are included in the model, the interpretation of the 2SLS estimand becomes more complicated, unless we impose further restrictions on effect heterogeneity. Below, we describe the interpretation of the estimand in different settings and then briefly discuss alternative estimands and estimation strategies.

E.1 Added Covariates

First, consider the setting with a binary treatment, binary instrument, and added covariates. The covariate-specific LATE is the average treatment effect for compliers at particular values of the covariates. To fix ideas, consider the Conley and Glauber (2006) study on the effects of sibship size. Suppose we estimate the treatment effect of sibship size using a 2SLS regression with a single additional binary covariate—an indicator for whether the mother obtained a college degree. In this situation, there are two covariate-specific LATEs: the LATE for families with college-educated mothers and the LATE for families with mothers who have not completed college. If these two sets of compliers have different average treatment effects, these covariate-specific LATEs will differ. For the case with a binary treatment and instrument, we can define a covariate-specific LATE as

$$\text{LATE}(\mathbf{x}) = E[Y_i(1) - Y_i(0) \mid D_i(1) - D_i(0) = 1, \mathbf{X}_i = \mathbf{x}],$$

where \mathbf{X} is possibly multi-dimensional.

Angrist and Imbens (1995) showed that, when covariates are added to the 2SLS model, the 2SLS estimand does not, in general, equal the LATE. They show that in a 2SLS model with a fully saturated first stage and discrete covariates, the estimand converges to a weighted average of covariate-specific LATEs, where weights are a function of how much variance the instrument induces in the treatment at particular values of the covariates.³⁵ For instance, if the instrument has a stronger effect on the treatment for families with college-educated mothers, then their LATE will be upweighted relative to the LATE for families with mothers without college degrees.

In practice, some of covariates may be continuous, and we may not have a large enough sample to precisely estimate a fully saturated model. Evdokimov and Kolesár (2019) generalizes the result to the case without a fully saturated model under the assumption that the first-stage and reduced-form models are correctly specified. Słoczyński (2020) generalize the Angrist and Imbens (1995) result under the assumption that the probability of receiving the instrument is linear in covariates \mathbf{X} .

³⁵This parallels a result for OLS (Angrist and Krueger, 1999; Aronow and Samii, 2016).

If we suspect that effect heterogeneity across covariates is small, this result will be of little consequence. Słoczyński (2020) offers a more critical perspective on the 2SLS estimand, showing that when the probability of receiving the instrument is small—that is, when few units are encouraged by the instrument to get treated—the LATE for *treated* units receives more weight.

E.2 Multi-Valued Treatments

Next, consider the case with a single binary instrument and a continuous or multi-valued treatment. The per-unit LATE is the effect of shifting the treatment by one unit for a particular subset of compliers. Suppose our treatment is years of schooling, and recall that we define a causal effect as a contrast between two potential outcomes. We can define a treatment effect for receiving 9 years of schooling rather than 8 years of schooling, a treatment effect for receiving 10 years of schooling rather than 9 years, and so on. For the case with a binary instrument, we can define a per-unit LATE as

$$\text{LATE}(j) = E[Y_i(j) - Y_i(j-1) \mid D_i(1) \geq j > D_i(0)],$$

where j denotes a particular value of the treatment. Note that the per-unit LATE applies not to *all* compliers, but to compliers at a particular level of the treatment. Smith is a complier at j if, in the absence of the instrument, he takes a value of the treatment less than j , but in the presence of the instrument, takes values of the treatment of at least j . For instance, if the instrument shifts Smith’s treatment from 11 years of schooling to 12, he is a complier at $j = 12$ but not at $j = 9$. Angrist and Imbens (1995) showed that the 2SLS estimand converges to a weighted average of per-unit LATEs, with weights given by

$$\omega_j = \frac{P(D(1) \geq j > D(0))}{\sum_{i=1}^J P(D(1) \geq i > D(0))}.$$

The weights are non-negative and sum to 1. The formula for the weights reveals that treatment effects for units who are more responsive to the instrument will be up-weighted. The numerator captures the proportion of units who are compliers at j . The more compliers we have at a particular value of the treatment, the larger weights their treatment effects receive. Furthermore, compliers for whom the instrument causes a multi-unit increase in the level of the treatment will see their effects upweighted because they will be compliers at more than one value of the treatment.

In some settings, we may expect different treatment effects for different per-unit contrasts. For instance, if we think completing high school has a large effect on wages, then the treatment effect of receiving 12 years of schooling rather than 11 may be larger than the treatment effect of receiving 11 years rather than 10. In settings with effect heterogeneity of this sort, researchers may wish to use the procedure outlined in Section 5 of Angrist and Imbens (1995) to see who contributes most to the weighted average treatment effect.

For truly continuous treatments, 2SLS recovers a weighted average of derivatives of the causal treatment–outcome function, with an analogous weighting scheme to the one described above (Angrist, Graddy and Imbens, 2000; Angrist and Pischke, 2008).

E.3 Multi-valued Instruments

Finally, consider the case with a single binary treatment and a multi-valued or continuous instrument. We can construct a pairwise LATE for any two values of the instrument z and z' where $z \neq z'$. To fix ideas, suppose we use proximity to the nearest college as an instrument for obtaining a college degree. Alice might be a complier at $Z = 0.5$ miles. That is, if Alice is 0.5 miles (or less) from the nearest college, she will obtain a degree, but if she is any further than 0.5 miles from the nearest college, she will not. Bob, in contrast, is a complier at $Z = 10$ miles. If Bob is 10 miles (or less) from the nearest college, he will obtain a degree, but if he is any further than 10 miles, he will not. We can define the pairwise LATE for a continuous or multi-valued instrument and binary treatment as follows:

$$\text{LATE}(z, z') = E[Y_i(1) - Y_i(0) \mid D_i(z) > D_i(z')].$$

The specific pairwise LATE we consider will affect whether Alice or Bob are included in the set of compliers for whom we are estimating treatment effects. If we consider the values $z = 11$ and $z' = 9$, Bob will be a complier but Alice will not. If we consider the values $z = 0.7$ and $z' = 0.3$, Alice will be a complier but Bob will not. If we consider the values $z = 7$ and $z' = 4$, neither will be compliers, and at $z = 11$ and $z' = 0.3$, both will be compliers.

If the instrument is discrete and multi-valued, a fully saturated 2SLS model will recover a weighted average of pairwise LATEs, where LATEs for pairs that induce more variation in the treatment will be upweighted (Angrist, Graddy and Imbens, 2000; Cornelissen et al., 2016). For continuous instruments, we get an analogous weighted average of derivatives.

It is worth noting that for continuous instruments, the positivity assumption is quite strong. It states that, across all values of the covariates, units have a positive probability of receiving *any* value of the instrument. Furthermore, continuous instruments require an additional monotonicity assumption that the instrument has a monotonic association with the treatment probability (Cornelissen et al., 2016). Without this assumption, the weights in the 2SLS estimand can be negative.

E.4 Beyond 2SLS

In the presence of covariates, various semi- and non-parametric estimators have been proposed to recover the LATE rather than a weighted average of covariate-specific LATEs (Abadie, 2003; Tan, 2006; Ogburn, Rotnitzky and Robins, 2015; Słoczyński, Uysal and Wooldridge, 2022). In addition to targeting a more easily interpreted causal

contrast, these estimators relax the linearity and additivity assumptions of 2SLS. Relaxing these parametric assumptions, of course, comes at the cost of higher variance.

Heckman and Vytlačil (2001) introduced two alternative estimands for the case with continuous instruments: the Marginal Treatment Effect (MTE) and the Policy-Relevant Treatment Effect (PRTE) (see Heckman and Vytlačil (2005) and Heckman, Urzua and Vytlačil (2006) as well). The MTE is given by

$$\text{MTE}(z) = E[Y_i(1) - Y_i(0) \mid D_i(z) = 1 \text{ and } D_i(z') = 0 \text{ for any } z' < z],$$

where we employ the notation of Zeng, Li and Ding (2020). The MTE captures the average treatment effect for those who are at the margin of receiving the treatment at instrument level z . The PRTE is a weighted aggregate of MTEs, where the weights are determined by a policy specified by the researcher. Heckman and Vytlačil (2001), for instance, calculate PRTEs for the economic returns to college based on different hypothetical policies that would change tuition costs, either by pushing everyone to a certain point in the distribution of tuition or by shifting the entire distribution of tuition. Kennedy, Lorch and Small (2019) develop semi-parametric estimators for estimating the MTE.