Political Analysis (2024), **0**, 1–20
doi:10.1017/pan.2024.2

PA

ARTICLE

How Much Should We Trust Instrumental Variable Estimates in Political Science? Practical Advice Based on 67 Replicated Studies

Apoorva Lal[1], Mackenzie Lockhart[2], Yiqing Xu[3] and Ziwen Zu[4]

[1]Independent Researcher; [2]Institution for Social and Policy Studies, Yale University, New Haven, CT 06511, USA; [3]Department of Political Science, Stanford University, Stanford, CA 94305, USA; [4]Department of Political Science, University of California, San Diego, La Jolla, CA 92093, USA

Corresponding author: Yiqing Xu; Email: yiqingxu@stanford.edu

Abstract

Instrumental variable (IV) strategies are widely used in political science to establish causal relationships, but the identifying assumptions required by an IV design are demanding, and assessing their validity remains challenging. In this paper, we replicate 67 articles published in three top political science journals from 2010 to 2022 and identify several concerning patterns. First, researchers often overestimate the strength of their instruments due to non-i.i.d. error structures such as clustering. Second, IV estimates are often highly uncertain, and the commonly used *t*-test for two-stage-least-squares (2SLS) estimates frequently underestimate the uncertainties. Third, in most replicated studies, 2SLS estimates are significantly larger in magnitude than ordinary-least-squares estimates, and their absolute ratio is inversely related to the strength of the instrument in observational studies—a pattern not observed in experimental ones—suggesting potential violations of unconfoundedness or the exclusion restriction in the former. We provide a checklist and software to help researchers avoid these pitfalls and improve their practice.

Keywords: instrumental variables; two-stage-least-squares; replications; weak instruments; unconfoundedness; exclusion restriction; publication bias; meta-analysis

Edited by: Jeff Gill

1. Introduction

The instrumental variable (IV) approach is a widely used empirical method in the social sciences, including political science, for establishing causal relationships. It is often used when selection on observables is implausible, experimentation is infeasible or unethical, and rule-based assignments that allow for sharp regression-discontinuity (RD) designs are unavailable. In recent years, there have been a growing number of articles published in top political science journals, such as the *American Political Science Review* (*APSR*), *American Journal of Political Science* (*AJPS*), and *The Journal of Politics* (*JOP*), that use IV as a primary causal identification strategy (Figure 1). This trend can be traced back to the publication of *Mostly Harmless Econometrics* (Angrist and Pischke 2009), which popularized the modern interpretation of IV designs, and Sovey and Green (2011), which clarifies the assumptions required by an IV design and provides a useful checklist for political scientists.

Despite its popularity, the IV approach has faced scrutiny from researchers who note that two-stage least-squares (2SLS) estimates are often much larger in magnitude than "naïve" ordinary-least-squares

© The Author(s), 2024. Published by Cambridge University Press on behalf of The Society for Political Methodology
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/pan.2024.2 Published online by Cambridge University Press

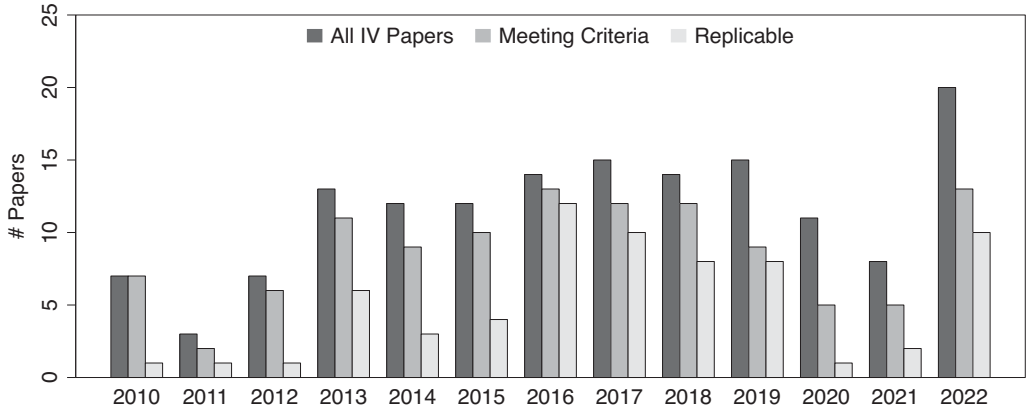


Figure 1. IV studies published in the *APSR*, *AJPS*, and *JOP*. Our criteria rule out IV models appearing in the Supplementary Material only, in dynamic panel settings, with multiple endogenous variables, and with nonlinear link functions. Non-replicability is primarily due to a lack of data and/or coding errors.

(OLS) estimates, even when the main concern with the latter is upward omitted-variables bias.[1] Others have raised concerns about the validity of the commonly used inferential method for 2SLS estimation (e.g., Lee *et al.* 2022; Young 2022).

These observations motivate our systematic examination of the use of IVs in the empirical political science literature. We set out to replicate all studies published in the *APSR*, *AJPS*, and *JOP* during the past 13 years (2010–2022) that use an IV design with a single endogenous variable as one of the main identification strategies.[2] Out of 114 articles meeting this criterion, 71 have complete replication materials online, which is concerning in itself. We successfully replicated at least one primary IV result in 67 out of the remaining 71 articles. Among the 67 articles, three articles feature two distinct IV designs, each yielding two separate replicable IV results.

Using data from these 70 IV designs, we conduct a programmatic replication exercise and find three troubling patterns. First, a significant number of IV designs in political science overestimate the first-stage partial F-statistic by failing to adjust standard errors (SEs) for factors such as heteroskedasticity, serial correlation, or clustering structure. Using the effective F-statistic (Olea and Pflueger 2013), we find that at least 11% of the published IV studies rely on what econometricians call "weak instruments," the consequences of which have been well documented in the literature. See Andrews, Stock, and Sun (2019) for a comprehensive review.

Second, obtaining valid statistical inferences for IV estimates remains challenging. Almost all studies we have replicated rely on t-tests for the 2SLS estimates based on analytic SEs and traditional critical values (such as 1.96 for statistical significance at the 5% level). Using analytic SEs, IV estimates are already shown to be much more imprecise than OLS estimates. When employing bootstrapping procedures, the AR test, or the tF procedure—an F-statistic-dependent t-test (Lee *et al.* 2022)—for hypothesis testing, we find that 17%–35% of the designs cannot reject the null hypothesis of no effect at the 5% level. In contrast, only 10% of studies based on originally reported SEs or p-values fail to reject the null hypothesis. This discrepancy suggests that many studies may have underestimated the uncertainties associated with their 2SLS estimates.

[1] For example, in the 2016 National Bureau of Economic Research–Political Economy Meeting, following a presentation of a study using an IV approach, the late political economist Alberto Alesina asked the audience: "How come 2SLS estimates are always five times bigger than OLS estimates in political economy?" We dedicate this paper to him for his seminal contributions to the field of political economy.

[2] Replication data and code are available at https://doi.org/10.7910/DVN/MM5THZ (Lal *et al.* 2024). Focusing on design with a single endogenous variable allows us to calculate the first-stage correlation coefficient (or R^2) and apply tools such as the Anderson–Rubin (AR) test and the tF test (when there is only a single instrument). Moreover, we find it difficult to justify the exclusion restriction in a multiple-treatment-multiple-instrument setting in the first place.

https://doi.org/10.1017/pan.2024.2 Published online by Cambridge University Press

What is even more concerning is that an IV approach can produce larger biases than OLS when a weak first-stage amplifies biases due to failures of IVs' unconfoundedness or exclusion restrictions. We observe that in 68 out of the 70 designs (97%), the 2SLS estimates have a larger magnitude than the naïve OLS estimates obtained from regressing the outcome on potentially endogenous treatment variables and covariates; 24 of these (34%) are at least five times larger. This starkly contrasts with the common rationale for using IV, which is to mitigate the upward bias in treatment effect estimates from OLS. Moreover, we find that the ratio between the magnitudes of the 2SLS and OLS estimates is strongly negatively correlated with the strength of the first stage among studies that use nonexperimental instruments, and the relationship is almost nonexistent among experimental studies. While factors such as heterogeneous treatment effects and measurement error might be at play, we contend that this phenomenon primarily stems from a combination of weak instruments and the failure of unconfoundedness or the exclusion restriction. Intuitively, because the 2SLS estimator is a ratio, an inflated numerator from invalid instruments paired with a small denominator due to a weak first stage leads to a disproportionately large estimate. Publication bias and selective reporting exacerbate this issue.

What do these findings imply for IV studies in political science? First, the traditional F-tests for IV strength, particularly when using classic analytic SEs, often mask the presence of weak instruments. Second, when operating with these weak instruments, particularly in overidentified scenarios, traditional t-tests do not adequately represent the considerable uncertainty surrounding the 2SLS estimates, paving the way for selective reporting and publication bias. Last but not least, many 2SLS estimates likely bear significant biases due to violations of unconfoundedness or the exclusion restriction, and a weak first-stage further exacerbates these biases. While we cannot pinpoint exactly which estimates are problematic, these issues seem to be pervasive across observational IV studies. The objective of this paper, however, is not to discredit existing IV research or dissuade scholars from using IVs. On the contrary, our intent is to caution researchers against the pitfalls of ad hoc justifications for IVs in observational research and provide constructive recommendations for future practices. These suggestions include accurately quantifying the strength of instruments, conducting valid inference for IV estimates, as well as implementing additional validation exercises, such as placebo tests, to bolster the identifying assumptions.

Our work builds on a growing literature evaluating IV strategies in social sciences and offering methods to improve empirical practice. Notable studies include Young (2022), which finds IV estimates to be more sensitive to outliers and conventional t-tests to understate uncertainties; Jiang (2017), which observes larger IV estimates in finance journals and attributed this to exclusion restriction violations and weak instruments; Mellon (2023), which emphasizes the vulnerability of weather instruments; Dieterle and Snell (2016), which develops a quadratic overidentification test and discovers significant nonlinearities in the first stage regression; Felton and Stewart (2022), which finds unstated assumptions and a lack of weak-instrument robust tests in top sociology journals; and Cinelli and Hazlett (2022), which proposes a sensitivity analysis for IV designs in an omitted variable bias framework. This study is the first comprehensive replication effort focusing on IV designs in political science and uses data to shed light on the consequences of a weak first-stage interacting with failures of unconfoundedness or the exclusion restriction.

2. Theoretical Refresher

In this section, we offer a brief overview of the IV approach, including the setup, the key assumptions, and the 2SLS estimator. We then discuss potential pitfalls and survey several inferential methods. To cover the vast majority of IV studies in political science, we adopt a traditional constant treatment effect approach, which imposes a set of parametric assumptions. For example, of our replication sample, 51 designs (73%) employ continuous treatment variables, and 49 (70%) use continuous IVs. Most of these studies make no reference to treatment effect heterogeneity and are ill-suited for the local average treatment effect framework (Angrist, Imbens, and Rubin 1996).

Apart from the canonical use of IVs in addressing noncompliance in experimental studies, we observe that in the majority of the articles we review, researchers use IVs to establish causality between

https://doi.org/10.1017/pan.2024.2 Published online by Cambridge University Press

a single treatment variable d and an outcome variable y in observational settings. The basic idea of this approach is to use a vector of instruments z to isolate "exogenous" variation in d (i.e., the variation in d that is not related to potential confounders) and estimate its causal effect on y. For simplicity, we choose not to include any additional exogenous covariates in our discussion. This is without loss of generality because, by the Frisch–Waugh–Lovell theorem, we can remove these variables by performing a regression of y, d, and each component of z on the controls and then proceeding with our analysis using the residuals instead.

2.1. Identification and Estimation

Imposing a set of parametric assumptions, we define a system of simultaneous equations:

$$\text{Structural equation:} \quad y = \tau_0 + \tau d + \varepsilon, \tag{2.1}$$

$$\text{First-stage equation:} \quad d = \pi_0 + \pi' z + \nu, \tag{2.2}$$

in which y is the outcome variable; d is a scalar treatment variable; z is a vector of instruments for d; and τ captures the (constant) treatment effect and is the key quantity of interest. The error terms ε and ν may be correlated. The endogeneity problem for τ in Equation (2.1) arises when d and ε are correlated, which renders $\hat{\tau}_{OLS}$ from a naïve OLS regression of y on d inconsistent. This may be due to several reasons: (1) unmeasured omitted variables correlated with both y and d, (2) measurement error in d, or (3) simultaneity or reverse causality, which means y may also affect d. Substituting d in Equation (2.1) using Equation (2.2), we have the reduced form equation:

$$\text{Reduced form:} \quad y = \underbrace{(\alpha + \tau\pi_0)}_{\gamma_0} + \underbrace{(\tau\pi)'}_{\gamma'} z + (\tau\nu + \varepsilon). \tag{2.3}$$

Substitution establishes that $\gamma = \tau\pi$, rearranging yields $\tau = \frac{\gamma}{\pi}$ (assuming a single instrument, but the intuition carries over to cases with multiple instruments). The IV estimate, therefore, is the ratio of the reduced-form and first-stage coefficients. To identify τ, we make the following assumptions (Greene 2003, Chapter 12).

Assumption 1 (**Relevance**). $\pi \neq 0$. *This assumption requires that the IVs can predict the treatment variable, and is therefore equivalently stated as* $d \not\perp z$.

Assumption 2 (**Exogeneity: unconfoundedness and the exclusion restriction**). $\mathbb{E}[\varepsilon] = 0$ *and* $Cov(z, \varepsilon) = 0$. *This assumption is satisfied when unconfoundedness (random or quasi-random assignment of z) and the exclusion restriction (no direct effect of z on y beyond d) are met.*

Conceptually, unconfoundedness and the exclusion restriction are two distinct assumptions and should be justified separately in a research design. However, because violations of either assumption lead to the failure of the 2SLS moment condition, $\mathbb{E}[z\varepsilon] = 0$, and produce observationally equivalent outcomes, we consider both to be integral components of Assumption 2.

Under Assumptions 1 and 2, the 2SLS estimator is shown to be consistent for the structural parameter τ. Consider a sample of N observations. We can write $\mathbf{d} = (d_1, d_2, \ldots, d_N)'$ and $\mathbf{y} = (y_1, y_2, \ldots, y_N)'$ as $(N \times 1)$ vectors of the treatment and outcome data, and $\mathbf{z} = (z_1, z_2, \ldots, z_N)'$ as an $(N \times p_z)$ matrix of instruments in which p_z is the number of instruments. The 2SLS estimator is written as follows:

$$\hat{\tau}_{2SLS} = \left(\mathbf{d}'\mathbf{P}_z\mathbf{d}\right)^{-1}\mathbf{d}'\mathbf{P}_z\mathbf{y}, \tag{2.4}$$

in which $\mathbf{P}_z = \mathbf{z}\left(\mathbf{z}'\mathbf{z}\right)^{-1}\mathbf{z}'$ is the hat-maker matrix from the first stage which projects the endogenous treatment variable $\mathbf{d}$ into the column space of $\mathbf{z}$, thereby in expectation preserving only the exogenous variation in $\mathbf{d}$ that is uncorrelated with ε. This formula permits the use of multiple instruments, in which case the model is said to be "overidentified." The 2SLS estimator belongs to a class of generalized method of moments (GMM) estimators taking advantage of the moment condition $\mathbb{E}[z\varepsilon] = 0$, including the two-step GMM (Hansen 1982) and limited information maximum likelihood estimators

https://doi.org/10.1017/pan.2024.2 Published online by Cambridge University Press

(Anderson, Kunitomo, and Sawa 1982). We use the 2SLS estimator throughout the replication exercise because of its simplicity and because every single paper in our replication sample uses it in at least one specification.

When the model is exactly identified, that is, the number of treatment variables equals the number of instruments, the 2SLS estimator can be simplified as the IV estimator: $\hat{\tau}_{\text{2SLS}} = \hat{\tau}_{\text{IV}} = \left(\mathbf{z}'\mathbf{d}\right)^{-1}\mathbf{z}'\mathbf{y}$. In the case of one instrument and one treatment, the 2SLS estimator can also be written as a ratio of two sample covariances: $\hat{\tau}_{2SLS} = \hat{\tau}_{IV} = \frac{\hat{\gamma}}{\hat{\pi}} = \frac{\widehat{\text{Cov}}(y,z)}{\widehat{\text{Cov}}(d,z)}$, which illustrates that the 2SLS estimator is a ratio between reduced-form and first-stage coefficients in this special case. This further simplifies to a ratio of the differences in means when z is binary, which is called a Wald estimator.

2.2. Potential Pitfalls in Implementing an IV Strategy

The challenges with 2SLS estimation and inference are mostly due to violations of Assumptions 1 and 2. Such violations can result in (1) significant uncertainties around 2SLS estimates and size distortion for t-tests due to weak instruments even when the exogeneity assumption is satisfied and (2) potentially larger biases in 2SLS estimates compared to OLS estimates when the exogeneity assumption is violated.

2.2.1. Inferential Problem Due to Weak Instruments

Since the IV coefficient is a ratio, the weak instrument problem is a "divide-by-zero" problem, which arises when $\text{Cov}(z,d) \approx 0$ (i.e., when the relevance assumption is violated). The instability of ratio estimators like $\widehat{\tau}_{\text{2SLS}}$ when the denominator is approximately zero has been extensively studied going back to Fieller (1954). The conventional wisdom in the past two decades has been that the first-stage partial F-statistic needs to be bigger than 10, and it should be clearly reported (Staiger and Stock 1997). The cutoff, as a rule of thumb, is chosen based on simulation results to meet two criteria under i.i.d. errors: (1) in the *worst* case, the bias of the 2SLS estimator does not exceed 10% of the bias of the OLS estimator and (2) a t-test based on the 2SLS estimator with a size of 5% does not lead to size over 15%. These problems are further exacerbated in settings where units belong to clusters with strong within-cluster correlation, where a small number of observations or clusters may heavily influence estimated results (Young 2022). Recently, however, Angrist and Kolesár (2023) argue that the conventional inference strategies are reliable in just-identified settings with independent errors. The weak instrument issue is indeed most concerning in heavily overidentified scenarios.

The literature has discussed at least three issues caused by weak instruments when the exogeneity assumption is satisfied. First, under i.i.d. errors, a weak first stage exacerbates the finite-sample bias of the 2SLS estimator toward the inconsistent OLS estimator, thereby reproducing the endogeneity problem that an IV design was meant to solve (Staiger and Stock 1997). Additionally, when the first stage is weak, the 2SLS estimator may not have a mean; its median is centered around the OLS coefficient (Hirano and Porter 2015). Second, the 2SLS estimates become very imprecise. To illustrate, a commonly used variance estimator for $\hat{\tau}_{IV}$ is $\hat{\mathbb{V}}(\hat{\tau}_{IV}) \approx \hat{\sigma}^2/(\sum_{i=1}^{N}(d_i - \overline{d})^2 R^2_{dz}) = \hat{\mathbb{V}}(\hat{\tau}_{OLS})/R^2_{dz}$, in which $\hat{\sigma}^2$ is a variance estimator for the error term and R^2_{dz} is the first stage R^2. $\hat{\mathbb{V}}(\hat{\tau}_{IV})$ is generally larger than $\hat{\mathbb{V}}(\hat{\tau}_{OLS})$ and increasing in $1/R^2_{dz}$. A third and related issue is that the t-tests are of the wrong size and the t-statistics do not follow a t-distribution (Nelson and Starz 1990). This is because the distribution of $\hat{\tau}_{2SLS}$ is derived from its linear approximation of $\hat{\tau}_{2SLS}$ in $(\hat{\gamma}, \hat{\pi})$, wherein normality of the two OLS coefficients implies the normality of their ratio. However, this normal approximation breaks down when $\hat{\pi} \approx 0$. Moreover, this approximation failure cannot generally be rectified by bootstrapping (Andrews and Guggenberger 2009); Young (2022) argues that it nevertheless allows for improved inference when outliers are present. Overall, valid IV inference relies crucially on strong IVs.

Generally, there are two approaches to conducting inference in an IV design: pretesting and direct testing. The pretesting approach involves using an F-statistic to test the first stage strength, and if it exceeds a certain threshold (e.g., $F > 10$), proceeding to test the null hypothesis about the treatment effect (e.g., $\tau = 0$). Nearly all reviewed studies employ this approach. The direct testing approach, in

https://doi.org/10.1017/pan.2024.2 Published online by Cambridge University Press

contrast, does not rely on passing a pretest. We examine four inferential methods for IV designs, with the first three related to pretesting and the last one being a direct test.

First, Olea and Pflueger (2013) propose the effective F-statistic for both just-identified and overidentified settings and accommodates robust or cluster-robust SEs. The effective F is a scaled version of the first-stage F-statistic and is computed as $F_{\text{Eff}} = \hat{\pi}'\hat{Q}_{ZZ}\hat{\pi}/\text{tr}(\hat{\Sigma}_{\pi\pi}\hat{Q}_{ZZ})$, where $\hat{\Sigma}_{\pi\pi}$ is the variance–covariance matrix of the first stage regression, and $\hat{Q}_{ZZ} = \frac{1}{N}\sum_{i=1}^{N} z_i z_i'$. In just-identified cases, F_{Eff} is the same as an F-statistic based on robust or cluster-robust SEs. The authors derive the critical values for F_{Eff} and note that the statistic and corresponding critical values are identical to the better-known robust F-statistic $\hat{\pi}\hat{\Sigma}_{\pi\pi}^{-1}\hat{\pi}$ and corresponding Stock and Yogo (2005) critical values. $F_{\text{Eff}} > 10$ is shown to be a reasonable rule of thumb under heteroskedasticity in simulations (Andrews, Stock, and Sun 2019; Olea and Pflueger 2013).

Second, Young (2022) recommends researchers report two types of bootstrap confidence intervals (CIs), *bootstrap-c* and *bootstrap-t*, for $\hat{\tau}_{2SLS}$ under non-i.i.d. errors with outliers, which is common in social science settings. They involve B replications of the following procedure: (1) sample n triplets (y_i^*, d_i^*, z_i^*) independently and with replacement from the original sample (with appropriate modifications for clustered dependence) and (2) on each replication, compute the 2SLS coefficient and SE, as well as the corresponding test statistic $t^* = \hat{\tau}^*_{2SLS}/\widehat{\text{SE}}(\hat{\tau}^*_{2SLS})$. The *bootstrap-c* method calculates the CIs by taking the $\alpha/2$ and $(1-\alpha/2)$ percentiles of the bootstrapped 2SLS coefficients $\hat{\tau}^*_{2SLS}$, while the *bootstrap-t* method calculates the percentile-t refined CIs by plugging in the $\alpha/2$ and $(1-\alpha/2)$ percentile of the bootstrapped t statistics into the expression $\hat{\tau}_{2SLS} \pm t^*_{\alpha|1-\alpha}\widehat{\text{SE}}(\hat{\tau}^*_{2SLS})$. Hall and Horowitz (1996) show that *bootstrap-t* achieves an asymptotic refinement over *bootstrap-c*. Note that t-tests based on bootstrapped SEs may be overly conservative (Hahn and Liao 2021) and are therefore not recommended.

Third, in just-identified single treatment settings, Lee *et al.* (2022) propose the *tF* procedure that smoothly adjusts the t-ratio inference based on the first-stage F-statistic, which improves upon the ad hoc screening rule of $F > 10$. The adjustment factor applied to 2SLS SEs is based on the first stage t-ratio $\hat{f} := \hat{\pi}/\hat{\mathbb{V}}(\hat{\pi})^{1/2}$, with the first stage $\hat{F} = \hat{f}^2$, and relies on the fact that the distortion from employing the standard 2SLS t-ratio $\hat{t} := \hat{\tau}/\hat{\mathbb{V}}(\hat{\tau})^{1/2}$ can be quantified in terms of an $\hat{F}$ statistic, which gives rise to a set of critical values for a given pair of $\hat{t}$ and $\hat{F}$. The authors also show that if no adjustment is made to the t-test's critical value (e.g., using 1.96 as the threshold for 5% statistical significance), a first stage $\hat{F}$ of 104.7 is required to guarantee a correct size of 5% for a two-sided t-test for the 2SLS coefficient.

Finally, where there is one endogenous treatment variable, the AR procedure, which is essentially an F-test on the reduced form, is a direct inferential method robust to weak instruments (Anderson and Rubin 1949; Chernozhukov and Hansen 2008). Without loss of generality, assume that we are interested in testing the null hypothesis that $\tau = 0$, which then implies that the reduced form coefficient from regressing y on z is zero, i.e., $\gamma = 0$. This motivates the following procedure: given a set $\mathcal{T}$ of potential values for $\widetilde{\tau}$, for each value $\widetilde{\tau}$, construct $\widetilde{y} = y - d\widetilde{\tau}$, and regress $\widetilde{y}$ on z to obtain a point estimate $\widetilde{\gamma}$ and (robust, or cluster robust) covariance matrix $\widetilde{\mathbb{V}}(\widetilde{\gamma})$, and construct a Wald statistic $\widetilde{W}_s(\widetilde{\gamma}) := \widetilde{\gamma}'\widetilde{\mathbb{V}}(\widetilde{\gamma})^{-1}\widetilde{\gamma}$. Then, the AR CI (or confidence set) is the set of $\widetilde{\gamma}$ such that $\widetilde{W}_s(\widetilde{\gamma}) \leq c(1-p)$ where $c(1-p)$ is the $(1-p)$th percentile of the χ^2_1 distribution. The AR test requires no pretesting and is shown to be the uniformly most powerful unbiased test in the just-identified case (Moreira 2009). It is less commonly used than pretesting procedures possibly because researchers are more accustomed to using t-tests than F-tests and reporting SEs rather than CIs. A potential limitation of the AR test is that its CIs can sometimes be empty or disconnected, and therefore lack a Bayesian interpretation under uninformative priors.[3]

2.2.2. *Bias Amplification and the Failure of the Exogeneity Assumption*

When the number of instruments is bigger than the number of endogenous treatments, researchers can use an overidentification test to gauge the plausibility of Assumption 2, the exogeneity assumption

[3]We thank Guido Imbens for highlighting this point.

https://doi.org/10.1017/pan.2024.2 Published online by Cambridge University Press

(Arellano 2002). However, such a test is often underpowered and has bad finite-sample properties (Davidson and MacKinnon 2015). In just-identified cases, Assumption 2 is not directly testable. When combined with instruments that are weak, even small violations of unconfoundedness or the exclusion restriction can produce inconsistency. This is because: plim $\hat{\tau}_{IV} = \tau + \frac{\text{Cov}(z,\varepsilon)}{\text{Cov}(z,d)}$. When $\text{Cov}(z,d) \approx 0$, even small violations of exogeneity, that is, $\text{Cov}(z,\varepsilon) \neq 0$, will enlarge the second term, resulting in large biases. Thus, the two identifying assumption failures exacerbate each other: having a weak first-stage compounds problems from confounding or exclusion restriction violations, and vice versa. With invalid instruments, it is likely that the asymptotic bias of the 2SLS estimator is much greater than that of the OLS estimator, that is, $\left|\frac{\text{Cov}(z,\varepsilon)}{\text{Cov}(z,d)}\right| \gg \left|\frac{\text{Cov}(d,\varepsilon)}{\mathbb{V}[d]}\right|$ in the single instrument case.[4]

While the inferential problem can be alleviated by employing alternative inferential methods as described above, addressing violations of unconfoundedness or the exclusion restriction is more challenging since it is fundamentally a research design issue that should be tackled at the design stage. Researchers often devote significant effort to arguing for unconfoundedness and exclusion restrictions in their settings. In Section A3 of the Supplementary Material (SM), we provide an exposition of the zero-first-stage (ZFS) test (Bound and Jaeger 2000), which is essentially a placebo test on a subsample where the instrument is expected to be uncorrelated with the treatment, to help researchers gauge the validity of their instruments. These estimates can then be used to debias the 2SLS estimate using the methods proposed in Conley, Hansen, and Rossi (2012).

3. Data and Types of Instruments

In this section, we first discuss our case selection criteria and replication sample, which is the focus of our subsequent analysis. We then describe the types of instruments in the replicable studies.

3.1. Data

We examine all empirical articles published in the *APSR*, *AJPS*, and *JOP* from 2010 to 2022 and identify studies that use an IV strategy as one of the main identification strategies, including articles that use binary or continuous treatments and that use a single or multiple instruments. We use the following criteria: (1) the discussion of the IV result needs to appear in the main text and support a main argument in the paper; (2) we consider linear models only; in other words, articles that use discrete outcome models are excluded from our sample;[5] (3) we exclude articles that include multiple endogenous variables in a single specification (multiple endogenous variables in separate specifications are included); (4) we exclude articles that use IV or GMM estimators in a dynamic panel setting because the validity of the instruments (e.g., y_{t-2} affects y_{t-1} but not y_t) is often not grounded in theories or substantive knowledge; these applications are subject to a separate set of empirical issues, and their poor performance has been discussed in the literature (e.g., Bun and Windmeijer 2010). These criteria result in 30 articles in the *APSR*, 33 articles in the *AJPS*, and 51 articles in the *JOP*. We then strive to find replication materials for these articles from public data-sharing platforms, such as the Harvard Dataverse, and the authors' websites. We are able to locate complete replication materials for 76 (62%) articles. However, code completeness and documentation quality vary widely. Since 2016–2017, data availability has significantly improved, thanks to new editorial policies that require authors to make replication materials publicly accessible (Key 2016). Starting in mid-2016 for *AJPS* and early-2021

[4] We are not the first to make this argument. According to Hahn and Hausman (2005, 34): "[T]he empirical finding that the 2SLS estimate increases compared to the OLS estimate may indicate that the instrument is not orthogonal to the stochastic disturbance. The resulting bias can be substantial. Indeed, it could exceed the OLS bias, leading to an increase in the estimated 2SLS coefficient over the estimated OLS coefficient."

[5] We expect the issue identified in this paper to be present, if not more pronounced, with nonlinear IVs. With nonlinear IVs, weak instruments correspond to weak identification in GMM estimations for some or all unknown parameters. Consequently, weak identification results in non-normal distributions even in large samples, rendering conventional IV or GMM inferences unreliable (Stock, Wright, and Yogo 2002).

https://doi.org/10.1017/pan.2024.2 Published online by Cambridge University Press

Table 1. Data availability and replicability of IV articles.

	#All articles	Incomplete data	Incomplete code	Replication error	Replicable
APSR	33	16	0	3	14 (42%)
AJPS	30	3	1	1	25 (83%)
JOP	51	19	3	1	28 (55%)
Total	114	38	4	5	67 (59%)

for *JOP*, both journals introduced a policy requiring third-party verification of full replicability as a prerequisite for publication, although not all data are made public. We view these measures as significant advancements.

Using data and code from the replication materials, we set out to replicate the main IV results in these 76 articles with complete data. Our replicability criterion is simple: As long as we can exactly replicate *one* 2SLS point estimate that appears in the paper, we deem the paper replicable. We do not aim at exactly replicating SEs, z-scores, or level of statistical significance for the 2SLS estimates because they involve the choice of the inferential method. After much effort and hundreds of hours of work, we are able to replicate the main results of 67 articles.[6] The low replication rate is consistent with what is reported in Hainmueller, Mummolo, and Xu (2019) and Chiu *et al.* (2023). The main reasons for failures of replication are incomplete data (38 articles), incomplete code or poor documentation (4 articles), and replication errors (5 articles). Table 1 presents summary statistics on data availability and replicability of IV articles for each of the three journals. The rest of this paper focuses on results based on these 67 replicable articles (and 70 IV designs).

3.2. Types of Instruments

Inspired by Sovey and Green (2011), in Table 2, we summarize the types of IVs in the replicable designs, although our categories differ from theirs to reflect changes in the types of instruments used in the discipline. These categories are ordered based on the strength of the design, in our view, for an IV study.

The first category is randomized experiments. These articles employ randomization, designed and conducted by researchers or a third party, and use 2SLS estimation to tackle noncompliance. With random assignment, our confidence in the exogeneity assumption increases because unconfoundedness is guaranteed by design and the direct effect of the instrument on the outcome is easier to rule out than without random assignment. For instance, Alt, Marshall, and Lassen (2016) use assignment to an information treatment as an instrument for economic beliefs to understand the relationship between economic expectations and vote choice. Compared to IV articles published before 2010, the proportion of articles using experiment-generated IVs has increased significantly (from 2.9% to 17.1%) due to the growing popularity of experiments.

Another category consists of instruments derived from explicit rules on observed covariates, creating quasi-random variations in the treatment. Sovey and Green (2011) refer to this category as "Natural Experiment." We avoid this terminology because it is widely misused and limit this category to two circumstances: fuzzy RD designs and variation in exposure to policies due to time of birth or eligibility. For example, Kim (2019) leverages a reform in Sweden that requires municipalities above a population threshold to adopt direct democratic institutions. Dinas (2014) uses eligibility to vote based on age at the time of an election as an instrument for whether respondents did vote. While rule-based IVs offer

[6]For three articles, we are able to produce the 2SLS estimates with perfectly executable code; however, our replicated estimates are inconsistent with what was reported in the original studies. We suspect the inconsistencies are caused by data rescaling or misreporting; hence, we keep them in the sample.

https://doi.org/10.1017/pan.2024.2 Published online by Cambridge University Press

Table 2. Types of instruments.

Type	#Articles	Percentage (%)
Experiment	12	17.1
Rules and policy changes	7	10.0
Fuzzy RD	4	5.7
Change in exposure	3	4.3
Theory	42	60.0
Weather/climate/geography	13	18.6
Treatment diffusion	2	2.9
History	11	15.7
Others	16	22.9
Econometrics	9	12.9
Interactions/"Bartik"	7	10.0
Lagged treatment	1	1.4
Empirical test	1	1.4
Total	70	100.0

a pathway to credible causal inference, recent studies have raised concerns about their implementation, highlighting issues of insufficient power in many RD designs (Stommes, Aronow, and Sävje 2023).

The next category is "Theory," where the authors justify unconfoundedness and the exclusion restriction using social science theories or substantive knowledge. Over a decade after Sovey and Green's (2011) survey, it remains the most prevalent category among IV studies in political science, at around 60%. We divide theory-based IVs into four subcategories: geography/climate/weather, treatment diffusion, history, and others. First, Many studies in the theory category justify the choices of their instruments based on geography, climate, or weather conditions. For example, Zhu (2017) uses weighted geographic closeness as an instrument for the activities of multinational corporations; Hager and Hilbig (2019) use mean elevation and distance to rivers to instrument equitable inheritance customs; Henderson and Brooks (2016) use rainfall around Election Day as an instrument for Democratic vote margins. Relatedly, several studies base their choices on regional diffusion of treatment. For example, Dube and Naidu (2015) use U.S. military aid to countries outside Latin America as an instrument for U.S. military aid to Colombia. Dorsch and Maarek (2019) use the regional share of democracies as an instrument for democratization in a country-year panel.[7] Third, historical instruments derive from past differences between units unrelated to current treatment levels. For example, Vernby (2013) uses historical immigration levels as an instrument for the current number of noncitizen residents. Finally, several articles rely on a unique instrument based on theories that we could not place in a category. Dower *et al.* (2018) use religious polarization as an instrument for the frequency of unrest and argue that religious polarization could only impact collective action through its impact on representation in local institutions.

We wish to clarify that our reservations regarding instruments in this category are not primarily about theories themselves. As a design-based approach, the IV strategy requires *specific* and *precise*

[7] While authors often argue that weather or geography is quasi-randomly imposed, it is typically harder to claim they only affect the outcome through the treatment variable. For example, Mellon (2023) contends that, while instruments like rain may be quasi-random, researchers have pinpointed several mechanisms through which it influences key political outcomes. Betz, Cook, and Hollenbach (2018) argue that spatial instruments are rarely valid because of cross-sectional interdependence and simultaneity. Inference also presents challenges.

https://doi.org/10.1017/pan.2024.2 Published online by Cambridge University Press

theories about the assignment process of the instruments and the exclusion restriction. We remain skeptical because many "theory"-driven instruments, in our view, do not genuinely uphold these assumptions, often appearing to be developed in an ad hoc or post hoc manner.

The last category of instruments are based on econometric assumptions. This category includes what Sovey and Green (2011) call "Lags." These are econometric transformations of variables argued to constitute instruments. For example, Lorentzen, Landry, and Yasuda (2014) use a measure of the independent variable from eight years earlier to mitigate endogeneity concerns. Another example is shift-share "Bartik" instruments-based. For example, Baccini and Weymouth (2021) use the interaction between job shares in specific industries and national employment changes to study the effect of manufacturing layoffs on voting. The number of articles relying on econometric techniques, including flawed empirical tests (such as regressing y on d and z and checking if the coefficient of z is significant), has decreased.

4. Replication Procedure and Results

In this section, we describe our replication procedure and report the main findings.

4.1. Procedure

For each paper, we select the main IV specification that plays a central role in supporting a main claim in the paper; it is either referred to as the baseline specification or appears in one of the main tables or figures. Focusing on this specification, our replication procedure involves the following steps. First, we compute the first-stage partial F-statistic based on (1) classic analytic SEs, (2) Huber White heteroskedastic-robust SEs, (3) cluster-robust SEs (if applicable and based on the original specifications), and (4) bootstrapped SEs.[8] We also calculate $F_{\texttt{Eff}}$.

We then replicate the original IV result using the 2SLS estimator and apply four different inferential procedures. First, we make inferences based on analytic SEs, including robust SEs or cluster-robust SEs (if applicable). Additionally, we use two nonparametric bootstrap procedures, as described in Section 2, *bootstrap-c* and *bootstrap-t*. For specifications with only a single instrument, we also employ the *tF* procedure proposed by Lee *et al.* (2022), using the 2SLS t-statistic and first-stage F-statistic based on analytic SEs accounting for the originally specified clustering structure. Finally, we conduct an AR procedure and record the p-values and CIs.

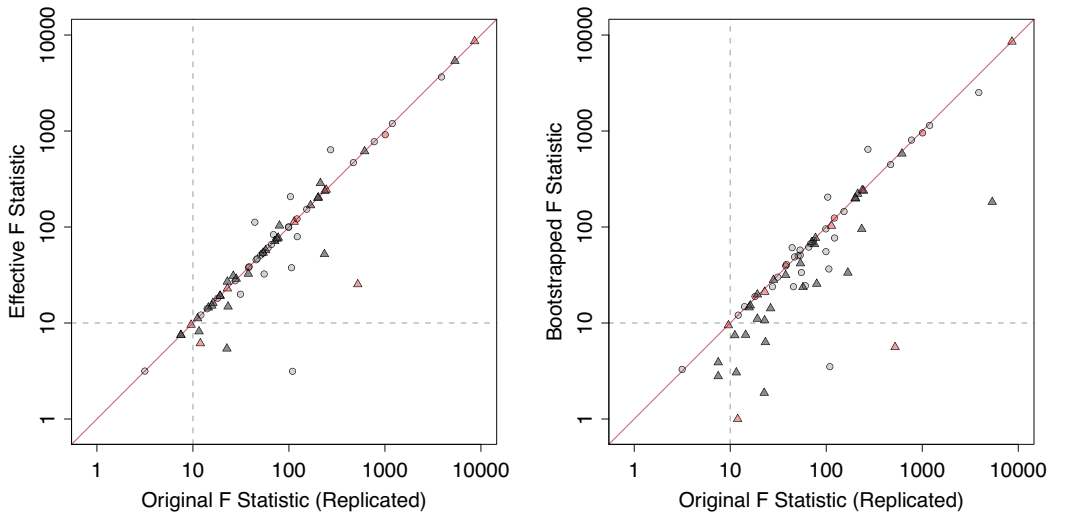
We record the point estimates, SEs (if applicable), 95% CIs, and p-values for each procedure (the point estimates fully replicate the reported estimates in the original articles and are the same across all procedures). In addition, we estimate a naïve OLS model by regressing the outcome variable on the treatment and covariates, leaving out the instrument. We calculate the ratio between the magnitudes of the 2SLS and OLS estimates, as well as the ratio of their analytic SEs. We also record other useful information, such as the number of observations, the number of clusters, the types of instruments, the methods used to calculate SEs or CIs, and the rationale for each paper's IV strategy. Our replication yields the following three main findings.

4.2. Finding 1. The First-Stage Partial F-Statistic

Our first finding regards the strengths of the instruments. To our surprise, among the 70 IV designs, 12 (17%) do not report this crucial statistic despite its key role in justifying the validity of an IV

[8]They are calculated by $F_{boot} = \hat{\tau}'_{2SLS}\hat{\mathbb{V}}_{boot}(\hat{\tau}_{2SLS})^{-1}\hat{\tau}_{2SLS}/p_z$, where p_z is the number of IVs and $\hat{\mathbb{V}}_{boot}(\hat{\tau}_{2SLS})$ is the estimated variance–covariance matrix based on a nonparametric bootstrap procedure, in which we repeatedly sample the rows of the data matrix with replacement. If the data have a clustered structure, we use cluster-bootstrapping instead (Colin Cameron and Miller 2015; Esarey and Menger 2019). We include F_{boot} as a reference to the classic F and effective F. In Section A2 of the SM, we compare the five types of F-statistics and show that the effective F and F based on bootstrapping are usually more conservative (smaller) than other F-statistics.

https://doi.org/10.1017/pan.2024.2 Published online by Cambridge University Press



(a) Original *F* vs. Effective *F* (b) Original *F* vs. Bootstrapped *F*

Figure 2. Original versus effective and bootstrapped *F*. Circles represent applications without a clustering structure and triangles represent applications with a clustering structure. Studies that do not report *F*-statistic are painted in red. The original *F*-statistics are obtained from the authors' original model specifications and choices of variance estimators in the 2SLS regressions. They may differ from those reported in the articles because of misreporting.

design. Among the remaining 58 studies that report *F*-statistic, 9 (16%) use classic analytic SEs, thus not adjusting for potential heteroskedasticity or clustering structure. In Figure 2, we plot the replicated first-stage partial *F*-statistic based on the authors' original model specifications and choices of variance estimators on the *x*-axis against (a) effective *F*-statistic or (b) bootstrapped *F*-statistic on the *y*-axis, both on a logarithmic scale.[9]

In the original studies, the authors used various SE estimators, such as classic SEs, robust SEs, or cluster-robust SEs. As a result, the effective *F* may be larger or smaller than the original ones. However, a notable feature of Figure 2 is that when a clustering structure exists, the original *F*-statistic tends to be larger than the effective *F* or bootstrapped *F*. When using the effective *F* as the benchmark, eight studies (11%) have $F_{\texttt{Eff}} < 10$. This number increases to 12 (17%) when the bootstrapped *F*-statistic is used. The median first-stage $F_{\texttt{Eff}}$ statistic is higher in experimental studies compared to nonexperimental ones (67.7 vs. 53.5). It is well known that failing to cluster the SEs at appropriate levels or using the analytic cluster-robust SE with too few clusters can lead to an overstatement of statistical significance (Cameron, Gelbach, and Miller 2008). However, this problem has received less attention when evaluating IV strength using the *F*-statistic.[10]

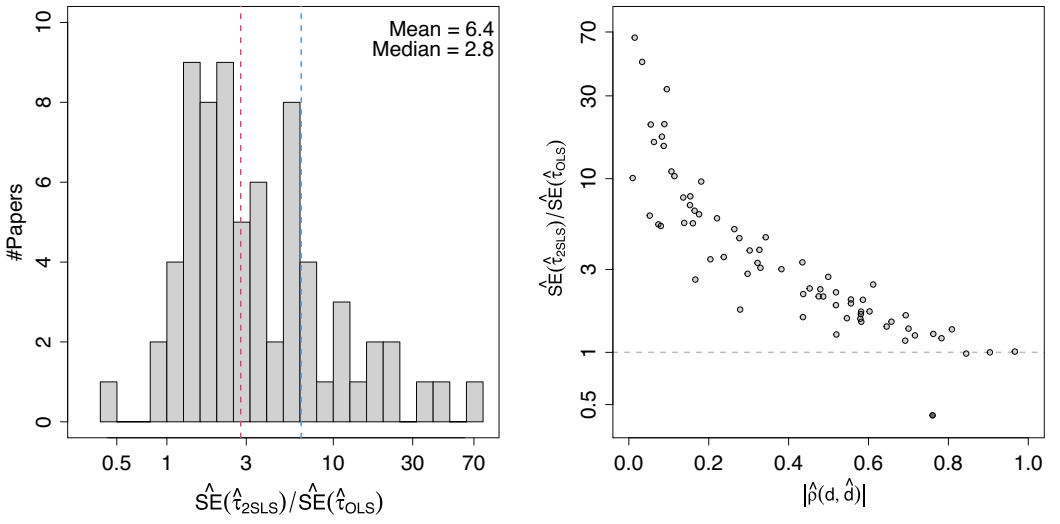
4.3. Finding 2. Inference

Typically, 2SLS estimates have higher uncertainties than OLS estimates. Figure 3 reveals that the 2SLS estimates in the replication sample are in general much less precise than their OLS counterparts, with

[9] We use the replicated *F*-statistics instead of the reported ones because some authors either do not report or misreport their *F*-statistics (see the SM for a comparison between the reported and replicated *F*-statistics).

[10] Abadie *et al.* (2020, 2023) delineate the differences between a traditional *sampling-based* view, where clustering arises from a two-stage sampling process (sampling clusters, then units within them), and a *design-based* view, where clustering stems from the clustered nature of treatment assignment. The key takeaway from both papers is the importance of clustering at the unit of randomization. They argue that finite-population standard errors, rooted in the design-based perspective, can be tighter than conventional cluster-robust errors. Given that the exact design is often unknown in many political science observational studies, clustering where the instrument is likely assigned offers a more reliable approach for valid inference. In the replication exercise, however, we cluster SEs according to the levels specified by the original authors.

https://doi.org/10.1017/pan.2024.2 Published online by Cambridge University Press



(a) Distribution of Ratio of 2SLS and OLS SEs

(b) First Stage and the Ratio

Figure 3. Comparison of 2SLS and OLS analytic SEs. Subfigure (a) shows the distribution of the ratio between $\hat{SE}(\hat{\tau}_{2SLS})$ and $\hat{SE}(\hat{\tau}_{OLS})$, both obtained analytically. Subfigure (b) plots the relationship between the absolute values of $\hat{\rho}(d,\hat{d})$, the estimated correlational coefficient between d and $\hat{d}$, and the ratio (on a logarithmic scale). In one study, the analytic $\hat{SE}(\hat{\tau}_{2SLS})$ is much smaller than $\hat{SE}(\hat{\tau}_{OLS})$; we suspect that the former severely underestimates the true SE of the 2SLS estimate, likely due to a clustering structure.

the median ratio of the analytic SEs equal to 3.8. This ratio decreases as the strength of the instrument, measured by the estimated correlation coefficient between the treatment and predicted treatment $|\hat{\rho}(d,\hat{d})|$, increases. This is not surprising because $\hat{\rho}(d,\hat{d})^2 = R^2_{dz}$, the first-stage partial R-squared. However, one important implication of large differences in SEs is that to achieve comparable levels of statistical significance, 2SLS estimates often need to be at least three times larger than OLS estimates—not to mention that t-testing based on analytical SEs for 2SLS coefficients is often overly optimistic. This difference in precision sets the stage for potential publication bias and p-hacking.

Next, we compare the reported and replicated p-values for the null hypothesis of no effect. For studies that do not report a p-value, we calculate it based on a standard normal distribution using the reported point estimates and SEs. The replicated p-values are based on (1) *bootstrap-c*, (2) *bootstrap-t*, and (3) the AR procedure. Since we can exactly replicate the point estimates for the articles in the replication sample, the differences in p-values are the result of the inferential methods used. Figure 4a–c plots reported and replicated p-values, from which we observed two patterns. First, most of the reported p-values are smaller than 0.05 or 0.10, the conventional thresholds for statistical significance. Second, consistent with Young's (2022) finding, our replicated p-values based on the bootstrap methods or AR procedure are usually bigger than the reported p-value (exceptions are mostly caused by rounding errors), which are primarily based on t-statistics calculated using analytic SEs. Using the AR test, we cannot reject the null hypothesis of no effect at the 5% level in 12 studies (17%), compared with 7 (10%) in the original studies. The number increases to 13 (19%) and 19 (27%) when we use p-values from the *bootstrap-t* and *-c* methods. Note that very few articles we review utilize inferential procedures specifically designed for weak instruments, such as the AR test (two articles), the conditional likelihood-ratio test (Moreira 2003) (one article), and confident sets (Mikusheva and Poi 2006) (none).

We also apply the *tF* procedure to 54 studies that use single IVs using $F_{\texttt{Eff}}$ statistics and t-statistics based on robust or cluster-robust SEs. Figure 4d shows that 19 studies (35%) are not statistically significant at the 5% level, and 7 studies (13%) deemed statistically significant when using the conventional fixed critical values for the t-test become statistically insignificant using the *tF* procedure, indicating that overly optimistic critical values due to weak instruments also contribute to overestimation of statistical

https://doi.org/10.1017/pan.2024.2 Published online by Cambridge University Press

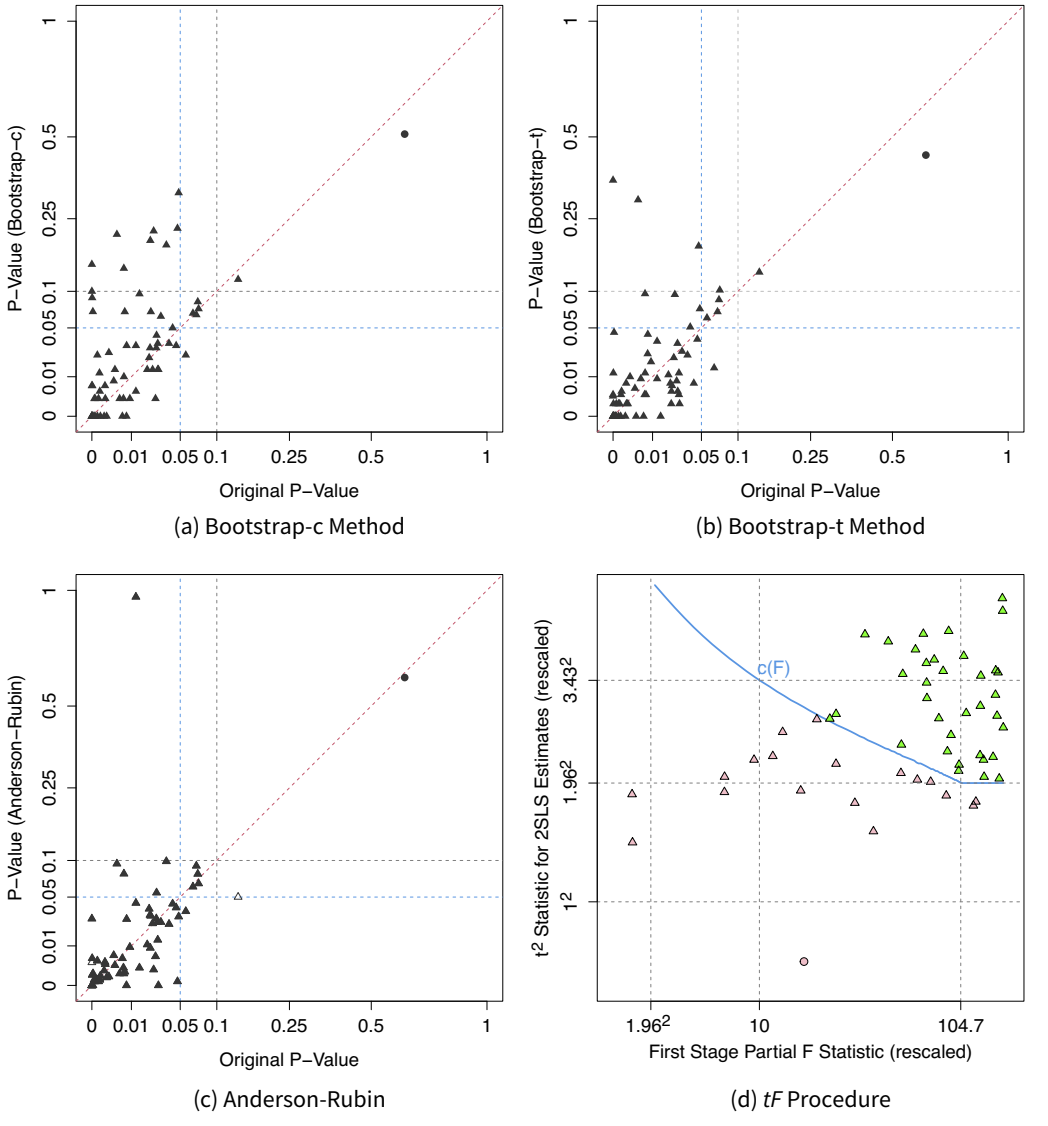


Figure 4. Alternative inferential methods. In subfigures (a)–(c), we compare original *p*-values to those from alternative inferential methods, testing against the null that $\tau = 0$. Both axes use a square-root scale. Original *p*-values are adapted from original articles or calculated using standard-normal approximations of *z*-scores. Solid circles represent Arias and Stasavage (2019), where the authors argue for a null effect using IV strategy. *Bootstrap-c* and *-t* represent percentile methods based on 2SLS estimates and *t*-statistics, respectively, using original model specifications. Hollow triangles in subfigure (c) indicate unbounded 95% CIs from the AR test using the inversion method. Subfigure (d) presents *tF* procedure results from 54 single instrument designs. Green and red dots represent studies remaining statistically significant at the 5% level using the *tF* procedure and those that do not, respectively. Subfigures (a)–(c) are inspired by Figure 3 in Young (2022), and subfigure (d) by Figure 3 in Lee *et al.* (2022).

power, but not as the primary factor. These results suggest that both weak instruments and non-i.i.d. errors have contributed to overstatements of power in IV studies in political science.

4.4. Finding 3. 2SLS–OLS Discrepancy

Finally, we investigate the relationship between the 2SLS estimates and naïve OLS estimates. In Figure 5a, we plot the 2SLS coefficients against the OLS coefficients, both normalized using reported OLS SEs.

https://doi.org/10.1017/pan.2024.2 Published online by Cambridge University Press

The shaded area indicates the range beyond which the OLS estimates are statistically significant at the 5% level. It shows that for most studies in our sample, the 2SLS estimates and OLS estimates share the same direction and that the magnitudes of the former are often much larger than those of the latter. Figure 5b plots the distribution of the ratio between the 2SLS and OLS estimates (in absolute terms). The mean and median of the absolute ratios are 12.4 and 3.4, respectively. In fact, in all but two designs (97%), the 2SLS estimates are bigger than the OLS estimates, consistent with Jiang's (2017) finding based on finance research. While it is theoretically possible for most OLS estimates in our sample to be biased toward zero, only 21% of the studies have researchers expressing their belief in downward biases of the OLS estimates. Meanwhile, 40% of the studies consider the OLS results to be their main findings. The fact that researchers use IV designs as robustness checks for OLS estimates due to concerns of upward biases is apparently at odds with the significantly larger magnitudes of the 2SLS estimates.

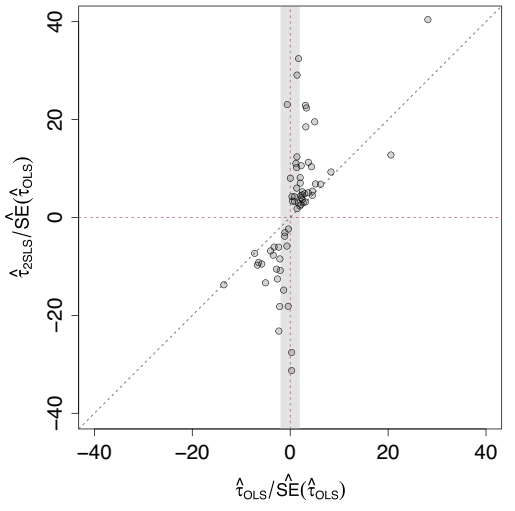
In Figure 5c, we further explore whether the 2SLS–OLS discrepancy is related to IV strength, measured by $|\hat{\rho}(d,\hat{d})|$. We find a strong negative correlation between $|\hat{\tau}_{2SLS}/\hat{\tau}_{OLS}|$ and $|\hat{\rho}(d,\hat{d})|$ among studies using nonexperimental instruments (gray dots). The adjusted R^2 is 0.264, with $p = 0.000$. However, the relationship is much weaker among studies using experiment-generated instruments (red dots). The adjusted R^2 is −0.014 with $p = 0.378$. At first glance, this result may seem mechanical: as the correlation between d and $\hat{d}$ increases, the 2SLS estimates naturally converge to the OLS estimates. However, the properties of the 2SLS estimator under the identifying assumptions do not predict the negative relationship (we confirm it in simulations in the SM), and such a relationship is not found in experimental studies. In Figure 5d, we limit our focus to the subsample in which the OLS estimates are statistically significant at the 5% level and researchers accept them as (part of) the main findings, and the strong negative correlation remains.

Several factors may be contributing to this observed pattern, including (1) failure of the exogeneity assumption, (2) publication bias, (3) heterogeneous treatment effects, and (4) measurement error in d. As noted earlier, biases originating from endogenous IVs or exclusion restriction failures can be magnified by weak instruments, that is, $\frac{|\text{Bias}_{IV}|}{|\text{Bias}_{OLS}|} = \left|\frac{\text{Cov}(z,\varepsilon)\mathbb{V}[d]}{\text{Cov}(d,\varepsilon)\text{Cov}(z,d)}\right| = \frac{|\rho(z,\varepsilon)|}{|\rho(d,\varepsilon)|\cdot|\rho(d,\hat{d})|} \gg 1$. In addressing large IV–OLS estimate ratios, Hahn and Hausman (2005) suggest two explanations: it could stem from a bias in OLS or from a bias in IV due to violations of the exogeneity assumption. Our empirical results, with particularly dubious IV to OLS estimate ratios in nonexperimental studies, seem to align with the latter explanation.

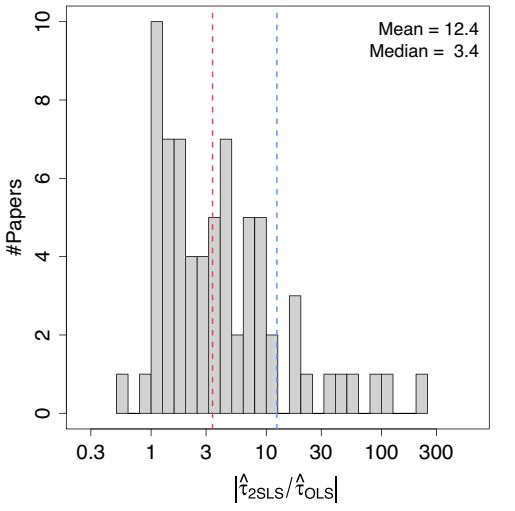
Publication bias may have also played a significant role. As shown in Figure 3, the variance of IV estimates increase as $|\hat{\rho}(d,\hat{d})|$ diminishes. If researchers selectively report only statistically significant results, or if journals have a tendency to publish such findings, it is not surprising that the discrepancies between IV and OLS estimates widen as the strength of the first stage declines, as shown in Figure 5c,d. This is because 2SLS estimates often need to be substantially larger than OLS estimates to achieve statistical significance. This phenomenon is known as Type-M bias and has been discussed in psychology and sociology literature (Felton and Stewart 2022; Gelman and Carlin 2014). Invalid instruments exacerbate this issue by providing ample opportunities for generating such large estimates.

Moreover, 30% of the replicated studies in our sample mention heterogeneous treatment effects as a possible explanation for this discrepancy. OLS and 2SLS place different weights on covariate strata in the sample, and therefore if compliers, those whose treatment status is affected by the instrument, are more responsive to the treatment than the rest of the units in the sample, we might see diverging OLS and 2SLS estimates. Under the assumption that the exclusion restriction holds, this gap can be decomposed into covariate weight difference, treatment-level weight difference, and endogeneity bias components using the procedure developed in Ishimaru (2021). In the SM, we investigate this possibility and find that it is highly unlikely that heterogeneous treatment effects *alone* can explain the difference in magnitudes between 2SLS and OLS estimates we observe in the replication data, that is, the variance in treatment effects needed for this gap is implausibly large.

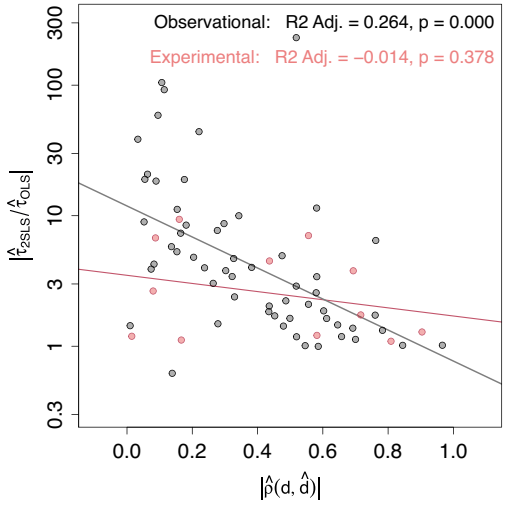
https://doi.org/10.1017/pan.2024.2 Published online by Cambridge University Press



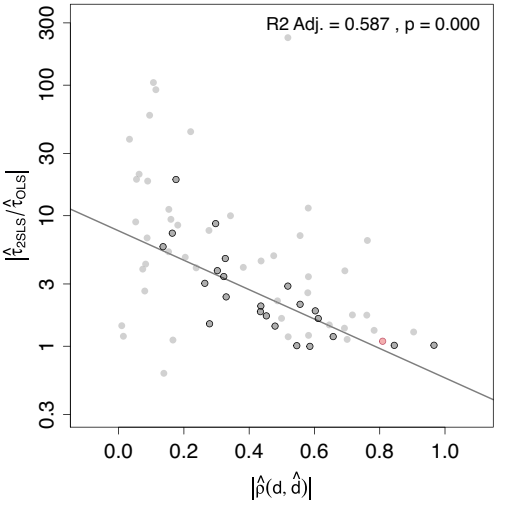
(a) OLS vs 2SLS Coefficients



(b) Distribution of Ratio of 2SLS and OLS Coefficients



(c) First Stage and the Ratio: Full Sample



(d) First Stage and Ratio: Subsample

Figure 5. Relationship between OLS and 2SLS estimates. In subfigure (a), both axes are normalized by reported OLS SE estimates with the gray band representing the $[-1.96, 1.96]$ interval. Subfigure (b) displays a histogram of the logarithmic magnitudes of the ratio between reported 2SLS and OLS coefficients. Subfigures (c) and (d) plot the relationship between $|\hat{\rho}(d, \hat{d})|$ and the ratio of 2SLS and OLS estimates. Gray and red circles represent observational and experimental studies, respectively. Subfigure (d) highlights studies with statistically significant OLS results at the 5% level, claimed as part of the main findings.

Finally, IV designs can correct for downward biases due to measurement errors in d, resulting in $|\hat{\tau}_{2SLS}/\hat{\tau}_{OLS}| > 1$. If the measurement error is large, this can weaken the relationship between d and $\hat{d}$, producing a negative correlation. We find it an unlikely explanation because only four articles (6%) attribute their use of IV to measurement errors, and the negative correlation is even stronger when we focus solely on studies where OLS estimates are statistically significant and regarded as the main findings.

https://doi.org/10.1017/pan.2024.2 Published online by Cambridge University Press

Table 3. Summary of replication results

	Experimental (12)	Observational (58)	All (70)
Panel A: First-stage F-statistic			
Unreported	.333	.138	.171
Effective $F < 10$	.083	.121	.114
Bootstrapped $F < 10$	.167	.172	.171
Median effective F	67.7	53.5	53.5
Panel B: Inference for 2SLS estimates			
Original $p > 0.05$	.250	.069	.100
AR $p > 0.05$	.417	.121	.171
Bootstrap-c $p > 0.05$	.417	.276	.300
Bootstrap-t $p > 0.05$	.333	.155	.186
tF procedure $p > 0.05$	.400	.341	.352
Panel C: 2SLS-OLS relationship			
$sign(\hat{\tau}_{2SLS}) = sign(\hat{\tau}_{OLS})$	1.00	.914	.929
$\lvert\hat{\tau}_{2SLS}/\hat{\tau}_{OLS}\rvert > 1$	1.00	.966	.971
$\lvert\hat{\tau}_{2SLS}/\hat{\tau}_{OLS}\rvert > 5$	.250	.362	.343
$\lvert\hat{\tau}_{2SLS}/\hat{\tau}_{OLS}\rvert > 10$	.000	.207	.171
Median $\lvert\hat{\tau}_{2SLS}/\hat{\tau}_{OLS}\rvert$	2.19	3.61	3.40

In Table 3, we present a summary of the main findings from our replication exercise. Observational studies, compared to experimental counterparts, generally have weaker first stages, often display larger increases in p-values when more robust inferential methods are used, and demonstrate bigger discrepancies between the 2SLS and OLS estimates. Based on these findings, we contend that a significant proportion of IV results based on observational data in political science either lack credibility or yield estimates that are too imprecise to offer insights beyond those provided by OLS regressions.

5. Recommendations

IV designs in experimental and observational studies differ fundamentally. In randomized experiments, the instruments' unconfoundedness is ensured by design, and researchers can address possible exclusion restriction violations at the design stage, for example, by testing potential design effects through randomization (Gerber and Green 2012, 140–141). Practices like power analysis, placebo tests, and preregistration also help reduce the improper use of IVs. In contrast, observational IV designs based on "natural experiments" require detailed knowledge of the assignment mechanism, making them more complex and prone to issues (Sekhon and Titiunik 2012).

Our findings suggest that using an IV strategy in observational settings is extremely challenging due to several reasons. First, truly random and strong instruments are rare and difficult to find. This is mainly because neither unconfoundedness nor the exclusion restriction is guaranteed by design, placing a greater burden of proof on researchers for the exogeneity assumption. Moreover, conducting placebo tests like the ZFS test for the exclusion restriction after data collection is not always feasible. Finally, increasing the sample size to achieve sufficient statistical power is often impractical. To prevent misuse of IVs in observational studies, we provide a checklist for researchers to consider when applying or contemplating an IV strategy with one endogenous treatment variable:

https://doi.org/10.1017/pan.2024.2 Published online by Cambridge University Press

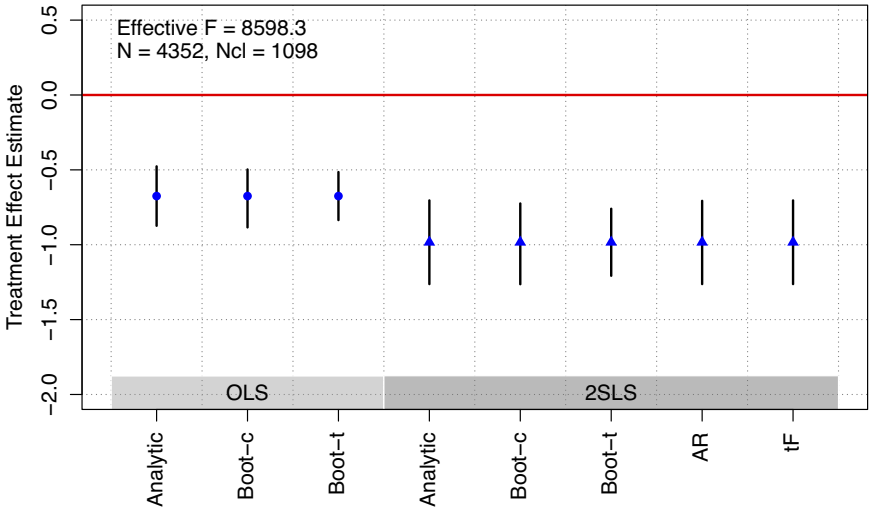


Figure 6. Replicated OLS and 2SLS estimates with 95% CIs (Rueda 2017, Table 5 column 1). The outcome is citizens' reports of voting buying. The treatment is the actual polling place size. The instrument is the size of the polling station predicted by the rules limiting the voters per polling station. The magnitude of the 2SLS estimate is slightly larger than that of the OLS estimate. Similar figures for each of the 70 IV designs are shown in the SM. This plot is made by ivDiag, an open-source package in R (Note to editor, please remove the Github link; please format the font of "R" properly).

Design

- Prior to using an IV strategy, consider how selection bias may affect treatment effect estimates obtained through OLS. If the main concern is underestimating an already statistically significant treatment effect, an IV strategy may be unnecessary.
- During the research design phase, consider whether the chosen instrument can realistically create random or quasi-random variations in treatment assignment while remaining excluded from the outcome equation.

Characterizing the first stage

- Calculate and report $F_{\texttt{Eff}}$ for the first stage, taking into account heteroscedasticity and clustering structure as needed. However, do not discard a design simply because $F_{\texttt{Eff}} < 10$.
- If both d and z are continuous, we recommend plotting d against its predicted values, $\hat{d}$, after accounting for covariates. Alternatively, plot both d and $\hat{d}$ against specific covariates that serve as the foundation for the rules used to derive the instruments. These visualizations are useful for detecting outliers and gaining insights into the sources of exogenous variation.

Hypothesis testing and inference

- *Option 1. t-test with $F_{\texttt{Eff}}$ pretesting.* If $F_{\texttt{Eff}} < 10$, choose Option 2 or 3. Utilize conservative methods like *bootstrap-t* and *bootstrap-c* if outliers or group structures are present.
- *Option 2. tF procedure.* For single treatment and instrument cases, adjust t-test critical values based on $F_{\texttt{Eff}}$.
- *Option 3. Direct testing.* Apply weak-instrument-robust procedures, such as the AR test.

Communicating your findings

- Present OLS and IV estimates alongside CIs from various inferential methods in a graphical format, like in Figure 6. These CIs may not concur on statistical significance, but they collectively convey the findings' robustness to different inferential approaches. In addition, they present the degree of uncertainty in both OLS and IV estimates in an intuitive manner.
- Remember to report first-stage and reduced-form estimation results, including 95% CIs for coefficients, as they offer insight into both instrument strength and statistical power.

https://doi.org/10.1017/pan.2024.2 Published online by Cambridge University Press

Additional diagnostics

- If you expect the OLS results to be upward biased, be concerned if the 2SLS estimator yields much larger estimates than OLS.
- If there is good reason to believe that treatment effects on compliers are significantly larger in magnitude than those on noncompliers, explain this through profiling of these principal strata (Abadie 2003; Marbach and Hangartner 2020).
- If it is possible to identify "never takers" or a subset of them, conduct a placebo test by estimating the effect of the instruments on the outcome in this ZFS sample. Using results from the ZFS test, obtain local-to-zero IV estimates and CIs and compare them to the original estimates. Section A3 of the SM provides detailed explanations and an empirical example.
- Conduct a sensitivity analysis as proposed by Cinelli and Hazlett (2022).

We provide an `R` package, ivDiag, to implement our recommended procedures. Stata tutorials for carrying out these procedures are also available on the corresponding author's website. Our aim is to address concerns regarding IVs in social science research and improve the quality of estimation and inference, especially for nonexperimental IV designs.

Acknowledgements. We thank Te Bao, Thomas Cao, Albert Chiu, Daniel Chen, Gary Cox, Charles Crabtree, Ted Enamorado, Hanming Fang, Avi Feller, Don Green, Justin Grimmer, Anna Grzymala-Busse, Jens Hainmueller, Guido Imbens, David Laitin, Adeline Lo, Justin McCrary, Jacob Montgomery, Doug Rivers, Henrik Sigstad, Brandon Stewart, Arthur Yu, and Xiang Zhou, as well as seminar participants at Stanford University, Washington University in St. Louis, APSA 2021, and Polmeth 2021, for their extremely helpful comments. We are also deeply grateful to four anonymous reviewers for PA and Editor Jeff Gill, as well as two anonymous reviewers for APSR—their invaluable suggestions have improved this paper significantly.

Data Availability Agreement. Replication data and code for this article have been published at Harvard Dataverse at https://doi.org/10.7910/DVN/MM5THZ.

Supplementary Material. For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2024.2.

References

Abadie, A. 2003. "Semiparametric Instrumental Variable Estimation of Treatment Response Models." *Journal of Econometrics* 113 (2): 231–263.

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge. 2020. "Sampling-Based Versus Design-Based Uncertainty in Regression Analysis." *Econometrica* 88 (1): 265–296.

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge. 2023. "When Should You Adjust Standard Errors for Clustering?" *The Quarterly Journal of Economics* 138 (1): 1–35.

Alt, J. E., J. Marshall, and D. D. Lassen. 2016. "Credible Sources and Sophisticated Voters: When Does New Information Induce Economic Voting?" *The Journal of Politics* 78 (2): 327–342.

Anderson, T. W., N. Kunitomo, and T. Sawa. 1982. "Evaluation of the Distribution Function of the Limited Information Maximum Likelihood Estimator." *Econometrica* 50 (4): 1009–1027.

Anderson, T. W., and H. 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, D. W. K., and P. Guggenberger. 2009. "Validity of Subsampling and 'Plug-in Asymptotic' Inference for Parameters Defined by Moment Inequalities." *Econometric Theory* 25 (3): 669–709.

Andrews, I., J. H. Stock, and L. Sun. 2019. "Weak Instruments in Instrumental Variables Regression: Theory and Practice." *Annual Review of Economics* 11 (1): 727–753.

Angrist, J., and M. Kolesár. 2023. "One Instrument to Rule Them All: The Bias and Coverage of Just-ID IV." *Journal of Econometrics* 240 (2): 105398.

Angrist, J. D., G. W. Imbens, and D. B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (434): 444–455.

Angrist, J. D., and J.-S. Pischke. 2009. *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press.

Arellano, M. 2002. "Sargan's Instrumental Variables Estimation and the Generalized Method of Moments." *Journal of Business & Economic Statistics* 20 (4): 450–459.

Arias, E., and D. Stasavage. 2019. "How Large Are the Political Costs of Fiscal Austerity?" *The Journal of Politics* 81 (4): 1517–1522.

https://doi.org/10.1017/pan.2024.2 Published online by Cambridge University Press

Baccini, L., and S. Weymouth. 2021. "Gone for Good: Deindustrialization, White Voter Backlash, and US Presidential Voting." *American Political Science Review* 115 (2): 550–567.
Betz, T., S. J. Cook, and F. M. Hollenbach. 2018. "On the Use and Abuse of Spatial Instruments." *Political Analysis* 26 (4): 474–479.
Bound, J., and D. A. Jaeger. 2000. "Do Compulsory School Attendance Laws Alone Explain the Association between Quarter of Birth and Earnings?" *Research in Labor Economics* 19 (4): 83–108.
Bun, M. J. G., and F. Windmeijer. 2010. "The Weak Instrument Problem of the System GMM Estimator in Dynamic Panel Data Models." *The Econometrics Journal* 13 (1): 95–126.
Cameron, A. C., J. B. Gelbach, and D. L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *The Review of Economics and Statistics* 90 (3): 414–427.
Chernozhukov, V., and C. Hansen. 2008. "The Reduced Form: A Simple Approach to Inference with Weak Instruments." *Economics Letters* 100 (1): 68–71.
Chiu, A., X. Lan, Z. Liu, and Y. Xu. 2023. "What to Do (and Not to Do) with Causal Panel Analysis under Parallel Trends: Lessons from a Large Reanalysis Study." https://ssrn.com/abstract=4490035.
Cinelli, Carlos, and Chad Hazlett. 2022. "An Omitted Variable Bias Framework for Sensitivity Analysis of Instrumental Variables." Available at https://ssrn.com/abstract=4217915.
Colin Cameron, A., and D. L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *The Journal of Human Resources* 50 (2): 317–372.
Conley, T. G., C. B. Hansen, and P. E. Rossi. 2012. "Plausibly Exogenous." *The Review of Economics and Statistics* 94 (1): 260–272.
Davidson, R., and J. G. MacKinnon. 2015. "Bootstrap Tests for Overidentification in Linear Regression Models." *Econometrics* 3 (4): 825–863.
Dieterle, S. G., and A. Snell. 2016. "A Simple Diagnostic to Investigate Instrument Validity and Heterogeneous Effects When Using a Single Instrument." *Labour Economics* 42: 76–86.
Dinas, E. 2014. "Does Choice Bring Loyalty? Electoral Participation and the Development of Party Identification." *American Journal of Political Science* 58 (2): 449–465.
Dorsch, M. T., and P. Maarek. 2019. "Democratization and the Conditional Dynamics of Income Distribution." *American Political Science Review* 113 (2): 385–404.
Dower, P. C., E. Finkel, S. Gehlbach, and S. Nafziger. 2018. "Collective Action and Representation in Autocracies: Evidence from Russia's Great Reforms." *American Political Science Review* 112 (1): 125–147.
Dube, O., and S. Naidu. 2015. "Bases, Bullets, and Ballots: The Effect of US Military Aid on Political Conflict in Colombia." *The Journal of Politics* 77 (1): 249–267.
Esarey, J., and A. Menger. 2019. "Practical and Effective Approaches to Dealing with Clustered Data." *Political Science Research and Methods* 7 (3): 541–559.
Felton, Chris, and Brandon M. Stewart. 2022. "Handle with Care: A Sociologist's Guide to Causal Inference with Instrumental Variables." https://osf.io/preprints/socarxiv/3ua7q.
Fieller, E. C. 1954. "Some Problems in Interval Estimation." *Journal of the Royal Statistical Society: Series B (Methodological)* 16 (2): 175–185.
Gelman, A., and J. Carlin. 2014. "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science* 9 (6): 641–651.
Gerber, A. S., and D. P. Green. 2012. *Field Experiments: Design, Analysis and Interpretation*. New York: W. W. Norton.
Greene, W. H. 2003. *Econometric Analysis*. Noida: Pearson Education India.
Hager, A., and H. Hilbig. 2019. "Do Inheritance Customs Affect Political and Social Inequality?" *American Journal of Political Science* 63 (4): 758–773.
Hahn, J., and J. Hausman. 2005. "Estimation with Valid and Invalid Instruments." *Annales d'Economie et de Statistique* 79/80: 25–57.
Hahn, J., and Z. Liao. 2021. "Bootstrap Standard Error Estimates and Inference." *Econometrica* 89 (4): 1963–1977.
Hainmueller, J., J. Mummolo, and Y. Xu. 2019. "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice." *Political Analysis* 27 (2): 163–192.
Hall, P., and J. L. Horowitz. 1996. "Bootstrap Critical Values for Tests Based on Generalized-Method-of-Moments Estimators." *Econometrica* 64 (4): 891–916.
Hansen, L. P. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50 (4): 1029–1054.
Henderson, J., and J. Brooks. 2016. "Mediating the Electoral Connection: The Information Effects of Voter Signals on Legislative Behavior." *The Journal of Politics* 78 (3): 653–669.
Hirano, K., and J. R. Porter. 2015. "Location Properties of Point Estimators in Linear Instrumental Variables and Related Models." *Econometric Reviews* 34 (6–10): 720–733.
Ishimaru, S.. 2021. "Empirical Decomposition of the IV-OLS Gap with Heterogeneous and Nonlinear Effects." *The Review of Economics and Statistics* 106 (2): 505–520.
Jiang, W. 2017. "Have Instrumental Variables Brought Us Closer to the Truth." *The Review of Corporate Finance Studies* 6 (2): 127–140.

https://doi.org/10.1017/pan.2024.2 Published online by Cambridge University Press

Key, E. M. 2016. "How Are We Doing? Data Access and Replication in Political Science." *PS: Political Science & Politics* 49 (2): 268–272.
Kim, J. H. 2019. "Direct Democracy and women's Political Engagement." *American Journal of Political Science* 63 (3): 594–610.
Lal, A., M. Lockhart, Y. Xu, and Z. Zu. 2024. "Replication Data for: How Much Should We Trust Instrumental Variable Estimates in Political Science? Practical Advice based on 67 Replicated Studies." Harvard Dataverse, V1. https://doi.org/10.7910/DVN/MM5THZ
Lee, D. S., J. McCrary, M. J. Moreira, and J. Porter. 2022. "Valid T-Ratio Inference for IV." *American Economic Review* 112 (10): 3260–3290.
Lorentzen, P., P. Landry, and J. Yasuda. 2014. "Undermining Authoritarian Innovation: The Power of China's Industrial Giants." *The Journal of Politics* 76 (1): 182–194.
Marbach, M., and D. Hangartner. 2020. "Profiling Compliers and Noncompliers for Instrumental-Variable Analysis." *Political Analysis* 28 (3): 435–444.
Mellon, J. 2023. "Rain, Rain, Go Away: 195 Potential Exclusion-Restriction Violations for Studies Using Weather as an Instrumental Variable." https://doi.org/10.31235/osf.io/9qj4f
Mikusheva, A., and B. P. Poi. 2006. "Tests and Confidence Sets with Correct Size When Instruments Are Potentially Weak." *The Stata Journal* 6 (3): 335–347.
Moreira, M. J. 2003. "A Conditional Likelihood Ratio Test for Structural Models." *Econometrica* 71 (4): 1027–1048.
Moreira, M. J. 2009. "Tests with Correct Size When Instruments Can Be Arbitrarily Weak." *Journal of Econometrics* 152 (2): 131–140.
Nelson, C., and R. Starz. 1990. "Some Further Results on the Exact Small Sample Properties of the Instrumental Variables Estimator." *Econometrica* 58 (41): 967–976.
Olea, J. L. M., and C. Pflueger. 2013. "A Robust Test for Weak Instruments." *Journal of Business & Economic Statistics* 31 (3): 358–369.
Rueda, M. R. 2017. "Small Aggregates, Big Manipulation: Vote Buying Enforcement and Collective Monitoring." *American Journal of Political Science* 61 (1): 163–177.
Sekhon, J. S., and R. Titiunik. 2012. "When Natural Experiments Are Neither Natural Nor Experiments." *American Political Science Review* 1: 35–57.
Sovey, A. J., and D. P. Green. 2011. "Instrumental Variables Estimation in Political Science: A Readers' Guide." *American Journal of Political Science* 55 (1): 188–200.
Staiger, D., and J. H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65 (3): 557–586.
Stock, J. H., J. H. Wright, and M. Yogo. 2002. "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments." *Journal of Business & Economic Statistics* 20 (4): 518–529.
Stock, J. H., and M. Yogo. 2005. "Asymptotic Distributions of Instrumental Variables Statistics with Many Instruments." In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, edited by D. W. K. Andrews and J. H. Stock, 109–120. Cambridge: Cambridge University Press.
Stommes, D., P. M. Aronow, and F. Sävje. 2023. "On the Reliability of Published Findings Using the Regression Discontinuity Design in Political Science." *Research & Politics* 10 (2): 1–12.
Vernby, K. 2013. "Inclusion and Public Policy: Evidence from Sweden's Introduction of Noncitizen Suffrage." *American Journal of Political Science* 57 (1): 15–29.
Young, A. 2022. "Consistency without Inference: Instrumental Variables in Practical Application." *European Economic Review* 147: 104–112.
Zhu, B. 2017. "MNCs, Rents, and Corruption: Evidence from China." *American Journal of Political Science* 61 (1): 84–99.

Cite this article: Lal, A., Lockhart, M., Xu, Y. and Zu, Z., (2024). How Much Should We Trust Instrumental Variable Estimates in Political Science? Practical Advice Based on 67 Replicated Studies. *Political Analysis*, 1–20. https://doi.org/10.1017/pan.2024.2

https://doi.org/10.1017/pan.2024.2 Published online by Cambridge University Press

Erratum

Page 9

Dower et al. (2018) use religious polarization as an instrument for the frequency of unrest and argue that religious polarization could only impact collective action through its impact on representation in local institutions.

It should be:

Dower et al. (2018) use religious polarization as an instrument for the frequency of unrest and argue that religious polarization could only impact representation in local institutions through collective action.

We apologize for the error misrepresenting the original paper.

Reference:

Dower, P. C., E. Finkel, S. Gehlbach, and S. Nafziger. 2018. "Collective Action and Representation in Autocracies: Evidence from Russia's Great Reforms." *American Political Science Review* 112 (1): 125–147.