How Much Should We Trust Instrumental Variable Estimates in Political Science? Practical Advice based on Over 60 Replicated Studies*

Apoorva Lal (Stanford) Mac Lockhart (UCSD)
Yiqing Xu (Stanford) Ziwen Zu (UCSD)

First Version: July 10, 2021
This Version: August 16, 2021

Abstract

Instrumental variable (IV) strategies are commonly used in political science to establish causal relationships, yet the identifying assumptions required by an IV design are demanding and it remains challenging for researchers to evaluate their plausibility. We replicate 61 papers published in three top journals in political science from the past decade (2011-2020) and document several troubling patterns: (1) researchers often miscalculate the first-stage F statistics, overestimating the strength of their IVs; (2) most researchers rely on classical asymptotic standard errors, which often severely underestimate the uncertainties around the two-stage-least-squared (2SLS) estimates; (3) in the majority of the replicated studies, the 2SLS estimates are much bigger than the ordinary-least-squared estimates, and their ratio is negatively correlated with the strength of the IVs in studies where the IVs are not experimentally generated, suggesting potential violations of the exclusion restriction; such a relationship is much weaker with experimentally generated IVs. To improve practice, we provide a checklist for researchers to avoid these pitfalls and recommend a zero-first-stage test and a local-to-zero procedure to guard against failure of the identifying assumptions.

Keywords: instrumental variables, two-stage-least-squared, replications, weak IV, exclusion restriction, zero-first-stage test

*Apoorva Lal, PhD Candidate, Stanford University; Email: apoorval@stanford.edu. Mac Lockhart, PhD Candidate, University of California, San Diego; Email: mwlockha@ucsd.edu. Yiqing Xu, Assistant Professor, Stanford University; Email: yiqingxu@stanford.edu; Ziwen Zu, PhD Student, University of California, San Diego; Email: zzu@ucsd.edu. We thank Te Bao, Gary Cox, Hanming Fang, Don Green, Justin Grimmer, Anna Grzymala-Busse, Jens Hainmueller, David Laitin, Avi Feller, Justin McCrary, Doug Rivers, and seminar participants at Stanford University and Polmeth 2021 for extremely valuable comments.

1. Introduction

The instrumental variable (IV) approach is a commonly used empirical method in the social sciences, including political science, to establish causal relationships. The IV approach has become a popular choice of applied researchers because of its broad applicability in estimating causal effects in both cross-sectional and panel settings where experimentation is infeasible or unethical, and rule-based assignments that make possible sharp regression-discontinuity designs are difficult to find. Since the publication the popular textbook *Mostly Harmless Econometrics* (Angrist and Pischke, 2008), which popularized the modern interpretation of IV designs, and Sovey and Green (2011), which clarifies the assumptions required by an IV approach and provides a useful checklist for political science researchers, over 100 papers that use IV as one of their main causal identification strategies have been published in three top journals in our discipline, the *American Political Science Review* (APSR), *American Journal of Political Science* (AJPS), and *Journal of Politics* (JOP).

Meanwhile, researchers also cast doubt on whether an IV estimate truthfully reveals a causal quantity of interest. Some complain that IV estimates are often considerably larger in magnitude than "naïve" OLS estimates even when the primary concern of the latter is upward selection bias.[1] A commonly used defense is that treatment effect heterogeneity is based on the LATE framework; the compliers in the study, units that whose value of the treatment x changes because of the instrument z, are more responsive to the treatment than the rest of the units in the sample—in fact, 19 (31%) papers in our replicated studies use this defense. However, the discrepancies in the effect magnitude are often too large to be justified by treatment effect heterogeneity and are often suggestive of an exclusion restriction violation instead (Hahn and Hausman, 2005). Some express serious concerns over whether

[1]For example, in the 2016 NBER Political Economy meeting, following a presentation of a study using an IV approach, the late political economist Alberto Alesina asked the audience: "How come IV estimates are always five times bigger than OLS estimates in political economy?"

the inferential methods being regularly used for IV estimation are valid (Young, 2017).

This observation motivates our systematic examination of the use of IVs in the empirical literature. We set out to replicate all papers published in the APSR, AJPS and JOP during the past decade (2011-2020) that use an IV design as one of the main identification strategies. We start with 115 papers, of which 65 have complete replication materials online, a finding that is a troubling pattern in its own right. Among the 65 papers, we successfully replicate at least of one of main IV results for 61 papers. Among them, three papers each have two separate IV designs, producing separate 2SLS results. Using data from these 64 IV designs, we conduct a programmatic replication exercise and find three troubling patterns. First, a large portion of the studies either do not report the first-stage partial F statistic or miscalculate it, for example, by not adjusting standard errors (SEs) for heteroskedasticity, serial correlation, or clustering structure. As a result, many published IV studies rely on what the econometrics literature calls "weak instruments" (Andrews, Stock and Sun, 2019).

Our second finding is on statistical inference: most studies we replicated rely on classical asymptotic SEs to quantify the uncertainties around the two-stage-least-squared (2SLS) estimates. Young (2017) shows that they severely underestimate the uncertainties in economic research and have led to over-rejection of the null hypotheses. We find a similar pattern: when replacing the reported SEs with bootstrapped SEs, in 26 of the 64 designs (41%), the 2SLS estimates become statistically insignificant at the 5% level while the number based on the SEs reported in the original papers is only 9 (14%).

Last but not least, our replications corroborate evidence from economics and finance that the 2SLS estimates are often much bigger in magnitude than the OLS estimates in political science (Jiang, 2017). In 59 out of the 64 designs (92%), the 2SLS estimates are bigger than OLS estimates in magnitude; among them, 21 (33%) are at least five times bigger. Even after we exclude 16 papers that explicitly claim to expect downward biases in OLS estimates, the percentages remain high (89% and 29%, respectively).

The first two patterns may be caused by researchers' unfamiliarity with IV methodology or under-utilization of statistical procedures for assumption-lean uncertainty quantification, such as bootstrapping. Therefore, researchers can avoid these problems by adopting better practice. The third finding, however, is the most concerning. We cannot explain it with weak IVs alone because when the other identifying assumptions are satisfied, weak IVs bias 2SLS estimates toward OLS estimates in finite samples (Bound, Jaeger and Baker, 1995), but what we observe is the opposite: the ratio between the magnitudes of the 2SLS and OLS estimates is strongly negatively correlated with the replicated partial correlation coefficient between the instrument and the treatment among studies that use non-experimental IVs; the relationship is weaker among studies with experimental IVs, though there may be too few data points to paint a definitive picture.

We suspect that this pattern is caused by a combination of weak IVs and the failure of the exclusion restriction, either because the IVs are not (conditionally) random or because they have directs effects on the outcome variables beyond the channel through the treatment. Intuitively, because the 2SLS estimator is a ratio, when the magnitude of the numerator is inflated due to the failure of the exclusion restriction while the magnitude of the denominator is a small number because of weak IVs, the 2SLS estimate explodes. Our finding points to unwarranted use of many IVs in observational research generated by what the authors claim to be "natural experiments" which are fundamentally different from real experiments (Sekhon and Titiunik, 2012) and seldom straightforward to analyze without detailed knowledge of the assignment mechanism.

What do these findings mean for the empirical literature using IV in political science? First, like Young (2017), we show that deflated SEs for the 2SLS estimates mask the fact most IV results are uninformative—because of the large SEs, their 95% confidence intervals often cover the OLS estimates and in many cases include zero. Second, and more importantly, many of the 2SLS estimates likely suffer from large biases due to failure of the identifying

assumptions, hence, are not credible. Although we cannot definitively say which estimates are problematic, the underlying issue seems to prevail in the literature.

The goal of this paper is not to discredit existing IV studies or dissuade researchers from ever using the IV approach. Quite on the contrary, our goals are two-fold: we want to caution researchers against the danger of justifying IVs in an ad-hoc fashion, especially in observational studies; and we want to provide researchers with practical advice that we hope can improve future practice. We recommend that authors correctly quantify the strength of IVs and uncertainties around 2SLS estimates and conduct an additional placebo test—the local-to-zero (LTZ) test—to corroborate instrument validity. We also provide sample code such that researchers can follow these steps easily in `R` or `Stata`.

We contribute to a growing literature that evaluates the use of IV strategies in empirical work across the social sciences and provides methods to improve empirical practice. Young (2017) replicates IV regressions from 32 articles published in economics journals and finds that IV estimates are more sensitive to outliers than their OLS counterparts, and that conventional IV significance tests systematically understate confidence intervals and therefore have high false discovery rates. Jiang (2017) surveys over 250 articles published in finance journals and finds that the vast majority of them report IV estimates that are larger than corresponding OLS estimates regardless of the sign of the potential omitted variables bias, and postulates that this bias can be attributed to exclusion restriction violations exacerbated by weak instruments, or local effects that are far from representative of the population treatment effect; however, the author does not conduct replications. Mellon (2020) proposes to use sensitivity analysis to quantify the vulnerability of using weather as IVs to exclusion restriction violations. Dieterle and Snell (2016) develop a quadratic over-identification test for the first stage 2SLS, apply it to 15 published papers that use linear first stage and find significant non-linearities in 10 of them, and suggest that evaluating the implied patterns of heterogeneity from their test relative to theoretical predictions can be used to gauge

the validity of an instrument. To the best of our knowledge, we are the first to link the discrepancy between IV and OLS estimates with the problem of weak IVs using large-scale replication data from the social sciences.

The organization of the paper is as follows. In Section 2, we briefly review the IV method under the traditional parametric framework, including the required identification assumptions, estimation strategies, and inferential methods. Section 3 details our case selection criteria and the resulting replication sample. Section 4 describes our replication procedure and presents the main results, Section 5 introduces diagnostic tools for potential exclusion restriction violations and illustrates them using a case study. The last section concludes the paper with a set of practical advice.

2. Theoretical Refresher

In this section, we offer a brief overview of the IV approach, including the setup, the identifying assumptions, as well as the point and variance estimators. We then discuss potential pitfalls. We follow a vast majority of IV studies in political science and adopt a textbook constant treatment effect linear setup for the IV approach. For simplicity, we do not include additional exogenous controls in the discussion without loss of generality. This is because, by the Frisch-Waugh-Lovell theorem, we can remove them by regressing the outcome, treatment, and instrument variables on the controls and using the residuals for all subsequent analyses.

Apart from addressing non-compliance in experimental encouragement designs, researchers often use IVs in observational studies in an attempt to convince readers that an observed correlation between a treatment variable x and an outcome variable y implies causality. The basic idea behind the method is to use an instrument z that isolates "exogenous" variation in x, i.e. the variation in x that is unrelated to potential confounders, in order to estimate its effect on y. The intuitiveness of the method masks the strong and untestable assumptions

that underlie its applications in most observational empirical settings. In this paper, we document patterns in the findings of IV designs in political science in the last decade, evaluate the plausibility of the assumptions that underpin their validity, and provide practical advice for researchers seeking to use IV designs [2].

Figure 1 shows a directed acyclic graph for an IV design, where ε represents the error term which captures all the unexplained variations in y. Because x and ε are correlated, an observed correlation between x and y does not identify the casual effect of x on y.

FIGURE 1. A DIRECTED ACYCLIC GRAPH OF AN IV DESIGN

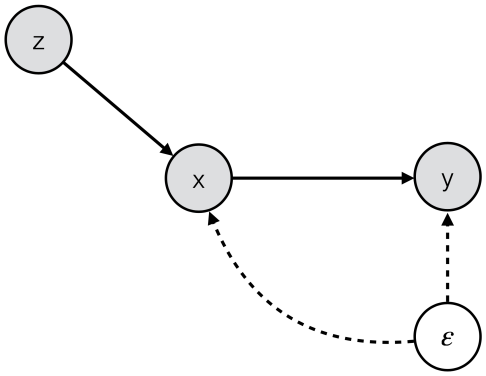


Figure 1 also illustrates that the IV approach relies on two key identifying assumptions: (1) z is correlated with x, or the *relevance* of the instrument, which is directly testable; and (2) z is uncorrelated with the error term, which implies two things: (i) z is quasi-randomly assigned, and (ii) it does not have a direct effect on y beyond the channel through x, referred to as the *exclusion restriction*.[3] As we will review in the next section, the traditional constant treatment effect framework for IVs formalizes these two assumptions with a set of parametric assumptions. On the other hand, the local average treatment effect (LATE) framework has a slightly different interpretation of these assumptions. It allows the treatment effect to vary by unit and adds an additional monotonicity assumption (Angrist, Imbens and Rubin,

[2]We also provide an `R` package to implement the statistical procedures in this paper at https://github.com/apoorvalal/ivDiag

[3]In this paper, we do not distinguish the instrument's (conditional) independence and the exclusion restriction because their failures lead to observational equivalent empirical results. In other words, the failure of the exclusion restriction may incorporate the failure of the conditional independence of the instrument.

1996). We adopt the former framework such that our replication effort can cover as many IV studies as possible in the political science literature.[4]

2.1. Estimation Strategies for the IV Approach

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

$$\text{Structural equation:} \quad y_i = \alpha + \beta x_i + \varepsilon_i \tag{2.1}$$

$$\text{First-stage equation:} \quad x_i = \pi_0 + \pi_1 z_i + \nu_i \tag{2.2}$$

in which y_i is the outcome for unit i ($i = 1, 2, \cdots, N$), x_i is a scalar treatment variable; z_i is a vector of instrument variables for x_i; β captures the (constant) treatment effect and is the key quantity of interest. Equations (2.1) and (2.2) are referred to as the structural equation and the first-stage equation. ε_i and ν_i represent idiosyncratic errors in each of the two equations and they may be correlated.

The identification problem of β arises when x_i and ε_i are correlated, which renders $\hat{\beta}_{OLS}$ from a naïve OLS regression of y_i on x_i inconsistent. The endogeneity problem may be due to one of the following reasons: (1) unmeasured omitted variables that are correlated with both y_i and x_i; (2) measurement error in x_i, and (3) simultaneity or reverse causality, which means y_i may also affect x_i. The IV approach addresses this problem by taking advantage of the exogenous variation in x_i brought by z_i. Substituting x_i in Equation (2.1) using Equation (2.2), we have the reduced form equation:

$$\text{Reduced form:} \quad y_i = \underbrace{(\alpha_0 + \beta\pi_0)}_{\gamma_0} + \underbrace{(\beta\pi_1)}_{\gamma_1} z_i + (\beta\nu_i + \varepsilon_i) \tag{2.3}$$

[4] 33 (57%) papers in our replicated studies employ continuous treatment variables and make no reference to treatment effect heterogeneity, hence, they are ill-suited for the LATE approach. Recent Marginal Treatment Effect (MTE) based methods (Heckman and Vytlacil, 2007; Mogstad, Santos and Torgovitsky, 2018; Mogstad and Torgovitsky, 2018) extend the LATE approach to cover such cases, but these methods have not yet been widely adopted in political science.

Since substitution establishes that $\gamma_1 = \beta\pi$, rearranging yields $\beta = \frac{\gamma_1}{\pi_1}$ (assuming that we only use one 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 β, we make the following assumptions (Imbens, 2014, Sec. 6).

Assumption 1 (Relevance): $\pi_1 \neq 0$.

This assumption requires that the instruments can predict the treatment variable.

Assumption 2 (Exclusion Restriction): $z_i \perp\!\!\!\perp \varepsilon_i$.

Because ε_i and ν_i may be correlated, this assumption implies that $z_i \perp\!\!\!\perp \nu_i$. In most realistic scenarios, it entails two requirements. First, the instruments are random or quasi-random conditional on exogenous covariates. In other words, there are no omitted variables in the first-stage and the IVs are not reversely affected by the treatment; moreover, conditional on the controls, the IVs are uncorrelated with unobservables in the structural equation. Second, it requires that the IVs do not affect the outcome except through the treatment. With Assumptions 1 and 2, we now review the most commonly used estimation strategies for an IV design.

The 2SLS estimator. We can write $\mathbf{x} = (x_1, x_2, \cdots, x_N)'$ and $\mathbf{y} = (y_1, y_2, \cdots, y_N)'$ as $(N \times 1)$ vectors of the treatment and outcome data, and $\mathbf{z} = (z_1, z_2, \cdots, z_N)'$ as $(N \times p_z)$ matrix of the instruments in which p_z is the number of instruments. To simplify mathematics, we residualize $\mathbf{x}$, $\mathbf{y}$, and each column of $\mathbf{z}$ against the exogenous covariates, obtaining $\mathbf{y}$, $\mathbf{x}$, and $\mathbf{z}$, respectively. The 2SLS estimator is written as follows:

$$\beta_{2SLS} = (\mathbf{x}'\mathbf{P}_z\mathbf{x})^{-1}\,\mathbf{x}'\mathbf{P}_z\boldsymbol{y} \tag{2.4}$$

in which $\mathbf{P}_z = \mathbf{z}\,(\mathbf{z}'\mathbf{z})^{-1}\,\mathbf{z}'$ is the hat-maker matrix from the first-stage which projects the endogenous treatment variable $\mathbf{x}$ into the column space of $\mathbf{z}$, thereby preserving only the exogenous variation in $\mathbf{x}$ that is uncorrelated with ε. This formula permits the use of more

than one 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 $\mathbb{E}[z_i\varepsilon_i] = 0$, including the two-step GMM (Hansen, 1982) and Limited Information Maximum Likelihood estimators (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 at least in one specification.

When the model is exactly identified, i.e., the number of treatment variables equals the number of instruments, the 2SLS estimator can be simplified as the IV estimator:

$$\widehat{\beta}_{2SLS} = \widehat{\beta}_{IV} = (\mathbf{z}'\mathbf{x})^{-1}\mathbf{z}'\mathbf{y} \tag{2.5}$$

In a case of one instrument variable z_i for one treatment variable x_i, the 2SLS estimator can also be written as a Wald estimator:

$$\widehat{\beta}_{2SLS} = \widehat{\beta}_{IV} = \hat{\beta}_{Wald} = \frac{\hat{\gamma}_1}{\hat{\pi}_1} = \frac{\widehat{\text{Cov}}(\mathbf{y},\mathbf{z})}{\widehat{\text{Cov}}(\mathbf{x},\mathbf{z})} \tag{2.6}$$

which clearly illustrates that the 2SLS estimator is a ratio between reduced-form and first-stage coefficients. This further simplifies to a ratio of difference in means when x and z are binary.

The 2SLS estimator is consistent and asymptotically normal, but has bad finite-sample properties and is biased toward the OLS result in small samples when identifying assumptions are invalid (Bound, Jaeger and Baker, 1995). The finite sample bias is seen clearly from the expectation of the simple IV estimator

$$\mathbb{E}\left[\widehat{\beta}_{IV}\right] = \mathbb{E}\left[\left(\sum_{i=1}^{N} z_i z_i'\right)^{-1}\sum_{i=1}^{N} z_i y_i\right] = \beta + \mathbb{E}\left[\left(\sum_{i=1}^{N} z_i x_i'\right)^{-1}\sum_{i=1}^{N} z_i \varepsilon_i\right]$$

The second term may not go to zero even when $\mathbb{E}[z_i\varepsilon_i] = 0$. This is because the denominator cannot pass through the conditional expectation as in the OLS case, since we are not con-

ditioning on $\mathbf{x}$, only on $\mathbf{z}$. This bias is decreasing in sample size, and is worsened by weak IVs and too many instruments. Hirano and Porter (2015) prove that unbiased estimation is infeasible for linear IV models with an unrestricted parameters space for the first-stage coefficients.

Inference. We use the IV estimator to illustrate why inference is more challenging with the 2SLS estimator than with the OLS estimator. A commonly used variance estimator for $\widehat{\beta}_{IV}$ can be written as:

$$\hat{\mathbb{V}}(\widehat{\beta}_{IV}) \approx \frac{\widehat{\sigma}^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2} \frac{1}{R^2_{xz}} = \hat{\mathbb{V}}(\widehat{\beta}_{OLS}) \frac{1}{R^2_{xz}} \tag{2.7}$$

in which $\widehat{\sigma}^2$ is a variance estimator for the error term and R^2_{xz} is the R-squared from the first-stage. The estimated variance is mechanically larger than the estimated variance of the OLS estimator as long as $R^2_{xz} < 1$. It is decreasing in R^2_{xz}, i.e. stronger instruments produce more precise IV estimates. Robust SEs can be computed using the IV analogue of the Huber-White sandwich formula (Bekker, 1994). We can also obtain uncertainty estimates for the 2SLS estimates using nonparametric bootstraps or block bootstraps.

2.2. Potential Pitfalls in IV Estimation

The main sources of difficulty in IV estimation and inference are failures of the two identifying assumptions. Their consequences include: (1) bias toward the OLS estimators when Assumption 1 fails but Assumption 2 is satisfied; (2) large uncertainty around the 2SLS estimate due to weak instruments; and (3) larger biases of the 2SLS estimates than the OLS estimates when both assumptions fail.

Weak instruments. Since the IV coefficient is a ratio in essence, the weak instruments problem is a "divide-by-zero" problem, which arises when $\text{Cov}(Z, X) \approx 0$. The instability of

ratio estimators like β_{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). Recently, Lee et al. (2020) show that the number needs to be as big as 104.7 for a conventional t-test to have a correct size.

The literature has discussed at least three issues caused by weak instruments. First, when Assumption 2 is valid, they exacerbate the finite bias of 2SLS estimator towards the inconsistent OLS estimator Second, the 2SLS estimates become very imprecise, which is shown in Equation (2.7). A third and related issue is that tests are of the wrong size and t-statistics don't follow a t-distribution (Charles and Starz, 1990). Weak instruments induce biases towards OLS estimates; this can be seen by examining the expectation of the IV estimator: $\mathbb{E}\left[\widehat{\beta}_{2SLS} - \beta\right] \approx \frac{\sigma_{\nu\varepsilon}}{\sigma^2_{\varepsilon}} \frac{1}{F+1}$, in which $\frac{\sigma_{\nu\varepsilon}}{\sigma^2_{\varepsilon}}$ is the bias from OLS and F is the first-stage F statistic (Angrist and Pischke, 2008, pp. 206-208). As a result, as $F \to 0$, the bias of the IV tends to the bias of the OLS, thereby reproducing the endogeneity problem that IV was meant to solve.

Issues relating to imprecision and test-statistic size arise from the fact that the distribution of $\widehat{\beta}$ is derived from its linear approximation of $\widehat{\beta}$ in $(\widehat{\gamma}, \widehat{\pi})$, wherein normality of the two OLS coefficients implies the normality of their ratio (Andrews, Stock and Sun, 2019). However, this normal approximation breaks down when $\widehat{\pi} \approx 0$. This approximation failure cannot be rectified by the bootstrap (Andrews and Guggenberger, 2009). Therefore, valid IV inference relies crucially on correctly identifying strong instruments.

Failure of the exclusion restriction. The validity of an IV design rests on the IV validity assumption (Assumption 2), which entails the conditional independence assumption and exclusion restriction. When the number of IVs is bigger than the number of endogenous variables, researchers can use an over-identification test to gauge its plausibility (Arellano,

2002). However, research has shown that such a test is often under-powered and has bad finite sample properties (Davidson and MacKinnon, 2015). In most cases, Assumption 2 is not directly testable. Because the consistency of 2SLS estimates depends on Assumption 2, especially the exclusion restriction, researchers usually spend a great amount of effort verbally arguing for its plausibility or by adding controls.

When combined with weak instruments, even small violations of the exclusion restriction can produce inconsistency. This is because plim $\widehat{\beta}_{IV} = \beta + \frac{\text{Cov}(Z,\varepsilon)}{\text{Cov}(Z,X)}$. When $\text{Cov}(Z, X) \approx 0$, even small violation of the exclusion restriction ($\text{Cov}(Z, \varepsilon) \neq 0$) can enlarge the second term, resulting in inconsistency. Thus, the two problems with instrumental variables exacerbate themselves: having weak instruments compounds problems from exclusion restriction violations, and vice versa.

With invalid instruments, it is likely that the asymptotic bias of the 2SLS estimator is greater than that of the OLS estimator. To see this, we write the ratio of biases of two estimators as follows (assuming $\hat{\beta}_{OLS}$ is inconsistent):

$$\frac{\text{plim } \widehat{\beta}_{IV} - \beta}{\text{plim } \widehat{\beta}_{OLS} - \beta} = \frac{\rho(Z, \varepsilon)}{\rho(Z, X)\rho(X, \varepsilon)} \tag{2.8}$$

in which $\rho(Z, \varepsilon)$, $\rho(Z, X)$ and $\rho(X, \varepsilon)$ are the correlation coefficients between Z and ε, between Z and X, and between X and ε, respectively. When $\rho(Z, X)$ in the denominator is small (i.e. when the instrument is weak), the magnitude of the ratio is likely to be large.

3. Data and Types of IVs

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

Data. We examined all empirical papers published in the APSR, AJPS, and JOP (printed versions) from 2011 to 2020 and identify studies that use an IV strategy as one of the main identification strategies, including papers that use binary or continuous treatments and that use a single or multiple IVs. Specifically, we use the following criteria: (1) the discussion of the IV result needs to appear in the main text and supports a main argument in the paper; (2) we consider linear models only; in other words, papers that use discrete outcome models are excluded from our sample; (3) we exclude papers that include multiple endogenous variables in a single specification because estimated treatment effects are correlated with one another (multiple endogenous variables in separate specifications are included); (4) we exclude papers that use IV or GMM estimators in a dynamic panel setting because they are subject to a separate set of empirical issues and their poor performance has been thoroughly discussed in the literature (Bun and Windmeijer, 2010). These criteria result in 38 papers in the APSR, 31 papers in the AJPS, and 46 papers in the JOP. We then strove to find replication materials for these papers from the public data sharing platforms, such as the Harvard Dataverse, and the authors' personal websites. We are able to locate replication materials for 70 (61%) papers.[5] However, code completeness and quality of documentation vary a great deal. Data availability has significantly improved since 2016-2017 following new editorial policies requiring authors to make replication materials publicly available, though none of the journal requires full replicability administrated by a third party as a condition for publication (Key, 2016), which would constitute a major improvement in our view.

Table 1. Data Availability and Replicability of Papers Using IVs

	#All Papers	Incomplete Data	Incomplete Code	Replication Error	Replicable
APSR	38	23	0	2	13 (34%)
AJPS	31	7	1	0	23 (74%)
JOP	46	17	2	2	25 (54%)
Total	115	47	3	4	61 (53%)

[5]Among them, two papers have incomplete data which make it impossible for us to replicate the results.

Using data and code from the replication materials, we set out to replicate the main IV results in the 70 papers. 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 replicating SEs, z-scores, or level of statistical significance for the 2SLS estimates because they involve the choice of a variance estimator, which we will discuss in the next section.

After much effort and hundreds of hours of work, we are able to replicate the main results of 61 papers.[6] The low replication rate is consistent with what is reported in Hainmueller, Mummolo and Xu (2019). The main reasons for failures of replication are incomplete data (47 papers), incomplete code or poor documentation (3 papers), and replication errors (4 papers). Table 1 presents summary statistics on data availability and replicability of IV papers for each of the three journals. The rest of this paper focuses on results based on these 61 replicable papers (and 65 IV designs).

Types of IVs. 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. As in Sovey and Green (2011), the biggest categories is "Theory," in which the authors justify validity of Assumption 1, including IVs' quasi-randomness and exclusion restriction, using social science theories or substantive knowledge. We further divide theory-based IVs into four subcategories: geography/climate/weather, history, treatment diffusion, and others.

Many studies in the theory category justify the choices of their IVs based on geography, climate, or weather conditions. For example, Zhu (2017) uses weighted geographic closeness as an instrument for the activities of multi-national corporations; Hager and Hilbig (2019)

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

use mean elevation and distance to rivers to instrument equitable inheritance customs; and Grossman, Pierskalla and Boswell Dean (2017*a*) use the number of distinct landmasses as an instrument for government fragmentation. Henderson and Brooks (2016) uses rainfall around Election Day as an instrument for democratic vote margins. The popularity of weather instruments for a whole host of outcomes necessarily implies that the exclusion restriction is especially tenuous in such cases (Mellon, 2020).

TABLE 2. TYPES OF IVS

IV Type	#Papers	Percentage%
Theory	40	62.5
Geography/climate/weather	10.5	16.4
History	10	15.6
Treatment diffusion	2.5	3.9
Others	17	26.6
Experiment	12	18.8
Rules & policy changes	5	7.8
Change in exposure	3	4.7
Fuzzy RD	2	3.1
Econometrics	7	10.9
Interactions/"Bartik"	5	7.8
Lagged treatment	1	1.6
Empirical test	1	1.6
Total	64	100.0

Note: One paper uses both geography-based instruments and an instrument based on treatment diffusion from neighbors. We count 0.5 for each category.

Historical IVs are based on historical differences between units that cannot be explained by current levels of the treatment. For example, Vernby (2013) uses historical immigration levels as an instrument for the current number of noncitizen residents. Similarly, Spenkuch and Tillmann (2018) use historical decisions by rulers in Europe over the religion of their region to instrument for the current religion of survey respondents. These studies use historical variation as instruments for current or modern variables.

Several studies base their choices on regional diffusion of treatment. For example, Dube and Naidu (2015) use US military aid to countries outside of Latin America as an instru-

ment for US military aid to Colombia. Grossman, Pierskalla and Boswell Dean (2017*b*) use over-time variation in the number of regional governments to instrument government fragmentation in sub-Saharan Africa. Dorsch and Maarek (2019*b*) use regional share of democracies as an instrument for democratization in a country-year panel.

Finally, a number of papers rely on a unique instrument based on theory that we could not place in a category. For example, Carnegie and Marinov (2017) use the rotating presidency of the Council of the European Union as an instrument for official development aid. They argue that countries that were colonized by the country that holds the presidency receive exogenously more aid than other countries. In another example, 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 it's impact on representation in local institutions. These papers employed IVs that were usually unique to the paper.

The second biggest category is randomized experiments. Articles in this category employ randomization, designed and conducted by researchers or a third party, to make causal inference and use IV estimation to address non-compliance issue in an encouragement design—the instrument normally is being encouraged to take the treatment. With random assignment, we have more confidence in the exclusion restriction because $z \perp\!\!\!\perp v$ by design, and the direct effect of an encouragement on the outcome is easier to rule out than without random assignment.[7]

Another category of IVs are based on explicit rules, which generate quasi-random variation in the treatment. Sovey and Green (2011) refer to this category "Natural Experiment." We avoid this terminology because it is widely misused (Sekhon and Titiunik, 2012). We limit this category to two circumstances: fuzzy regression discontinuity (RD) designs and variation in exposure to policies due to time of birth or eligibility.[8] For example, Kim (2019)

[7]We also include Healy and Malhotra (2013) in this category in which the instrument is a biological "lottery" of a newborn's gender.

[8]The difference between the two is subtle: for the latter, the gap in the forcing variable, such as birth cohort,

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. Respondents who were 18 as of election day could vote while those who were 17 could not, making this a problem of one sided non-compliance as some respondents who were over 18 did not vote.

The last category of IVs are based on econometric assumptions. This category includes what Sovey and Green (2011) called lags. These are econometric transformations of variables the author argues constitute an instrument. For example, Lorentzen, Landry and Yasuda (2014) use a measure of the independent variable from 8 years earlier to mitigate endogeneity concerns. Another example of this are instruments that rely on a transformation of variables to satisfy the assumptions. For example, Dorsch and Maarek (2019*a*) use the sum of neighboring countries with similar institutions that are democracies as an instrument for democratization within a country. Shift-share "Bartik" instruments that are based on interactions between multiple variables are also included in this category (Goldsmith-Pinkham, Sorkin and Swift, 2020).

Compared with IV papers published before 2010, there is a significant increase in the proportion of papers using experiment-generated IVs (from 2.9% to 18.8%) thanks to an increased popularity of survey and field experiments in the field. In stark contrast, the number of papers relying on econometric techniques or flawed empirical tests (such as regression y on x and z in one regression and check whether the coefficient of z is significant) to justify potential IVs has decreased thanks to improving empirical practice in the discipline. The percentage of papers using theory-justified IVs remain almost exactly the same around 60%.

4. Replication Procedure and Results

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

is fixed and cannot be arbitrarily small.

Procedure. For each paper, we select a 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 asymptotic SEs, (2) Huber White robust SEs, (3) cluster-robust SEs if there is a clustering structure according to the authors, and (4) bootstrap SEs. For example, the bootstrap SEs are calculated by

$$F_{boot} = \hat{\beta}'_{2SLS} \widehat{\text{Var}}_{boot}(\hat{\beta}_{2SLS})^{-1} \hat{\beta}_{2SLS}/p_z$$

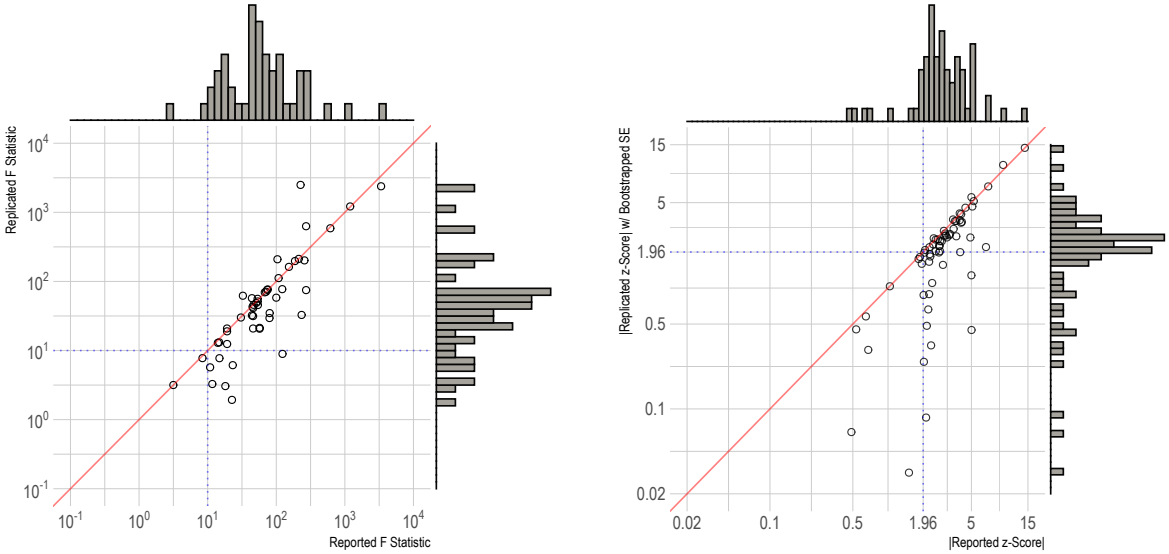
where p_z is the number of IVs and $\widehat{\text{Var}}_{boot}(\hat{\beta}_{2SLS})$ is the estimated variance-covariance matrix based on a nonparametric bootstrap procedure, in which we sample the rows of the data matrix with replacement.[9] If the data have a clustered structure, we use block bootstrapping instead by sampling with replacement each cluster of data (Colin Cameron and Miller, 2015; Esarey and Menger, 2019). In Supplementary Materials (SM), we compare the four types of F statistics. Because bootstrap and cluster-bootstrap SEs are usually more conservative than other SE estimators, we use the partial F statistic based on these SEs in most comparisons.

Second, we replicate the original IV result using the 2SLS estimator with both analytic SEs and bootstrapped SEs. We record the point estimates, the SEs, and z-scores. We also estimate a naïve OLS model by regressing the outcome variable on the treatment and control variables, leaving out the instrument. We then calculate the ratio between the magnitudes of the 2SLS and OLS estimates. We also record other useful information, such the number of observations, the number of clusters, the types of IVs, the way SEs are calculated, and the rationale each paper uses to justify its IV strategy.

[9] In a just-identified case where there is only one instrument, the F statistic is equivalent to the square of the t-statistic for the coefficient on the instrument in the first stage regression based on the bootstrap SE.

Results. We present three main findings based on our replication data. First, we study the strengths of the IVs by replicating the first-stage partial F statistics. To our surprise, among the 64 IV designs, 14 (22%) do not report this crucial statistic despite its key role in justifying the validity of an IV design. Among the remaining 50 studies that report F statistics, 10 (20%) use classic asymptotic SEs, thus not adjusting for potential heteroskadasticity or clustering structure. In Figure 2(a), we plot the of reported and replicated first-stage partial F statistic (on logarithmic scales) for these studies and overlay it their respective histograms on the top and to the right. Because the authors use a variety of SE estimators, such as the

FIGURE 2. DISTRIBUTIONS OF F STATISTICS AND z SCORES: REPORTED VS. REPLICATED



(a) First-Stage Partial F-Statistic

(b) z-Score for 2SLS Estimates

Huber White robust SE and the cluster-robust SEs, in the original studies, our replicated F statistics based on (clustered) bootstrap SEs are sometimes larger (15 studies, or 30%) and sometimes smaller (35 studies, or 70%) than the reported ones.[10] However, among the 12 studies that have replicated F statistics smaller than 10, three do not report the F

[10] As shown in the Supplementary Material, F statistics based on classic asymptotic SEs can sometimes be smaller than F statistics based on bootstrap SEs; however, this rarely happens when there is a clustering structure in data.

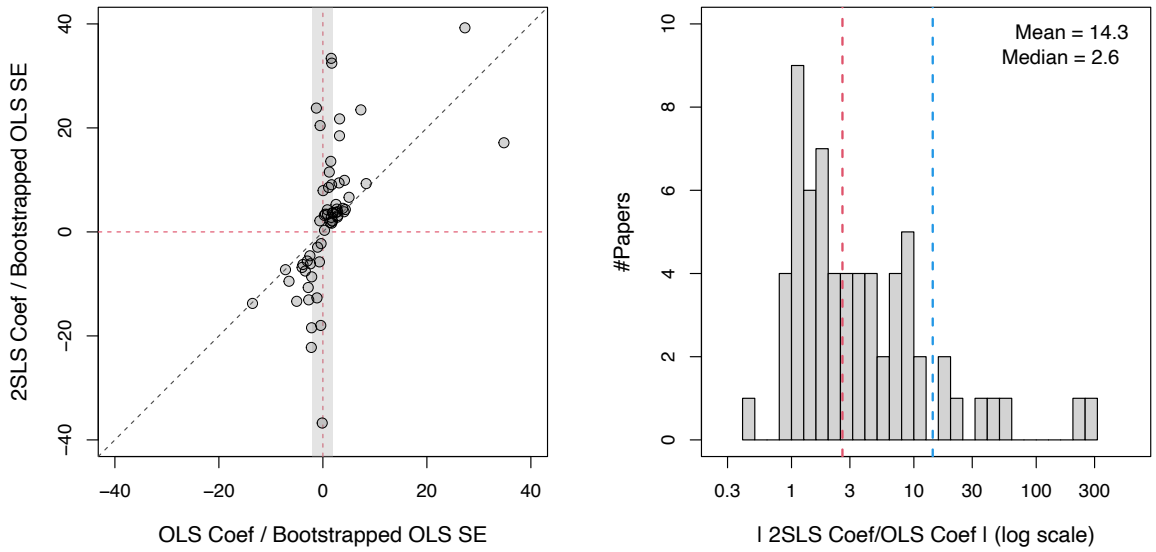
statistics and seven report F statistics bigger than 10 in the original paper. Although 81% of the studies have replicated $F \geq 10$, the number decreases to only 31% if we impose a more stringent requirement in order for a conventional t-test to have a correct test size, i.e., $F \geq 104.7$ (Lee et al., 2020) (see Table 3 for a breakdown by journal).

By comparing the F statistics based on different SE estimators in SM, we demonstrate that the overestimation of the F statistics when the instruments are weak is primarily caused by underestimation of uncertainties in studies with clustered data structure. It is well known that by either not clustering SE at a proper level or using the asymptotic clustered SE with too few clusters, researchers risk severely overstating the level of statistical significance (Cameron, Gelbach and Miller, 2008), but so far the same problem has received little attention when it comes to evaluating weak instruments using the F statistic.

Next, we compare the reported and replicated z-scores for the 2SLS estimates, which is defined as $z = \hat{\beta}_{2SLS}/\widehat{SE}(\hat{\beta}_{2SLS})$. Because we can replicate the point estimates for the papers in the replication sample, the only differences are the SE estimates. Figure 2(b) plots reported and replicated z-scores (on logarithmic scales), from which we observed two patterns. First, there is clear evidence of "p-hacking": reported z-scores "bunch" around 1.96, corresponding to 5% statistical significance in a standard two-sided tests for a normally distributed test score. Second, consistent with Young (2017)'s finding, our replicated z-scores based on bootstrapped SEs are smaller than reported z-scores, which are primarily based on asymptotic SEs. Using the replicated z-scores, estimates in 26 studies (41%) become statistically insignificant at the 5% level, compared with 9 (14%) in the original studies.

Lee et al. (2020) argue that if we maintain $F \geq 10$ as a criterion for sufficiently strong instruments, which they deem too lenient, z-scores need to be greater than 3.43 to have a test size of 5%; only 14 studies (21.9%) meet this requirement. Moreover, very few papers resort to tests or methods specifically designed for weak instruments, such as the Anderson-Rubin test (2 papers), the conditional likelihood-ratio test (Moreira, 2003) (1 paper), and

FIGURE 3. RELATIONSHIP BETWEEN OLS AND 2SLS ESTIMATES



(a) Scatterplot (b) Ratio Histogram

constructs confident sets (Mikusheva and Poi, 2006) (none).

Finally, we investigate the relationship between the 2SLS estimates and naïve OLS estimates. In Figure 3(a), we plot the 2SLS estimates against the OLS estimates, both of which are normalized using bootstrapped SEs for the OLS estimates. The shaded area indicates the range within which the OLS estimates are statistically significant at the 5% level. Figure 3(a) shows that for most studies in our sample, the 2SLS estimates and OLS estimates share the same sign (94%) and that the magnitudes of the 2SLS estimates are often much bigger than those of the OLS estimates. Figure 3(b) plots the distribution of the ratio between the 2SLS and OLS estimates (in absolute terms). The mean and median of the absolute ratios are 14.3 and 2.6, respectively. In fact, in 59 of the 64 designs (92%), the 2SLS estimates are bigger than the OLS estimates, consistent with Jiang (2017)'s finding based on finance research. While this is theoretically possible when omitted variables bias biases towards zero, it is difficult to evaluate whether this is a valid explanation because researchers seldom state their beliefs regarding the sign of the bias in OLS.

We then explore whether this ratio is related to the strength of the instruments, captured by $|\rho(Z,X)|$. Because plim $\widehat{\beta}_{2SLS} = \beta + Bias_{IV}$ and plim $\widehat{\beta}_{OLS} = \beta + Bias_{OLS}$, we have

$$\text{plim} \left| \frac{\widehat{\beta}_{2SLS} - \widehat{\beta}_{OLS}}{\widehat{\beta}_{OLS}} \right| = \left| \frac{Bias_{2SLS}}{\beta + Bias_{OLS}} \right| \leq \left| \frac{Bias_{2SLS}}{Bias_{OLS}} \right| = \left| \frac{\rho(Z,\varepsilon)}{\rho(X,\varepsilon)} \right| \cdot \frac{1}{|\rho(X,\hat{X})|}, \tag{4.1}$$

in which $\hat{X}$ is the predicted value of X in the first-stage. The inequality holds in (4.1) if β and $Bias_{OLS}$ are of the same sign (i.e., positive selection). Positive selection is why many researchers adopt an IV strategy in the first place.[11] Because $\left| \frac{\widehat{\beta}_{2SLS} - \widehat{\beta}_{OLS}}{\widehat{\beta}_{OLS}} \right|$ is bounded by $\left| \frac{\rho(Z,\varepsilon)}{\rho(X,\varepsilon)} \right| \cdot \frac{1}{|\rho(X,\hat{X})|}$, when exclusion restriction holds, i.e., $\rho(Z,\varepsilon) = 0$, in theory, there should be no relationship between $\left| \frac{\widehat{\beta}_{2SLS} - \widehat{\beta}_{OLS}}{\widehat{\beta}_{OLS}} \right|$ and $|\rho(X,\hat{X})|$. However, when we plot their relationship in Figure 4(a) using the replication data , we find a strong negative correlation among IV designs using non-experimental IVs (gray dots; adjusted $R^2 = 0.274$); the relationship is much weaker among studies using experiment-generated IVs (red dots; adjusted $R^2 = 0.086$).[12] Even after we limit our focus to the subsample in which the reported OLS estimates are statistically significant at 5% and share the same signs with the 2SLS estimates, the negative correlation remains (Figure 4(b)). This means that even when researchers do not show explicit concerns of their OLS results and use IVs as robustness checks, the OLS-IV discrepancy is related to IV strength. These results lend strong support to our conjecture that the large discrepancies between 2SLS and OLS estimates are due to a combination

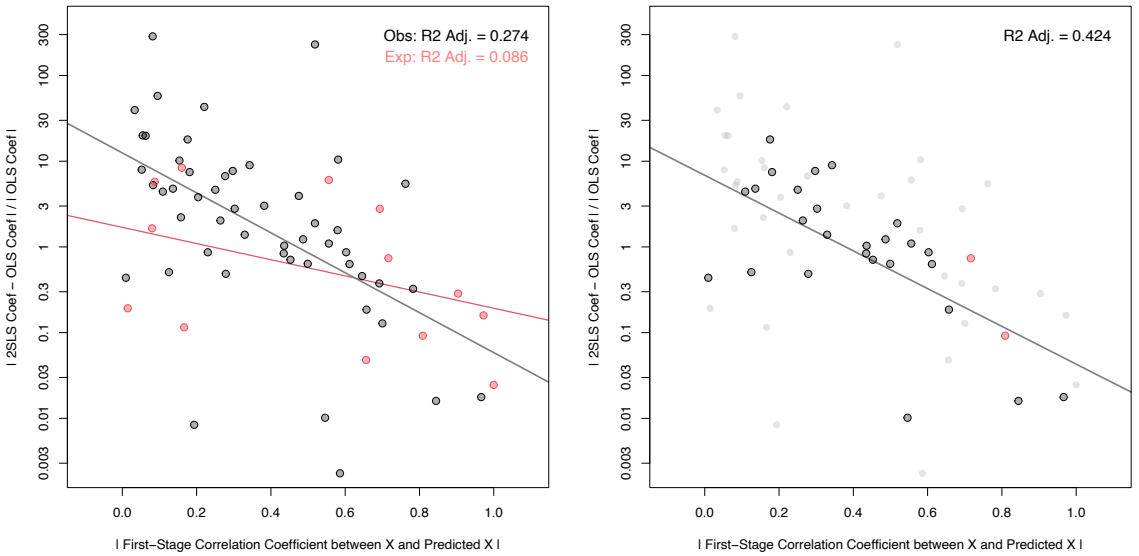
[11] When the treatment effects are heterogeneous and the local average treatment effect (LATE) does not equal the average treatment effect (ATE), the above derivation becomes:

$$\text{plim} \left| \frac{\widehat{\beta}_{IV} - \widehat{\beta}_{OLS}}{\widehat{\beta}_{OLS}} \right| \leq \text{plim} \left| \frac{\Delta_{LATE}}{\widehat{\beta}_{OLS}} \right| + \text{plim} \left| \frac{Bias_{IV}}{\widehat{\beta}_{OLS}} \right| \leq \left| \frac{\Delta_{LATE}}{\beta + Bias_{OLS}} \right| + \left| \frac{\rho(Z,\varepsilon)}{\rho(X,\varepsilon)} \right| \cdot \frac{1}{|\rho(X,\hat{X})|},$$

in which Δ_{LATE} is the difference between the LATE and ATE. $\left| \frac{\Delta_{LATE}}{\beta + Bias_{OLS}} \right|$ may be decreasing in the strength of the instrument as well because as the instrument predicts more variation in the treatment, the LATE converges to the ATE. However, it is highly unlikely that treatment effect heterogeneity can explain the difference in magnitudes between 2SLS and OLS estimates we observe in the replication data. See SM for additional simulation results.

[12] The negative relationship can be explained by exclusion restriction violations (possible, but less likely), compilers being more responsive to treatments, or publication bias.

FIGURE 4. IV STRENGTH AND OLS-IV DISCREPANCY



(a) Full Sample (b) Subsample with Significant OLS Results

of failures of the identifying assumptions, namely, biases caused by exclusion restriction violations are amplified by weak IVs.[13]

We summarize our main findings from the replication study in Table 3.

5. Diagnostic Tools for Exclusion Restriction Violations

The problems of overestimating IV strength and understating uncertainties around 2SLS estimates can be alleviated by employing more conservative uncertainty estimators, such as those based on bootstrapping. The exclusion restriction failure, however, is much harder to address. This is because it is inherently a research design issue and should ultimately be tackled at the research design stage. In this section, we provide a set of diagnostic tools to help research gauge the validity of Assumption 2. These tests are most well-suited for observational studies where justifying the exclusion restriction remains challenging.

[13] Another possibility is publication bias, namely, because IV estimates tend to have large standard errors, when the estimates are small and statistically insignificant, they end up not getting published. We thank Hanming Fang for raising this point.

Table 3. Summary of Replication Results

(%)	APSR (13)	AJPS (23)	JOP (25)	All (61)
Panel A				
First-stage F Statistic Unreported	13.3	20.8	28.0	21.9
F Statistic Miscalculated	15.4	47.4	50.0	40.0
Replicated F Statistic $\geq$ 10.0	86.7	79.2	80.0	81.2
Replicated F Statistic $\geq$ 104.7	33.3	37.5	24.0	31.2
Panel B				
Reported SE $<$ 0.8$*$Bootstrapped SE	26.7	33.3	40.0	34.4
Reported SE $<$ 0.5$*$Bootstrapped SE	26.7	20.8	28.0	25.0
Reported SE $<$ 0.3$*$Bootstrapped SE	13.3	16.7	12.0	14.1
Replicated $\|z\| \geq 1.96$	53.3	62.5	60.0	59.4
Replicated $\|z\| \geq 3.43$	33.3	29.2	8.0	21.9
Panel C				
$\|\hat{\beta}_{2SLS}/\hat{\beta}_{OLS}\| > 1$	86.7	91.7	96.0	92.2
$\|\hat{\beta}_{2SLS}/\hat{\beta}_{OLS}\| > 3$	46.7	41.7	52.0	46.9
$\|\hat{\beta}_{2SLS}/\hat{\beta}_{OLS}\| > 5$	33.3	33.3	32.0	32.8
$\|\hat{\beta}_{2SLS}/\hat{\beta}_{OLS}\| > 10$	13.3	16.7	16.0	15.6

Note: "F Statistic Miscalculated" is based on a subset of studies that report first-stage F statistics and defined as reported first-stage partial F being at least 30% smaller than the bootstrapped F statistics. Lee et al. (2020) show that, to have a correct test size at 5%, researchers need to either require$F \geq 104.7$ or $|z| \geq 3.43$.

Zero-first-stage tests. The exclusion restriction is a strong and generally untestable assumption that underlies the validity of the instrument; indeed, researchers typically spend considerable effort in papers and seminars arguing that the condition is satisfied in their particular setting. However, some placebo tests have recently become popular as a way to argue for instrument validity, especially in observational settings where the choice of instrument is guided by detailed domain knowledge. Bound and Jaeger (2000) first suggest using an auxiliary regression on a subsample where the instrument is not expected to influence treatment assignment, known as "zero-first-stage" (ZFS) tests. The primary intuition is that in a subsample that one has a strong prior that the first stage is zero, the reduced form effect should also be zero if the exclusion restriction is satisfied. In other words, motivated by a substantive prior that the first-stage effect of the instrument is likely zero for a subsample of the population (henceforth, the "ZFS subsample"), the researcher than proceeds to show

that the reduced-form coefficient for the instrument (by regression Y on Z) is approximately zero *in the ZFS subsample*, which is suggestive evidence in favor of instrument validity. Most observational instruments ought to yield some a ZFS subsample based on substantive knowledge of the assignment mechanism.

This style of placebo is particularly popular in studies of historical political economy, where particular historical or geographic features are argued to be valid instruments for treatment assignment, and thus they are unlikely to be driving treatment assignment outside of a specific context. For example, Nunn (2008) studies the effects of the slave-trade on modern-day development in Africa using sailing distance from each country to the nearest locations of demand for slave labor as an instrument for normalised number of slaves taken. The author then argues that distance to demand locations in the New World are likely to be a valid instrument by using a placebo test that the first stage effect (the instrument regressed on the outcome, modern day GDP) is approximately zero for countries outside Africa, where the posited mechanism (that places close to demand locations exported more slaves only in the transatlantic slave trade) has no traction, thereby providing a candidate ZFS sample. In a related paper, Nunn and Wantchekon (2011) use the same strategy to show that distance to slave-trade ports does not predict modern-day trust attitudes in the Asiabarometer, while they do in the Afrobarometer (which is the primary study population).

While this is a useful heuristic check that we advise most observational IV papers adopt, it is an informal test and provides no test statistics. van Kippersluis and Rietveld (2018) demonstrate that the ZFS test can be fruitfully combined with the "plausibly exogenous" method suggested by Conley, Hansen and Rossi (2012) (henceforth, CHR 2012).[14] To illustrate the method, we first rewrite the IV simultaneous equations in CHR (2012)'s notation:

[14]We do not use other strategies proposed by CHR (2012), such as union of confidence intervals, because they require more use discretion.

$$Y = X\beta + Z\gamma + \varepsilon; \quad X = Z\Pi + \nu, \tag{5.1}$$

where the instrument Z also enters the structural equation, the exclusion restriction amounts to a dogmatic prior that $\gamma = 0$. CHR (2012) suggest that this assumption can be relaxed, and replaced with a user-specified assumption on a plausible value, range, or distribution for γ depending on the researcher's beliefs regarding the degree of exclusion restriction violation. They propose three different approaches for inference that involve specifying the range of values for γ, a prior distributional assumption for γ, and a fully-Bayesian analysis that requires priors over all model parameters and corresponding parametric distributions. We focus on the second method, which CHR (2012) call the "local to zero" (LTZ) approximation, which considers "local" violations of the exclusion restriction [15] and requires a prior over γ alone. CHR (2012) show that replacing the standard assumption that $\gamma = 0$ with the weaker assumption that $\gamma \sim \mathbb{F}$ implies distribution for $\widehat{\beta}$ in Equation (5.2).

$$\widehat{\boldsymbol{\beta}} \sim^a \mathcal{N}\left(\boldsymbol{\beta}, \mathbb{V}_{2SLS}\right) + \mathbf{A}\gamma \;\text{ where } \mathbf{A} \equiv (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z} \tag{5.2}$$

$$\widehat{\boldsymbol{\beta}} \sim^a \mathcal{N}\left(\boldsymbol{\beta} + \mathbf{A}\mu_\gamma, \mathbb{V}_{2SLS} + \mathbf{A}\boldsymbol{\Omega}\mathbf{A}'\right) \tag{5.3}$$

where the original 2SLS asymptotic distribution is inflated by the additional term. While a simulation-based approach can be used to implement Equation (5.2) for an arbitrary distribution for γ, the distribution takes its most convenient form when one uses a Gaussian prior over $\gamma \sim \mathcal{N}\left(\mu_\gamma, \Omega_\gamma\right)$, which simplifies Equation (5.2) to Equation (5.3), with a posterior being a Gaussian centered at $\boldsymbol{\beta} + \mathbf{A}\mu_\gamma$.

CHR (2012) suggest that researchers use domain knowledge to choose $\mu_\gamma, \Omega_\gamma$, since they often hold strong priors about instruments anyway (which presumably motivates the choice of instrument). van Kippersluis and Rietveld (2018) suggest that a principled method to

[15] $\gamma = C/\sqrt{N}$ for some constant C and sample size N

choose μ_γ is to estimate Equation (5.1) on the ZFS population (wherein Π is assumed to be zero), and use this estimate $\widehat{\gamma}_{ZFS}$ as μ_γ. This approach combines the informal ZFS test with the plausibly-exogenous method in a straightforward manner, and software to implement it is available in both `R` (accompanying this paper) and `STATA` (Clarke, 2014). We provide a simulation-based illustration of the method in the supplementary appendix and illustrate the application of this method to a published empirical paper next.

A case study. We illustrate the diagnostics described above by applying it to the instrumental-variables analysis in Guiso, Sapienza and Zingales (2016) (henceforth GSZ 2016), who re-visit Leonardi, Nanetti and Putnam (2001)'s celebrated conjecture that Italian cities that achieved self-government in the middle-ages have higher modern-day levels of social capital. More specifically, they study the effects of free-city state status on social capital as measured by the number of non-profits and organ donation per capita, and a measure of whether students cheat in mathematics.

TABLE 4. REPLICATION OF GSZ (2016) TABLE 6
REDUCED FORM REGRESSIONS

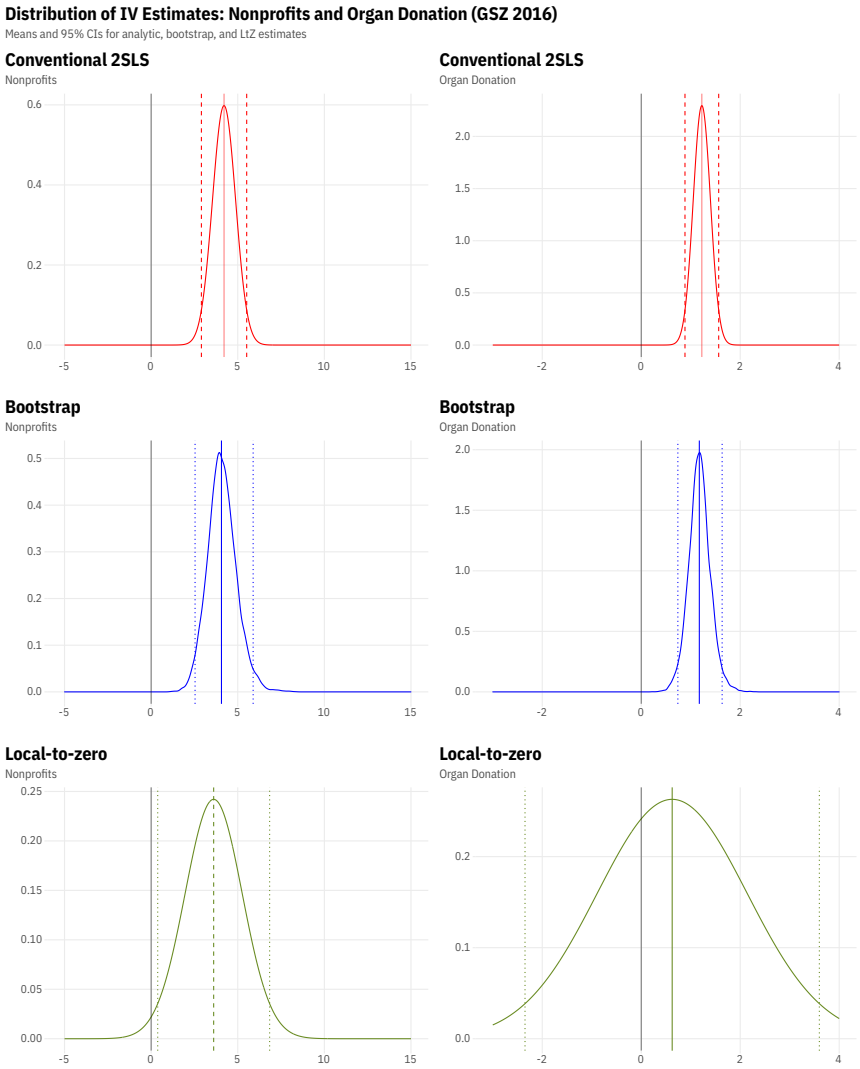
	North		South (ZFS)	
Outcome Variables	Nonprofit	Organ Donation	Nonprofit	Organ Donation
	(1)	(2)	(3)	(4)
Bishop (IV)	1.612	0.472	0.178	0.189
	(0.219)	(0.047)	(0.137)	(0.065)
Observations	5,357	5,535	2,175	2,178

Note: Bootstrap SEs are in the parentheses.

GSZ (2016) use whether the city was the seat of a bishop in the middle ages based on historical accounts of coordination preceding commune formation in the middle ages as an instrument for the "free-city experience" (Section 5). They argue that conditional on a host of geographic covariates, this instrument influences contemporary social capital solely through its increasing the likelihood of commune formation. As suggestive evidence for the validity

of their instrument, they estimate the reduced-form effect of medieval bishop presence of contemporary social capital measures separately in the north (where the instrument is conjectured to have an effect) and the south (where it is conjectured to be irrelevant). They fail to reject the null of no effects in the south, conclude that the instrument appears to have face-validity, and proceed to use bishop presence as an instrument for their IV estimates.

FIGURE 5. IV COEFFICIENTS FOR NON-PROFITS PER-CAPITA AND ORGAN DONATION



We begin by calculating the first-stage partial F-statistic based on bootstrapped SEs for the north sample, which is 67.3. Because there were no "free cities" in the south, the F

statistic for the south is zero by definition. We then replicate their reduced-form estimates in Table 4. The separate north and south reduced-form estimates in GSZ (2016) can be readily used for a LTZ test described above. Since the authors substantively believe that the south is a ZFS sample where bishop presence is irrelevant for treatment assignment, we can use the reduced-form estimates of 0.178 and 0.189 in the south for nonprofits per capita and organ donation (columns 3-4 in Table 4 as the prior μ_γ for the direct effect of the instrument on the outcome. Finally, we report the resultant analytic, bootstrap, and LTZ IV estimates in Figure 5. We find that conventional robust SEs understate the uncertainty of the estimates relative to the bootstrap, and that accounting for direct effect using LTZ attenuates GSZ (2016)'s estimates somewhat and substantially increases the standard error of the estimate for the nonprofit outcome. For organ donation, however, where we suspect an exclusion restriction violation because the reduced form effect is statistically distinguishable from zero, the use of the LTZ method to account for this exclusion restriction violation yields a smaller and substantially more uncertain estimate whose confidence interval contains 0.

This example shows how researchers may take advantage of the ZFS test and the LTZ technique to gauge the robustness of their findings based on an IV strategy.

6. Concluding Remarks

In this paper, we replicate 65 IV designs published in three top journals in political science that use IVs as one of the main identification strategies. We find that researchers often overestimate the strength of their IVs and underestimate the uncertainties around the 2SLS estimates. When using a bootstrap procedure to obtain the uncertainties, we find many 2SLS estimates become uninformative—they are often statistically indistinguishable from the naïve OLS estimates and often times 0. Moreover, we show that the 2SLS estimates are often much larger in magnitude than the OLS estimates, and their difference is negatively correlated with the strength of the IVs. We argue that this is because weak IVs amplify the

biases from exclusion restriction failures. This result suggest that IV estimates likely have much bigger biases than the OLS estimates.

How do we avoid these pitfalls when using an IV approach? A key principle is to start from a design-based perspective, as "design trumps analysis" in causal inference (Rubin, 2008). The IV approach should first and foremost been seen as a method addressing two-sided noncompliance in randomized experiments. Furthermore, since conducting power analysis, designing placebo tests, and preregistering research protocols have become widely-accepted norms in experimental work (see, for example, McDermott 2014), similar transparency-enhancing efforts will also likely reduce the risks of improper use of IVs.

The results of this paper suggests that employing an IV strategy in an observational setting is much more challenging. Because Assumption 2 is not guaranteed by design, researchers are under much heavier burden of proof for the validity of an IV design. On the one hand, truly random (and strong) instruments are rare in our daily lives; on the other hand, placebo tests for the exclusion restriction are difficult to construct after the data are collected. Moreover, researchers often cannot easily expand the sample size to obtain sufficient statistical power. Below we provide a checklist for researchers to reference to when applying or considering to apply an IV strategy with observational data:

- Think clearly whether potential selection is likely inflating or deflating the treatment effect estimates using OLS and whether an IV approach is necessary.
- At the research design stage, think over whether the IV of choice can plausibly create (quasi-)random shocks to the treatment variable.
- After running the first-stage regression, plot X against $\hat{X}$ (covariates and fixed effects should have already been partialed out) and check the IV strength with an eyeball test.
- Calculate the first-stage partial F statistic using bootstrapped SEs. A cluster-bootstrap procedure should be used when the data are clustered or have a group structure. Pro-

ceed if the F statistic is reasonably large.

- Similarly, use bootstrap methods to obtain SEs and confidence intervals (CIs) for the 2SLS estimates.
- If we expect the OLS estimates have upward biases for good reasons, be concerned if the 2SLS estimates are even (much) bigger the OLS ones.
- Try to identify observational analogues of the three principal strata of always takers, never takers, compliers in the purported natural experiment. The "never takers" are a reasonable candidate for a ZFS sample. Then conduct a placebo test by estimating the effect of the IV on the outcome of interest in a ZFS sample.
- Using results from the ZFS test, obtain LTZ IV estimates and CIs, and compare them with original estimates and CIs.

It is our hope that these recommendations can address concerns of using IV strategies in establishing causality in social science research and improve the quality of inference, especially when the IVs are not generated by experiments.

References

Anderson, Theodore W, Naoto Kunitomo and Takamitsu Sawa. 1982. "Evaluation of the distribution function of the limited information maximum likelihood estimator." *Econometrica: Journal of the Econometric Society* pp. 1009–1027.

Andrews, Donald WK and Patrik Guggenberger. 2009. "Validity of subsampling and" plug-in asymptotic" inference for parameters defined by moment inequalities." *Econometric Theory* pp. 669–709.

Andrews, Isaiah, James Stock and Liyang Sun. 2019. "Weak instruments in instrumental variables regression: Theory and practice." *Annual Review of Economics* 11:727–753.

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.

Arellano, Manuel. 2002. "Sargan's intrumental variables estimation and the generalized method of moments." *Journal of Business & Economic Statistics* 20(4):450–459.

Bekker, Paul A. 1994. "Alternative approximations to the distributions of instrumental variable estimators." *Econometrica: Journal of the Econometric Society* pp. 657–681.

Bound, John and David 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.

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.

Bun, Maurice JG and Frank 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 Colin, Jonah B Gelbach and Douglas L Miller. 2008. "Bootstrap-based improvements for inference with clustered errors." *The Review of Economics and Statistics* 90(3):414–427.

Carnegie, Allison and Nikolay Marinov. 2017. "Foreign aid, human rights, and democracy promotion: evidence from a natural experiment." *American Journal of Political Science* 61(3):671–683.

Charles, Nelson and Richard Starz. 1990. "Some Further Results on the Exact Small Sample Properties of the Instrumental Variables Estimator." *Econometrica* 58(41):967–976.

Clarke, Damian. 2014. "PLAUSEXOG: Stata module to implement Conley et al's plausibly exogenous bounds." Statistical Software Components, Boston College Department of Economics.
URL: *https://ideas.repec.org/c/boc/bocode/s457832.html*

Colin Cameron, A and Douglas L Miller. 2015. "A practitioner's guide to cluster-robust inference." *The Journal of Human Resources* 50(2):317–372.

Conley, Timothy G, Christian B Hansen and Peter E Rossi. 2012. "Plausibly exogenous." *The review of economics and statistics* 94(1):260–272.

Davidson, Russell and James G MacKinnon. 2015. "Bootstrap tests for overidentification in linear regression models." *Econometrics* 3(4):825–863.

Dieterle, Steven G and Andy Snell. 2016. "A simple diagnostic to investigate instrument validity and heterogeneous effects when using a single instrument." *Labour Economics* 42:76–86.

Dinas, Elias. 2014. "Does choice bring loyalty? Electoral participation and the development of party identification." *American Journal of Political Science* 58(2):449–465.

Dorsch, Michael T. and Paul Maarek. 2019*a*. "Democratization and the cnditional dynamics of income distribution." *American Political Science Review* 113(2):385–404.

Dorsch, Michael T. and Paul Maarek. 2019*b*. "Democratization and the conditional dynamics of income distribution." *American Political Science Review* 113(2):385–404. Publisher: Cambridge University Press.

Dower, Paul Castañeda, Evgeny Finkel, Scott Gehlbach and Steven Nafziger. 2018. "Collective action and representation in autocracies: Evidence from Russia's great reforms." *American Political Science Review* 112(1):125–147.

Dube, Oeindrila and Suresh 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. Publisher: University of Chicago Press Chicago, IL.

Esarey, J and A Menger. 2019. "Practical and effective approaches to dealing with clustered data." *Political Science Research and Methods* 7(3):541–559.

Fieller, Edgar C. 1954. "Some problems in interval estimation." *Journal of the Royal Statistical Society: Series B (Methodological)* 16(2):175–185.

Goldsmith-Pinkham, Paul, Isaac Sorkin and Henry Swift. 2020. "Bartik instruments: What, when, why, and how." *American Economic Review* 110(8):2586–2624.

Grossman, Guy, Jan H. Pierskalla and Emma Boswell Dean. 2017*a*. "Government fragmentation and public goods provision." *The Journal of Politics* 79(3):823–840.

Grossman, Guy, Jan H. Pierskalla and Emma Boswell Dean. 2017*b*. "Government fragmentation and public goods provision." *The Journal of Politics* 79(3):823–840. Publisher: University of Chicago Press Chicago, IL.

Guiso, Luigi, Paola Sapienza and Luigi Zingales. 2016. "Long-term persistence." *Journal of the European Economic Association* 14(6):1401–1436.

Hager, Anselm and Hanno Hilbig. 2019. "Do inheritance customs affect political and social inequality?" *American Journal of Political Science* 63(4):758–773.

Hahn, Jinyong and Jerry Hausman. 2005. "Estimation with valid and invalid instruments." *Annales d'Economie et de Statistique* pp. 25–57.

Hainmueller, Jens, Jonathan Mummolo and Yiqing Xu. 2019. "How much should we trust estimates from multiplicative interaction models? simple tools to improve empirical practice." *Political Analysis* .

Hansen, Lars Peter. 1982. "Large sample properties of generalized method of moments estimators." *Econometrica: Journal of the Econometric Society* pp. 1029–1054.

Healy, Andrew and Neil Malhotra. 2013. "Childhood socialization and political attitudes: Evidence from a natural experiment." *The Journal of Politics* 75(4):1023–1037. Publisher: Cambridge University Press New York, USA.

Heckman, James J and Edward J Vytlacil. 2007. "Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation." *Handbook of econometrics* 6:4779–4874.

Henderson, John and John Brooks. 2016. "Mediating the electoral connection: The information effects of voter signals on legislative behavior." *The Journal of Politics* 78(3):653–669.

Hirano, Keisuke and Jack R Porter. 2015. "Location properties of point estimators in linear instrumental variables and related models." *Econometric Reviews* 34(6-10):720–733.

Imbens, Guido W. 2014. "Instrumental Variables: An Econometrician's Perspective.". **URL:** *http://arxiv.org/abs/1410.0163*

Jiang, Wei. 2017. "Have instrumental variables brought us closer to the truth." *The Review of Corporate Finance Studies* 6(2):127–140.

Key, Ellen M. 2016. "How are we doing? Data access and replication in political science." *PS: Political Science & Politics* 49(2):268–272.

Kim, Jeong Hyun. 2019. "Direct democracy and women's political engagement." *American Journal of Political Science* 63(3):594–610. Publisher: Wiley Online Library.

Lee, David L, Justin McCrary, Marcelo J Moreira and Jack Porter. 2020. "Valid t-ratio Inference for IV.".

Leonardi, Robert, Raffaella Y Nanetti and Robert D Putnam. 2001. *Making democracy work: Civic traditions in modern Italy*. Princeton university press Princeton, NJ.

Lorentzen, Peter, Pierre Landry and John Yasuda. 2014. "Undermining authoritarian innovation: The power of China's industrial giants." *The Journal of Politics* 76(1):182–194.

McDermott, Rose. 2014. "Research transparency and data archiving for experiments." *PS: Political Science & Politics* 47(1):67–71.

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

Mikusheva, Anna and Brian P Poi. 2006. "Tests and Confidence Sets with Correct Size when Instruments are Potentially Weak." *The Stata journal* 6(3):335–347.

Mogstad, Magne and Alexander Torgovitsky. 2018. "Identification and extrapolation of causal effects with instrumental variables." *Annual review of economics* 10(1):577–613.
URL: *https://doi.org/10.1146/annurev-economics-101617-041813*

Mogstad, Magne, Andres Santos and Alexander Torgovitsky. 2018. "Using instrumental variables for inference about policy relevant treatment parameters." *Econometrica: journal of the Econometric Society* 86(5):1589–1619.
URL: *https://www.econometricsociety.org/doi/10.3982/ECTA15463*

Moreira, Marcelo J. 2003. "A Conditional Likelihood Ratio Test for Structural Models." *Econometrica: journal of the Econometric Society* 71(4):1027–1048.

Nunn, Nathan. 2008. "The long-term effects of Africa's slave trades." *The Quarterly Journal of Economics* 123(1):139–176.

Nunn, Nathan and Leonard Wantchekon. 2011. "The slave trade and the origins of mistrust in Africa." *American Economic Review* 101(7):3221–52.

Rubin, Donald B. 2008. "For Objective Causal Inference, Design Trumps Analysis." *The Annals of Applied Statistics* 2(3):808–840.

Sekhon, Jasjeet S and Rocio Titiunik. 2012. "When natural experiments are neither natural nor experiments." *American Political Science Review* pp. 35–57.

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

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

Staiger, Douglas and James H Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica: journal of the Econometric Society* 65(3):557–586.

van Kippersluis, Hans and Cornelius A Rietveld. 2018. "Beyond plausibly exogenous." *The econometrics journal* 21(3):316–331.
URL: *https://academic.oup.com/ectj/article/21/3/316/5145983*

Vernby, Kare. 2013. "Inclusion and public policy: Evidence from Sweden's introduction of noncitizen suffrage." *American Journal of Political Science* 57(1):15–29.

Young, Alwyn. 2017. "Consistency without inference: Instrumental variables in practical application." *Unpublished manuscript, London: London School of Economics and Political Science. Retrieved from: http://personal. lse. ac. uk/YoungA* .
URL: *https://personal.lse.ac.uk/YoungA/ConsistencyWithoutInference.pdf*

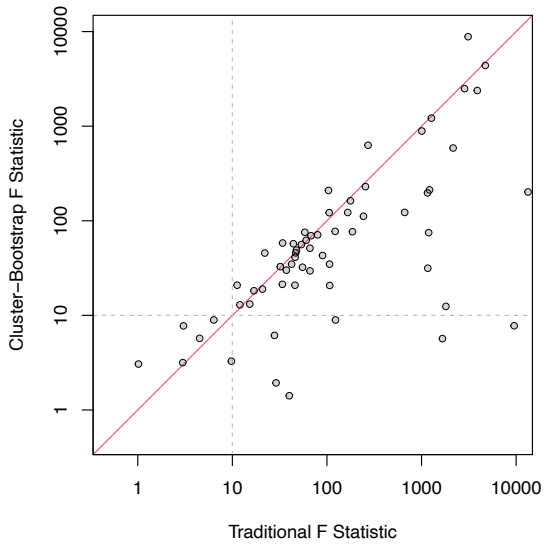
Zhu, Boliang. 2017. "MNCs, rents, and corruption: Evidence from China." *American Journal of Political Science* 61(1):84–99.

A. Supplementary Materials – Appendix A

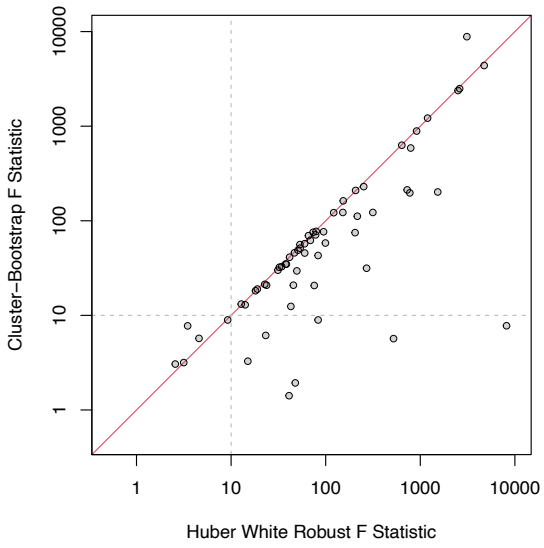
A.1. Additional Replication Results

In Figure A1, we plot the F statistics based on bootstrap or cluster-bootstrap SEs against F statistics based on classic asymptotic SEs, Huber White robust SEs, and cluster-robust SEs, respectively.

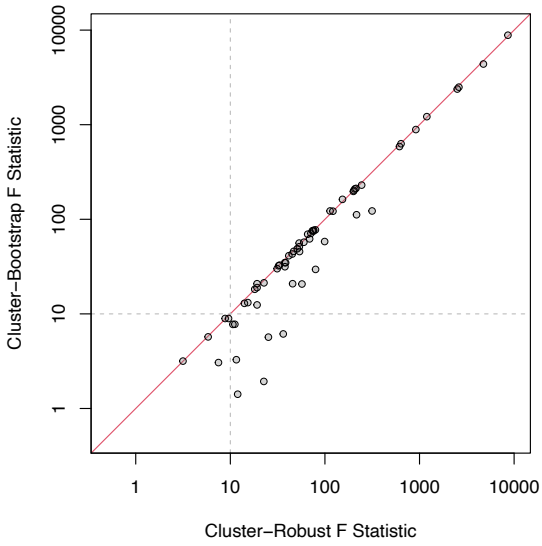
FIGURE A1. COMPARISON OF F STATISTICS BASED ON DIFFERENCE SE ESTIMATORS



(a) Classic Asymptotic SE vs. Cluster-Bootstrap SE



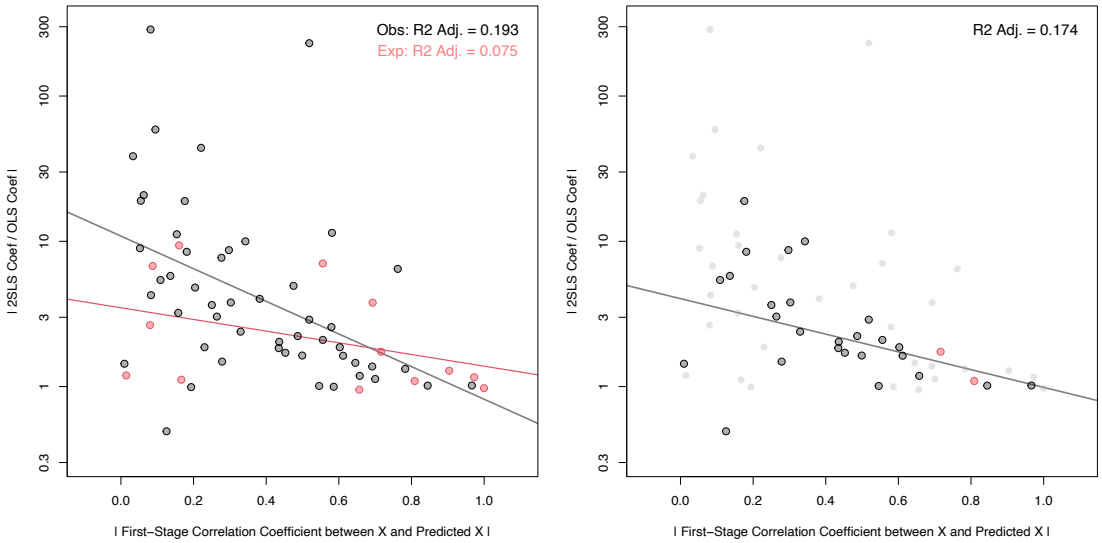
(b) H. White Robust SE vs. Cluster-Bootstrap SE



(c) Cluster-Robust SE vs. Cluster-Bootstrap SE

Instead of using $\left|\frac{\widehat{\beta}_{2SLS}-\widehat{\beta}_{OLS}}{\widehat{\beta}_{OLS}}\right|$ on the y-axis (as in Figure 4 in the main text), Figure A2 plots the absolute value of the ratio between 2SLS estimates and OLS estimates, i.e., $\left|\widehat{\beta}_{2SLS}/\widehat{\beta}_{OLS}\right|$, against $|\rho(X,\hat{X})|$. The results are similar: a strong negative correlation is observed only among studies using non-experimental IVs.

FIGURE A2. IV STRENGTH AND THE OLS-IV DISCREPANCY



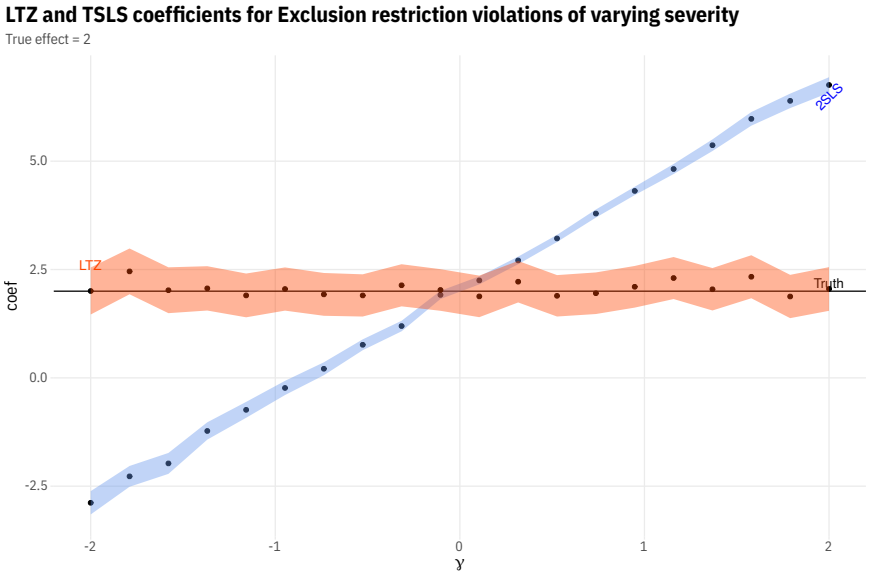
(a) Full Sample

(b) Subsample with Significant OLS Results

A.2. Simulation Study to Demonstrate the LTZ Method

Consider the following DGP, $Y_i = \beta_i D_i + \gamma Z_i + \varepsilon_i$, in which $Z_i \sim$ Bernoulli(0.5) is a binary instrument, $\pi_i \sim \mathsf{U}[1.5, 2.5]$, $\alpha_i \sim \mathcal{N}(-1, 1)$, $\varepsilon_i \sim \mathcal{N}(0, 1)$, $\beta_i \sim \mathcal{N}(1, 0.25)$. The treatment is assigned 1 if a latent variable $D_i^* > 0$ and 0 otherwise; and $D_i^* = \alpha_i + \pi_i Z_i + \varepsilon_i$. We generate the outcome Y_i with Z_i entering directly into the structural equation, which allows us to vary the magnitude of the exclusion restriction violation. We then estimate $\widehat{\beta}_{2SLS}$ using conventional two-staged-least-squares on this data. As we vary γ, $\hat{\beta}_{2SLS}$ is inconsistent for all values except when $\gamma = 0$.

FIGURE A3. IV AND LTZ ESTIMATES FOR VARYING DEGREES OF EXCLUSION RESTRICTION VIOLATIONS



To illustrate the LTZ method, we set $\pi = 0$ for the last 20% observations of the simulated data. We then estimate the reduced-form regression on this (known) sub-sample, and use the coefficient as a prior for μ_γ, and compute the LTZ IV estimate. The results are shown in figure A3. Unlike the 2SLS estimator (in blue), the LTZ estimator uncovers the true value of $\beta = 2$ even for large values of γ.

A.3. Treatment Effect Heterogeneity

One may argue that the observed strong negative relationship in Figure 4 in the main text is driven by treatment effect heterogeneity, combined with the possibility that responsiveness to instrument is positively correlated with treatment effect.[A1] We conduct simulation studies to investigate this argument. Below is the DGP.

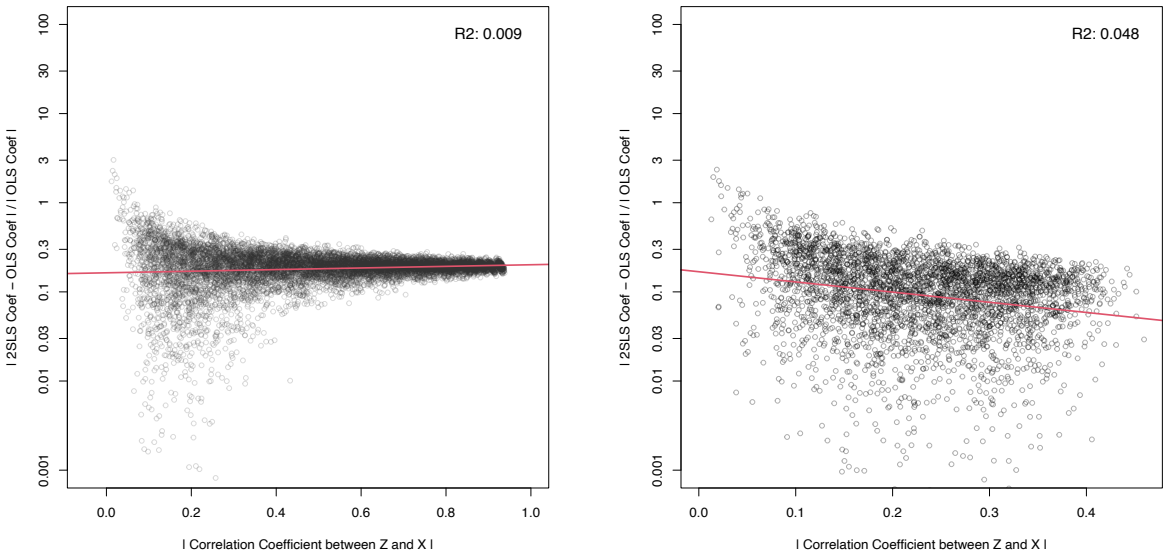
$$
\begin{aligned}
&y_i = 5 + \beta_i x_i + (u_i + b_i) \\
&x_i^* = (\kappa\pi_i) z_i + \left(0.2 v_i + \sqrt{1 - (\kappa\pi_i)^2} \cdot a_i\right) \\
&x_i = x_i^*, \; z_i \overset{i.i.d.}{\sim} N(0,1) \qquad \text{(continuous-continuous case)} \\
&\text{or} \\
&x_i = 1\{x_i^* > 0\}, \; z_i \overset{i.i.d.}{\sim} Bern(0, 0.5) \qquad \text{(binary-binary case)}
\end{aligned}
$$

in which z is the instrument, x is the treatment, and y is the outcome. We consider two scenarios: (1) both x and z are continuous; and (2) both are binary. Correlated errors $\begin{bmatrix} u_i \\ v_i \end{bmatrix} \overset{i.i.d.}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$; $a_i \overset{i.i.d.}{\sim} N(0,1)$, $b_i \overset{i.i.d.}{\sim} N(0,1)$ are the independent errors; first stage and reduced form coefficients may be correlated, i.e., $\begin{bmatrix} \pi_i \\ \beta_i \end{bmatrix} \overset{i.i.d.}{\sim} N\left(\begin{bmatrix} 2 \\ 1 \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & \lambda \\ \lambda & 0.5 \end{bmatrix}\right)$, in which σ controls the amount of heterogeneity in β_i and π_i while λ controls their correlation. In addition, we use κ to control the strength of the first stage.

The results are shown in Figure A4 (for the continuous-continuous case) and Figure A5 (for the binary-binary case). They show that, in both scenarios, when the treatment effect is constant ($\beta_i = \beta, \pi_i = \pi$), in expectation there is no mechanical negative relationship between the correlation coefficient between z and x and the discrepancy between the OLS and 2SLS estimates. However, when the treatment effects are heterogeneous and β_i and π_i are positively correlated, such a negative correlation exists, but its magnitude is much smaller (under our simulation parameters) than what we observed in the replication results in our paper. This suggests that the observed strong negative relationship in Figure 4 cannot solely be explained by treatment effect heterogeneity and responsiveness to the instrument.

[A1] For example, under selection-on-gains type settings, which are typically considered in generalized Roy models underlying MTE approaches to IV.

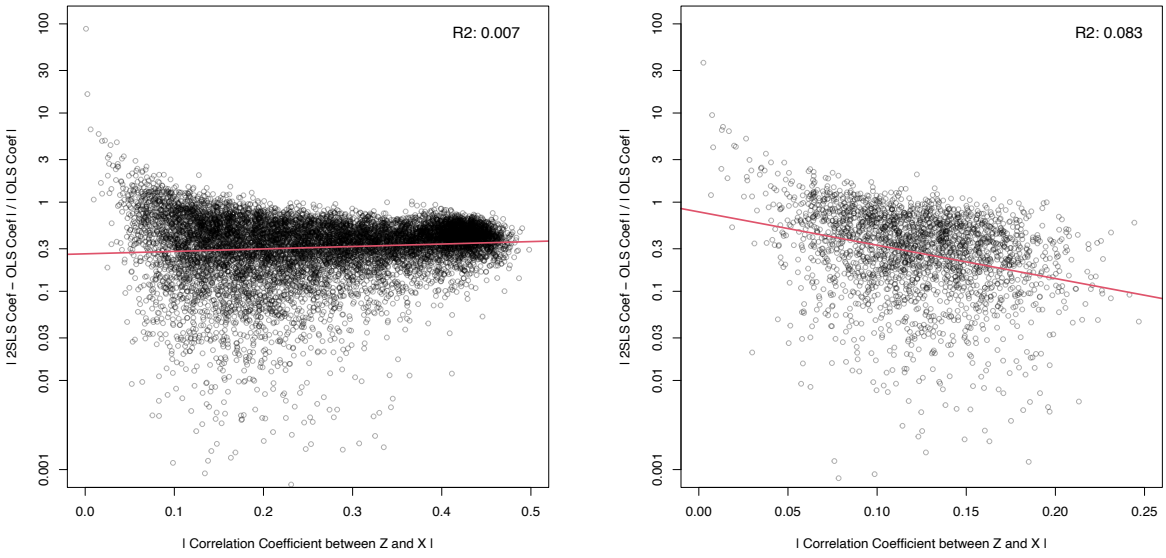
Figure A4. Treatment Strength and OLS-IV Discrepancy Continuous-Continuous Case



(a) Constant Effect ($\sigma^2 = 0$)

(b) Heterogeneous Effect ($\sigma^2 = 0.1, \lambda = 0.7$)

Figure A5. Treatment Strength and OLS-IV Discrepancy Binary-Binary Case



(a) Constant Effect ($\sigma^2 = 0$)

(b) Heterogeneous Effect ($\sigma^2 = 0.1, \lambda = 0.7$)

A.4. Comparing Tests for Detecting Weak Instruments

We conduct a simulation study with a clustered DGP in order to evaluate the relative performance of analytic and bootstrap F-tests to detect weak instruments. We simulate data from the following DGP

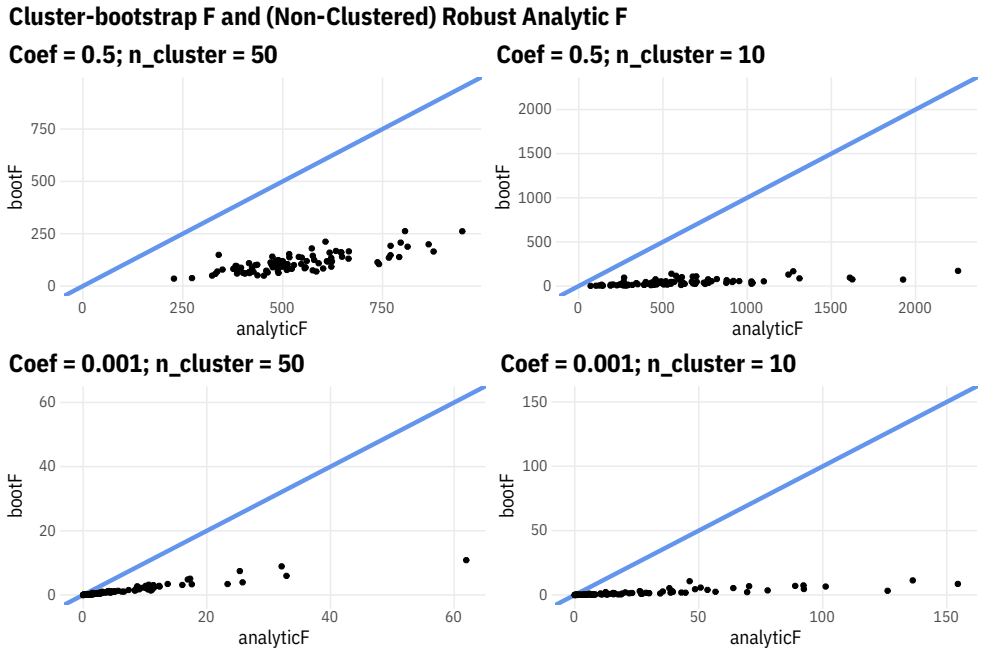
$$\begin{aligned} \text{clustered instrument and error components} \quad & \nu_j, \eta_j \sim \mathcal{N}(0, 0.5) \\ \text{instrument} \quad & z_i \sim \mathcal{N}(0,1) + \nu_j \\ \text{error} \quad & \varepsilon_i \sim \mathcal{N}(0,1) + \eta_j \\ \text{endogenous variable} \quad & x_i = \pi z_i + \varepsilon_i \end{aligned}$$

with errors and instrument components drawn from J clusters. This DGP ensures that the data has dependent structure within each cluster j. We then evaluate the strength of the instrument analytically by computing the t-statistic for $H_0 : \pi = 0$, or by using the corresponding bootstrap statistic $\frac{\pi^2}{\widehat{\sigma}^2}$ where $\widehat{\sigma}^2$ is the bootstrap estimate of the variance of π. We evaluate the analytic and bootstrap F statistics for various values of π and J for 100 replications of the above DGP in Figure (A6).

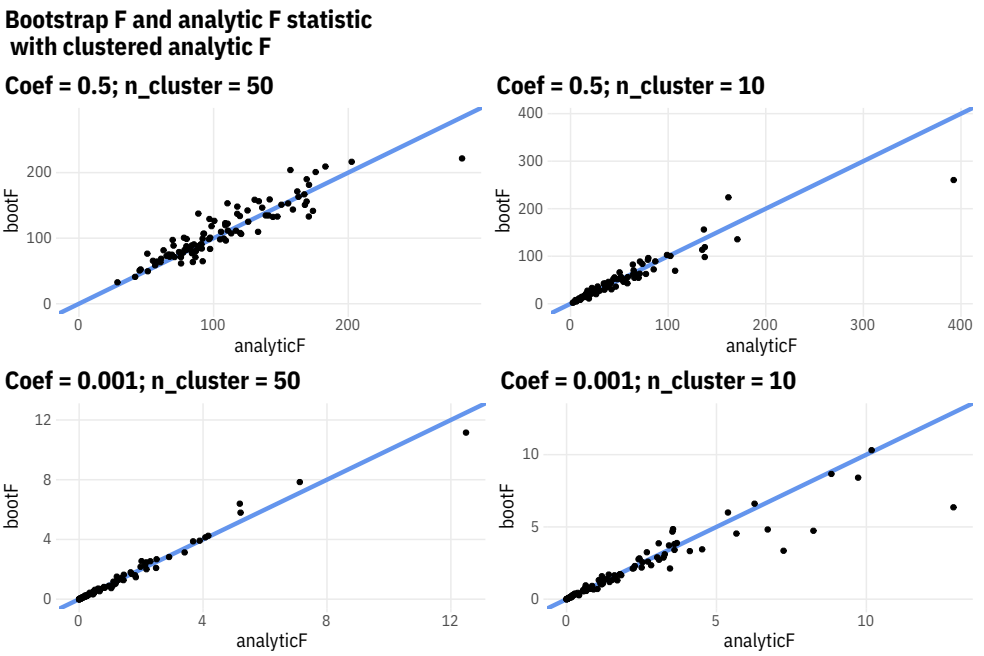
As seen in panel A, when robust analytic standard errors ignore the clustered structure, they vastly over-estimate the strength of the instrument relative to the block-bootstrap, with both 'small' (10) and 'many' (50) clusters and with 'strong' ($\pi = 0.5$) and 'weak' ($\pi = 0.001$) instruments. With appropriate clustered analytic SEs, however, the F-statistic is typically comparable to the bootstrap based equivalent (panel B), although the bootstrap F is marginally more conservative with a small number of clusters and weak instrument.

In summary, we find that cluster-bootstrap F statistic and the cluster-robust F statistic, which is equivalent to the "effective" F (Olea and Pflueger, 2013) in just-identified settings such as this one, are comparable in detecting weak instruments, and recommend reporting these statistics in applied settings. We also recommend reporting Anderson-Rubin confidence intervals for the IV coefficient, as it is robust to arbitrarily weak instruments (Andrews, Stock and Sun, 2019; Kang et al., 2020).

FIGURE A6. COMPARISONS OF F STATISTICS



(a) Cluster-bootstrap F statistic vs. Huber-white (non-clustered) F statistic



(b) Cluster-bootstrap F statistic vs. cluster-robust analytic F statistic

A.5. Summary of Replicated Papers

Table A1. Summary of Replicated Papers

Paper	Instrument	Treatment	Outcome	IV Type	Justification for IV Validity
			APSR		
Gerber, Huber and Washington (2010)	Being sent mail	Aligning party identification with latent partisanship	Voting and party alignment scale	Experiment	NA
Gerber et al. (2011)	Assigned TV and ratio advertisement	Actual TV and ratio advertisement	Voter preference	Experiment	NA
Meredith (2013)	Governor's home county	Democratic governor	Down-ballot Democratic candidates' vote share	Theory (Other)	"The validity of the instruments hinges on the assumption that, conditional on the control variables, coattail effects are the only channel through which the place of birth or residence of a party's gubernatorial candidate affects the vote shares received by its down-ballot candidates."(p.745)
Blattman, Hartman and Blair (2014)	Assignment to treatment blocks	Mass education campaign for dispute resolution	Serious land dispute	Experiment	NA
Laitin and Ramachandran (2016)	Geographic distance from the origins of writing	Language choice	human development index	Theory (Geography)	"[T]he distance from these sites of invention should have no independent impact on socioeconomic development today, except through the channel of affecting the probability of possessing a writing tradition." (p. 470)

Ritter and Conrad (2016)	Rainfall	Mobilized dissent	Repression	Theory (Weather)	"[R]ainfall is an exogenous predictor of dissent onset, meeting the key criteria for the instrumental analysis to allow for causal inference."(p.89)
Croke et al. (2016)	Access to the secondary education	Education attainment	Political participation	Rules & policy changes (Change in exposure)	"There are, however, good reasons to believe that the secondary education reform only affects participation through its effect on educational attainment."(p.592)
Dower et al. (2018)	Religious polarization and level of serfdom	Frequency of unrest	Peasant representation and unrest	Theory (History)	"After conditioning on these covariates, we are left with that portion of serfdom largely determined by idiosyncratic variation in land grants to the nobility decades or centuries before the zemstvo reform of 1864." (p. 133)
Nellis and Siddiqui (2018)	Narrow victory by secular parties in a district	The proportion of MNA seats in a district won by secularist candidates	Religious violence	Theory (Election)	"Our identifying assumption is that the outcomes of such close elections are as good as randomly decided." (p. 50)
Kapoor and Magesan (2018)	Changes in entry costs.	Number of independent candidates	Voter turnout	Rules & policy changes (Change in exposure)	"It is worth reiterating that the deposit increases had nothing to do with historical differences in voter and candidate participation across reserved and open constituencies." (p. 681)

Colantone and Stanig (2018*a*)	Imports from China to the United States × local industrial structure	Regional-level import shock from China	Leave support in Brexit	Econometrics (Interaction)	"[The]instrument is meant to capture the variation in Chinese imports, which is due to the exogenous changes in supply conditions in China, rather than to domestic factors in the United Kingdom that could be correlated with electoral outcomes." (p. 206)
Dorsch and Maarek (2019)	Regional share of democracies	Democratization events	Gini coefficient	Theory (Diffusion)	"Intuitively, we expect that what happens in the regional countries is not related to the degree of inequality in the domestic country i, except through its influence on domestic political institutions." (p. 390)
Hager, Krakowski and Schaub (2019)	Distance to the nearest location where armored military vehicles were stolen	Ethnic riots (destruction)	Prosocial behavior	Theory (Other)	"[W]e present a falsification test which corroborates that the instrument is unrelated to prosocial behavior in a sample of 136 nearby villages, thus underlining the exclusion restriction." (p. 1037)
		AJPS			
Kocher, Pepinsky and Kalyvas (2011)	Past insurgent control	Aerial bombing	Changes in local control	Theory (Other)	"Because instrumental variables require only conditional independence between instruments and the error term, we need only assume that there are no unobserved hamlet-specific variables that affected insurgent control in July, August, and December 1969, but not in September of that year as well." (p. 212)

Vernby (2013)	Immigration Inflow 1940–1950; immigration Inflow 1960–1967	Share of noncitizens in the electorate	Municipal education and social spending	Theory (History)	"Furthermore, it is unlikely that the initial locations of these refugees were affected by the level of local public services, suggesting that the instrument is also valid." (p. 25)
Tajima (2013)	Distance to health station	Distance to police posts (as a proxy for exposure to military intervention)	Incidence of communal violence	Theory (Geography)	"According to a Health Department official, primary health stations must be located in every subdistrict at their population centers, regardless of the propensity for violence of those locations" (p. 112)
De La O (2013)	Random assignment to early coverage	Early coverage of Conditional Cash Transfer	Incumbent party's vote share	Experiment	NA
McClendon (2014)	Assignment to treatment	Reading social esteem promising email	Participation in LGBTQ events	Experiment	NA
Gerber et al. (2015)	Assignment to treatment	Nonreturned mails	Voter turnout	Experiment	NA
Barth, Finseraas and Moene (2015)	Adjusted bargaining coverage and effective number of union confederations	Wage inequality	Welfare support	Theory (Other)	"Yet conditional on union density and country fixed effects, we argue that certain properties of the bargaining system are likely to affect wages, but not union involvement in politics." (p. 574)
Stokes (2016)	Wind speed	Turbine location	Vote turnout	Theory (Climate)	"Wind speed is theoretically orthogonal to precinct boundaries but predicts the placement of wind turbine locations." (p. 965)

Coppock and Green (2016)	Mailing showing 2005 Vote	Voting in November 2007 municipal elections	Voting in the 2008 presidential primary	Experiment	NA
Trounstine (2016)	The number of waterways in a city combined with logged population	Racial segregation	Direct general expenditures	Theory (Geography)	"I focus on waterways (including large streams and rivers), which vary in number across cities and are arguably exogenous to segregation and spending." (p. 717)
Carnegie and Marinov (2017)	Being a former colony of one of the Council members	Foreign aid	CIRI Human Empowerment index	Theory (History)	"In 1965, the EU stipulated that countries would hold the presidency for 6 months at a time [...] and would rotate alphabetically according to each member state's name as spelled in its own language. " (p. 676)
Zhu (2017)	Weighted geographic distance from economic centers	MNC activity	Corruption	Theory (Geography)	"This instrumental variable (IV) is rooted in the gravity models of international trade and FDI flows." (p. 90)
Rueda (2017)	The size of the polling station	Actual polling place size	Citizens' reports of electoral manipulation	Rules & policy changes (Fuzzy RD)	"The institutional rule predicts sharp reductions in the size of the average polling station of a municipality every time the number of registered voters reaches a multiple of the maximum number of voters allowed to vote in a polling station." (p. 173)

Lelkes, Sood and Iyengar (2017)	State-level ROW index	Number of providers	Affective polarization (partisan hostility)	Theory (Other)	"[A]n index of state regulation of right-of-way laws strongly predicts the number of providers in a county, which, as we discuss later, is a good proxy for broadband uptake." (p. 4).
Goldstein and You (2017)	Direct flight from city to Washington DC	Lobbying spending	Total earmarks or grants awarded	Theory (Other)	"The existence of a direct flight captures the convenience of travel to Washington, DC, from each city." (p. 865)
Spenkuch and Tillmann (2018)	Individual princes' decisions concerning whether to adopt Protestantism	Religion of voters living in the same areas more than three and a half centuries later	Nazi vote share	Theory (History)	"The historical record, however, suggests that princes' decisions may plausibly satisfy this exogeneity assumption, especially after controlling for economic conditions at the end of the Weimar Republic as well as all factors known to have influenced rulers." (p. 27)
Escriba-Folch, Meseguer and Wright (2018)	Time trend for received remittances in high-income OECD countries and a country's average distance from the coast.	Remittances	Protests	Theory (Geography)	"Remittances received in high-income OECD countries are unlikely to directly influence political change in remittance-receiving non-OECD countries except through their indirect effect on remittances sent to other countries." (p. 895)

Colantone and Stanig (2018*b*)	Chinese imports to the United States ×regional industrial structure	Regional import shock from China	Economic nationalism	Econometrics (Interaction)	"This instrument is meant to capture the variation in Chinese imports due to exogenous changes in supply conditions in China, rather than to domestic factors that could be correlated with electoral outcomes." (p. 6)
Hager and Hilbig (2019)	Mean elevation; Roman rule	Equitable inheritance customs	Female representation	Theory (Geography; History)	"Rivers are exogenous, but no longer should have a strong effect on inequality other than through the treatment." (p. 767)
López-Moctezuma et al. (2020)	Assignment to treatment	Town-hall meetings	Voting behavior	Experiment	NA
Chong et al. (2019)	Treatment assignment in get-out-to-vote campaigns	Actual proportion of households treated in the locality	Voted in 2013 presidential election	Experiment	NA
Kim (2019)	Population threshold	Democratic institutions	Women political engagement	Rule & rules & policy changes (Fuzzy RD)	"[L]ocalities with a population greater than 1,500 must create a municipal council [...] whereas those with a population below that threshold were free to choose between the status quo direct democracy and representative democracy." (p. 6).
Sexton, Wellhausen and Findley (2019)	Soldier fatalities	Health budget	Welfare outcome	Theory (Other)	"We substantiate [the exclusion restriction] below by ruling out the key alternative channel that local insecurity could affect citizens' use of health services." (p. 359)

			JOP		
Gehlbach and Keefer (2012)	Whether the first ruler in a nondemocratic episode is a military leader	Age of ruling party less leader years in office	Private investment/GDP	Theory	"[D]ictators who come to power with the backing of the military require less popular support to remain in power and are therefore less likely to promote private investment by allowing supporters to organize." (p. 628)
Healy and Malhotra (2013)	Whether the younger sibling is a sister	The share of a respondent's siblings who are female	1973 gender-role attitude	"Experiment" (Biology)	"However, under Assumption 1, all siblings have an impact only through the overall gender makeup of the household." (p. 1027)
Dube and Naidu (2015)	US military aid to countries outside of Latin America	US military aid to Colombia	The number of paramilitary attacks	Theory (Diffusion)	"The instrument is valid since US funding to the rest of the world is determined by the broad geopolitical outlook of the American government, reflecting factors such as the party of the president or other major world events, and can thus be considered exogenous to the conflict in Colombia."(p.256)
Flores-Macias and Kreps (2013)	Lagged values of country's energy production	Trade volume	Foreign policy convergence	Theory (Other)	"The logic is that trade and trade salience in Africa and Latin America are significantly related to countries' energy production, but there is no reason to believe that either of them is correlated with the error term in the equation predicting foreign policy convergence" (p. 365)

Charron and Lapuente (2013)	Consolidation of clientelistic networks in regions where rulers have historically less constraints to their decisions	Clientelism	Quality of government	Theory (History)	"[W]e also find that constraints are directly correlated with current regional institutional quality (yet in his analysis regional GDP and GDP growth are used), thus rendering it an imperfect instrument for clientelism"(p.576)
Kriner and Schickler (2014)	Number of days Congress is in session	Committee investigations	Presidential approval	Theory (Other)	"[T]here is no theoretical reason drawn from existing literatures to expect the calendar to be independently correlated with presidential approval." (p. 525)
Lorentzen, Landry and Yasuda (2014)	Large firm dominance in 1999	Large firm dominance in 2007	Pollution information transparency index	Econometrics (Lagged treatment)	"[The instrument was measured] well before transparency reforms were a major focus of discussion." (p. 187)
Dietrich and Wright (2015)	Constructed "internal" excluded instrument	Economic aid	transitions to multipartyism	Econometrics (Lewbel instrument)	"[We]show that the excluded instruments are generally uncorrelated with alternative channels through which they might influence the outcome variables." (p. 223)
Feigenbaum and Hall (2015)	Chinese exports to other economies ×local exposure	Localized trade shocks in congressional districts	Trade score	Econometrics (Interaction)	"[We] use an instrument that depends [...] on Chinese import growth to other rich, Western economies" and "the lagged version is unaffected by Chinese trade shock." (p.1019)
Alt, Marshall and Lassen (2016)	Assignment to receiving an aggregate unemployment forecast	Unemployment expectations	Vote intention	Experiment	NA

Johns and Pelc (2016)	Trade stake of the rest of the world	The number of other countries that became third parties	Becoming a third party	Theory (Other)	"[E]ach state's participation decision is not directly affected by the trade stake of other countries. The trade stake of other countries matters only to the extent that it shapes a player's belief about how other countries will behave." (p. 99)
Acharya, Blackwell and Sen (2016)	Measures of the environmental suitability for growing cotton	Slave proportion in 1860	proportion Democrat	Theory (History)	"We present results from this analysis showing that, outside the South, the relationship between cotton suitability and political attitudes is either very small or in the opposite direction as in the South." (p. 628)
Schleiter and Tavits (2016)	Prime Minister dissolution power	Opportunistic election calling	Vote share of Prime Minister's party	Theory (Other)	"The instrument correlates directly with the treatment of interest—opportunistic election calling—without being linked to anticipated incumbent electoral performance." (p. 840)
Henderson and Brooks (2016)	Rain around Election Day	Democratic vote margins	Incumbent roll call positioning	Theory (Weather)	"Rain several days before an election may dampen the willingness to make plans, arrange transportation, and schedule time off work to go to the polls."(p.657)

Rozenas (2016)	Proximity-weighted economic shock	Economic crises	measure of office insecurity	Theory (Other)	"The analysis assumes that, conditional on the covariates, country-level effects and time trends, electoral manipulation is not directly correlated with the instrument (independence) and the instrument affects electoral manipulation only through domestic economy (exclusion restriction)."(p.244)
Charron et al. (2017)	Proportion of Protestant residents in a region; aggregate literacy in 1880	More developed bureaucracy	Percent of single bidders	Theory (History)	"[C]ross-country data show that, while the least corrupted countries in the world all have had near universal literacy for decades, other countries considered highly corrupt, [...] have, for the entire postwar era, also been some of the most highly literate places in the world." (p.97)
West (2017)	IEM (prediction market) price	Obama win	Policy efficacy	Theory (Other)	"The identifying assumption is that there is no unobservable factor that simultaneously affects black (female) political efficacy and perceptions of the likelihood of an Obama (Clinton) victory." (p.352)

Stewart and Liou (2017)	Log total border length and the total number of that state's neighbors	Foreign territorial control	Civilian casualties	Theory (Geography)	"[T]he longer a state's borders or the greater its number of neighbors, the more accessible border regions in neighboring states will be to rebels, independent of the dynamics of their conflict with the government. Further, total border length or the number of bordering states is not likely to affect rebel targeting of civilians other than through their effects on the likelihood of rebel group's controlling foreign territory." (p. 291)
Lerman, Sadin and Trachtman (2017)	Born 1946 or 1947	Public (p 1) versus only private (p 0) health insurance	Support ACA	Rules & policy changes (Change in exposure)	"We can confirm across a host of observable covariates that these two age groups are similar on almost every dimension, with the exception of insurance." (p. 631)
Grossman, Pierskalla and Boswell Dean (2017)	The number of distinct landmasses; length of medium and small streams; over-time variation in the number of regional governments	Government fragmentation	Public goods provision	Theory (Geography / diffusion)	"Territorial structure of neighboring countries will affect the local discourse on institutional reforms and increase the likelihood that a country will adopt similar reforms" and "The other two instruments build on the fact that administrative and political boundaries are drawn around geographic landmarks." (p. 831)
Cirone and Van Coppenolle (2018)	Random assignment of budget incumbents to bureaux	Budget committee service	Legislator sponsorship on a budget bill	Theory (Other)	"Conceptually, the competitiveness of the randomly assigned group acts similarly to a form of encouragement design." (p. 953)

Bhavnani and Lee (2018)	Early-career job assignment to districts	Bureaucrats' embeddedness	Proportion of villages with high schools	Theory (Other)	"[T]he IAS posting orders that we obtained suggest that heuristics such alphabetical order and serial number—which are arbitrary and orthogonal to district and officer characteristics—are used to match officers to districts." (p. 78)
Pianzola et al. (2019)	Random assignment of the e-mail treatment	Smartvote use	Vote intention	Experiment	NA
Arias and Stasavage (2019)	Trade shock × UK bond yield	Government expenditures	Regular leader turnover	Econometrics (Interaction)	"The logic here is that when costs of external borrowing are high, a government experiencing a trade shock is more likely to cut expenditures because the option of borrowing to maintain or increase expenditures is too costly. This interaction term is the excluded instrument while the Trade Shock variable is included in both the first- and the second-stage estimates" (p. 1519)
Ziaja (2020)	Constructed instrument	Number of democracy donors	Democracy scores	Econometrics (Interaction)	"[T]here is no reason to believe that the gender composition of a donor country's parliament should affect democracy in a recipient country directly."(p.439)

Note: Justifications are omitted in the case of randomized controlled trials.

References

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. "The political legacy of American slavery." *The Journal of Politics* 78(3):621–641.

Alt, James E., John Marshall and David D. Lassen. 2016. "Credible sources and sophisticated voters: when does new information induce economic voting?" *The Journal of Politics* 78(2):327–342. Publisher: University of Chicago Press Chicago, IL.

Andrews, Isaiah, James Stock and Liyang Sun. 2019. "Weak instruments in instrumental variables regression: Theory and practice." *Annual Review of Economics* 11:727–753.

Arias, Eric and David Stasavage. 2019. "How large are the political costs of fiscal austerity?" *The Journal of Politics* 81(4):1517–1522.

Barth, Erling, Henning Finseraas and Karl O. Moene. 2015. "Political reinforcement: how rising inequality curbs manifested welfare generosity." *American Journal of Political Science* 59(3):565–577. Publisher: Wiley Online Library.

Bhavnani, Rikhil R. and Alexander Lee. 2018. "Local embeddedness and bureaucratic performance: evidence from India." *The Journal of Politics* 80(1):71–87. Publisher: University of Chicago Press Chicago, IL.

Blattman, Christopher, Alexandra C. Hartman and Robert A. Blair. 2014. "How to promote order and property rights under weak rule of law? An experiment in changing dispute resolution behavior through community education." *American Political Science Review* p. 100–120. Publisher: JSTOR.

Carnegie, Allison and Nikolay Marinov. 2017. "Foreign aid, human rights, and democracy promotion: Evidence from a natural experiment." *American Journal of Political Science* 61(3):671–683. Publisher: Wiley Online Library.

Charron, Nicholas, Carl Dahlström, Mihaly Fazekas and Victor Lapuente. 2017. "Careers, connections, and corruption risks: Investigating the impact of bureaucratic meritocracy

on public procurement processes." *The Journal of Politics* 79(1):89–104. Publisher: University of Chicago Press Chicago, IL.

Charron, Nicholas and Victor Lapuente. 2013. "Why do some regions in Europe have a higher quality of government?" *The Journal of Politics* 75(3):567–582. Publisher: Cambridge University Press New York, USA.

Chong, Alberto, Gianmarco León-Ciliotta, Vivian Roza, Martín Valdivia and Gabriela Vega. 2019. "Urbanization patterns, information diffusion, and female voting in rural Paraguay." *American Journal of Political Science* 63(2):323–341. Publisher: Wiley Online Library.

Cirone, Alexandra and Brenda Van Coppenolle. 2018. "Cabinets, committees, and careers: the causal effect of committee service." *The Journal of Politics* 80(3):948–963. Publisher: University of Chicago Press Chicago, IL.

Colantone, Italo and Piero Stanig. 2018*a*. "Global competition and Brexit." *American political science review* 112(2):201–218.

Colantone, Italo and Piero Stanig. 2018*b*. "The trade origins of economic nationalism: Import competition and voting behavior in Western Europe." *American Journal of Political Science* 62(4):936–953. Publisher: Wiley Online Library.

Coppock, Alexander and Donald P. Green. 2016. "Is voting habit forming? New evidence from experiments and regression discontinuities." *American Journal of Political Science* 60(4):1044–1062. Publisher: Wiley Online Library.

Croke, Kevin, Guy Grossman, Horacio A. Larreguy and John Marshall. 2016. "Deliberate disengagement: How education can decrease political participation in electoral authoritarian regimes." *American Political Science Review* 110(3):579–600. Publisher: Cambridge University Press.

De La O, Ana L. 2013. "Do conditional cash transfers affect electoral behavior? Evidence from a randomized experiment in Mexico." *American Journal of Political Science* 57(1):1–14. Publisher: Wiley Online Library.

Dietrich, Simone and Joseph Wright. 2015. "Foreign aid allocation tactics and democratic change in Africa." *The Journal of Politics* 77(1):216–234. Publisher: University of Chicago Press Chicago, IL.

Dorsch, Michael T. and Paul Maarek. 2019. "Democratization and the conditional dynamics of income distribution." *American Political Science Review* 113(2):385–404. Publisher: Cambridge University Press.

Dower, Paul Castaneda, Evgeny Finkel, Scott Gehlbach and Steven Nafziger. 2018. "Collective action and representation in autocracies: Evidence from Russia's great reforms." *American Political Science Review* 112(1):125–147. Publisher: Cambridge University Press.

Dube, Oeindrila and Suresh 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. Publisher: University of Chicago Press Chicago, IL.

Escriba-Folch, Abel, Covadonga Meseguer and Joseph Wright. 2018. "Remittances and protest in dictatorships." *American Journal of Political Science* 62(4):889–904. Publisher: Wiley Online Library.

Feigenbaum, James J. and Andrew B. Hall. 2015. "How legislators respond to localized economic shocks: Evidence from Chinese import competition." *The Journal of Politics* 77(4):1012–1030. Publisher: University of Chicago Press Chicago, IL.

Flores-Macias, Gustavo A. and Sarah E. Kreps. 2013. "The foreign policy consequences of trade: China's commercial relations with Africa and Latin America, 1992–2006." *The Journal of Politics* 75(2):357–371. Publisher: Cambridge University Press New York, USA.

Gehlbach, Scott and Philip Keefer. 2012. "Private investment and the institutionalization of collective action in autocracies: ruling parties and legislatures." *The Journal of Politics* 74(2):621–635. Publisher: Cambridge University Press New York, USA.

Gerber, Alan S., Gregory A. Huber and Ebonya Washington. 2010. "Party affiliation, partisanship, and political beliefs: A field experiment." *American Political Science Review* 104(4):720–744. Publisher: Cambridge University Press.

Gerber, Alan S., Gregory A. Huber, Marc Meredith, Daniel R. Biggers and David J. Hendry. 2015. "Can incarcerated felons be (Re) integrated into the political system? Results from a field experiment." *American Journal of Political Science* 59(4):912–926. Publisher: Wiley Online Library.

Gerber, Alan S., James G. Gimpel, Donald P. Green and Daron R. Shaw. 2011. "How large and long-lasting are the persuasive effects of televised campaign ads? Results from a randomized field experiment." *American Political Science Review* p. 135–150. Publisher: JSTOR.

Goldstein, Rebecca and Hye Young You. 2017. "Cities as lobbyists." *American Journal of Political Science* 61(4):864–876. Publisher: Wiley Online Library.

Grossman, Guy, Jan H. Pierskalla and Emma Boswell Dean. 2017. "Government fragmentation and public goods provision." *The Journal of Politics* 79(3):823–840. Publisher: University of Chicago Press Chicago, IL.

Hager, Anselm and Hanno Hilbig. 2019. "Do inheritance customs affect political and social inequality?" *American Journal of Political Science* 63(4):758–773. Publisher: Wiley Online Library.

Hager, Anselm, Krzysztof Krakowski and Max Schaub. 2019. "Ethnic riots and prosocial behavior: Evidence from kyrgyzstan." *American Political Science Review* 113(4):1029–1044.

Healy, Andrew and Neil Malhotra. 2013. "Childhood socialization and political attitudes: Evidence from a natural experiment." *The Journal of Politics* 75(4):1023–1037. Publisher: Cambridge University Press New York, USA.

Henderson, John and John Brooks. 2016. "Mediating the electoral connection: The information effects of voter signals on legislative behavior." *The Journal of Politics* 78(3):653–669.

Johns, Leslie and Krzysztof J. Pelc. 2016. "Fear of crowds in world trade organization disputes: Why don't more countries participate?" *The Journal of Politics* 78(1):88–104. Publisher: University of Chicago Press Chicago, IL.

Kang, Hyunseung, Yang Jiang, Qingyuan Zhao and Dylan S Small. 2020. "Ivmodel: an R package for inference and sensitivity analysis of instrumental variables models with one endogenous variable." *arXiv preprint arXiv:2002.08457* .

Kapoor, Sacha and Arvind Magesan. 2018. "Independent candidates and political representation in India." *American Political Science Review* 112(3):678–697. Publisher: Cambridge University Press.

Kim, Jeong Hyun. 2019. "Direct democracy and women's political engagement." *American Journal of Political Science* 63(3):594–610. Publisher: Wiley Online Library.

Kocher, Matthew Adam, Thomas B. Pepinsky and Stathis N. Kalyvas. 2011. "Aerial bombing and counterinsurgency in the Vietnam War." *American Journal of Political Science* 55(2):201–218. Publisher: Wiley Online Library.

Kriner, Douglas L. and Eric Schickler. 2014. "Investigating the president: Committee probes and presidential approval, 1953–2006." *The Journal of Politics* 76(2):521–534. Publisher: Cambridge University Press New York, USA.

Laitin, David D. and Rajesh Ramachandran. 2016. "Language policy and human development." *American Political Science Review* 110(3):457–480. Publisher: Cambridge University Press.

Lelkes, Yphtach, Gaurav Sood and Shanto Iyengar. 2017. "The hostile audience: The effect of access to broadband internet on partisan affect." *American Journal of Political Science* 61(1):5–20. Publisher: Wiley Online Library.

Lerman, Amy E., Meredith L. Sadin and Samuel Trachtman. 2017. "Policy uptake as political behavior: evidence from the Affordable Care Act." *The American Political Science Review* 111(4):755. Publisher: Cambridge University Press.

López-Moctezuma, Gabriel, Leonard Wantchekon, Daniel Rubenson, Thomas Fujiwara and Cecilia Pe Lero. 2020. "Policy deliberation and voter persuasion: Experimental evidence from an election in the Philippines." *American Journal of Political Science* .

Lorentzen, Peter, Pierre Landry and John Yasuda. 2014. "Undermining authoritarian innovation: the power of China's industrial giants." *The Journal of Politics* 76(1):182–194. Publisher: Cambridge University Press New York, USA.

McClendon, Gwyneth H. 2014. "Social esteem and participation in contentious politics: A field experiment at an LGBT pride rally." *American Journal of Political Science* 58(2):279–290. Publisher: Wiley Online Library.

Meredith, Marc. 2013. "Exploiting friends-and-neighbors to estimate coattail effects." *American Political Science Review* p. 742–765. Publisher: JSTOR.

Nellis, Gareth and Niloufer Siddiqui. 2018. "Secular party rule and religious violence in Pakistan." *The American Political Science Review* 112(1):49. Publisher: Cambridge University Press.

Olea, José Luis Montiel and Carolin Pflueger. 2013. "A robust test for weak instruments." *Journal of business & economic statistics: a publication of the American Statistical Association* 31(3):358–369.

Pianzola, Joëlle, Alexander H. Trechsel, Kristjan Vassil, Guido Schwerdt and R. Michael Alvarez. 2019. "The impact of personalized information on vote intention: Evidence from a randomized field experiment." *The Journal of Politics* 81(3):833–847. Publisher: The University of Chicago Press Chicago, IL.

Ritter, Emily Hencken and Courtenay R. Conrad. 2016. "Preventing and responding to dissent: The observational challenges of explaining strategic repression." *American Political Science Review* 110(1):85–99. Publisher: Cambridge University Press.

Rozenas, Arturas. 2016. "Office insecurity and electoral manipulation." *The Journal of Politics* 78(1):232–248.

Rueda, Miguel R. 2017. "Small aggregates, big manipulation: Vote buying enforcement and collective monitoring." *American Journal of Political Science* 61(1):163–177. Publisher: Wiley Online Library.

Schleiter, Petra and Margit Tavits. 2016. "The electoral benefits of opportunistic election timing." *The Journal of Politics* 78(3):836–850. Publisher: University of Chicago Press Chicago, IL.

Sexton, Renard, Rachel L. Wellhausen and Michael G. Findley. 2019. "How government reactions to violence worsen social welfare: evidence from Peru." *American Journal of Political Science* 63(2):353–367. Publisher: Wiley Online Library.

Spenkuch, Jörg L. and Philipp Tillmann. 2018. "Elite influence? Religion and the electoral success of the Nazis." *American Journal of Political Science* 62(1):19–36. Publisher: Wiley Online Library.

Stewart, Megan A. and Yu-Ming Liou. 2017. "Do good borders make good rebels? Territorial control and civilian casualties." *The Journal of Politics* 79(1):284–301. Publisher: University of Chicago Press Chicago, IL.

Stokes, Leah C. 2016. "Electoral backlash against climate policy: A natural experiment on retrospective voting and local resistance to public policy." *American Journal of Political Science* 60(4):958–974. Publisher: Wiley Online Library.

Tajima, Yuhki. 2013. "The institutional basis of intercommunal order: Evidence from Indonesia's democratic transition." *American Journal of Political Science* 57(1):104–119. Publisher: Wiley Online Library.

Trounstine, Jessica. 2016. "Segregation and inequality in public goods." *American Journal of Political Science* 60(3):709–725. Publisher: Wiley Online Library.

Vernby, Kare. 2013. "Inclusion and public policy: Evidence from Sweden's introduction of noncitizen suffrage." *American Journal of Political Science* 57(1):15–29.

West, Emily A. 2017. "Descriptive representation and political efficacy: Evidence from Obama and Clinton." *The Journal of Politics* 79(1):351–355. Publisher: University of Chicago Press Chicago, IL.

Zhu, Boliang. 2017. "MNCs, rents, and corruption: Evidence from China." *American Journal of Political Science* 61(1):84–99.

Ziaja, Sebastian. 2020. "More donors, more democracy." *The Journal of Politics* 82(2):433–447. Publisher: The University of Chicago Press Chicago, IL.

Supplemental Materials
Appendix B

How Much Should We Trust Instrumental Variable Estimates in Political Science? Practical Advice based on Over 60 Replicated Studies

Apoorva Lal* Mac Lockhart† Yiqing Xu‡ Ziwen Zu§

14 August 2021

Contents

*PhD Candidate, Stanford University; Email:apoorval@stanford.edu.
†PhD Candidate, University of California, San Diego; Email:mwlockha@ucsd.edu.
‡Assistant Professor, Stanford University; Email:yiqingxu@stanford.edu.
§PhD Student, University of California, San Diego; Email:zzu@ucsd.edu.

APSR

Gerber et al. (2010)

Replication Summary	
Unit of analysis	individual
Treatment	aligning party identification with latent partisanship
Instrument	being sent mail
Outcome	voting and party alignment scale

```
df <- readRDS("./data/apsr_Gerber_etal_2010.rds")
D <-"pt_id_with_lean"
Y <- "pt_voteevalalignindex"
Z <- "treat"
controls <- c("pre_lean_dem", "age", "age2" ,"regyear" ,
              "regyearmissing", "twonames", "combined_female",
              "voted2006", "voted2004", "voted2002", "voted2000",
              "voted1998", "voted1996", "interest", "pre_aligned_vh",
              "pre_direct_unemp", "pre_direct_econ","pre_direct_bushap",
              "pre_direct_congapp")
cl <- NULL
FE <- NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 10.106 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.5658     0.1712     0.1735     0.2120     0.8890
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     3.8231     2.7241    23.5963   -13.0950    21.6720
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##     2.9926     3.1563         NA     3.1796
##
## $p_iv
## [1] 1
##
## $N
## [1] 411
##
## $N_cl
```

```
## NULL
##
## $rho
## [1] 0.0873
```

This paper argues that aligning party identification with latent partisanship increases voting and party alignment. Our replication finds that the OLS estimate is 0.566 (with a standard error of 0.171), and 2SLS estimate is 3.823 (with a standard error of 2.724). The replicated first-stage partial F statistic is 3.18. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Gerber et al. (2011) (a)

Replication Summary	
Unit of analysis	media market
Treatment	actual TV advertisement
Instrument	assigned TV advertisement
Outcome	voter preference

```
df<- readRDS("./data/apsr_Gerber_etal_2011.rds")
Y<- "perry"
D<- "grp_mean"
Z<- "grp_assign"
controls <- c("stray_gr", "cks_rad", "partyvot")
cl <- NULL
FE <- c("week","dma","strata")
weights<-"number"
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 18.599 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     4.5945     1.7686     2.5252     0.1270     9.8300
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     5.3253     1.8226     2.7365     1.1570    11.5430
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##   664.0407   315.1504         NA   144.3485
##
## $p_iv
## [1] 1
##
```

```
## $N
## [1] 72
##
## $N_cl
## NULL
##
## $rho
## [1] 0.9726
```

This paper argues that televised ads have strong (but short-lived) positive effects on voting preferences. Our replication finds that the OLS estimate is 4.595 (with a standard error of 1.769), and 2SLS estimate is 5.325 (with a standard error of 1.823). The replicated first-stage partial F statistic is 144.35. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Gerber et al. (2011) (b)

Replication Summary	
Unit of analysis	media market
Treatment	actual radio advertisement
Instrument	assigned radio advertisement
Outcome	voter preference

```
df<- readRDS("./data/apsr_Gerber_etal_2011.rds")
Y<- "perry"
D<- "radio_gr"
Z<- "rad_assign"
controls <- c("stray_gr", "cks_rad", "partyvot")
cl <- NULL
FE <- c("week","dma","strata")
weights<-"number"
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 18.609 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     3.3461     6.6598    10.0783   -17.0360    22.7000
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     3.2639     6.6611    10.0697   -17.2250    22.4740
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##   98207.78  128147.31         NA   56544.56
```

```
##
## $p_iv
## [1] 1
##
## $N
## [1] 72
##
## $N_cl
## NULL
##
## $rho
## [1] 0.9998
```

This paper argues that radio ads have strong (but short-lived) positive effects on voting preferences. Our replication finds that the OLS estimate is 3.346 (with a standard error of 6.660), and 2SLS estimate is 3.264 (with a standard error of 6.661). The replicated first-stage partial F statistic is 56544.56. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Meredith (2013)

Replication Summary	
Unit of analysis	down-ballot race
Treatment	Democratic governor
Instrument	governor's home county
Outcome	down-ballot Democratic candidates' vote share

```
df <-readRDS("./data/apsr_Meredith_2013.rds")
Y <- "DemShareDB"
D<-"DemShareGOV"
Z <- "HomeGOV"
controls <- c("HomeDB")
cl <- "fips"
FE<- c("fips","RaceID")
weights<-"Weight"
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 31.599 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.4387     0.0079     0.0121     0.4170     0.4650
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.2163     0.0650     0.0842     0.0280     0.3640
##
```

```
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##   186.0359    94.9238    76.0563    73.3119
##
## $p_iv
## [1] 1
##
## $N
## [1] 14562
##
## $N_cl
## [1] 2756
##
## $rho
## [1] 0.1255
```

This article argues that Democratic gubernatorial coatails increase down-ballot Democratic candidates' vote shares. Our replication finds that the OLS estimate is 0.439 (with a standard error of 0.008), and 2SLS estimate is 0.216 (with a standard error of 0.065). The replicated first-stage partial F statistic is 73.31. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Blattman et al. (2014)

Replication Summary	
Unit of analysis	resident
Treatment	mass education campaign for dispute resolution
Instrument	assignment to treatment blocks
Outcome	serious land dispute

```
df <- readRDS("./data/apsr_Blattman_etal_2014.rds")
df$district <- 0
for (i in 1:15) {df$district[which(df[,paste0("district",i)]==1)] <- i}
D <-"months_treated"
Y <- "fightweap_dummy"
Z <- c("block1", "block2", "block3")
controls <- c("ageover60", "age40_60", "age20_40",
 "yrs_edu", "female", "stranger", "christian",
 "minority", "cashearn_imputedhst", "noland",
 "land_sizehst", "farm_sizehst", "lndtake_dum",
 "housetake_dum", "vsmall", "small",
 "small2", "small3", "quartdummy", "cedulevel_bc",
 "ctownhh_log_el", "cwealthindex_bc", "cviol_experienced_bc",
 "clndtake_bc", "cviol_scale_bc", "clandconf_scale_bc",
 "cwitchcraft_scale_bc", "cpalaviol_imputed_bc",
 "cprog_ldr_beliefs_bc", "cattitudes_tribe_bc",
 "crelmarry_bc", "trainee")
cl <- NULL
```

```r
FE <- "district"
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 22.220 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##      7e-04      4e-04      5e-04      0e+00      2e-03
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##      9e-04      5e-04      5e-04      0e+00      2e-03
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##   2756.385   2472.285         NA   2542.566
##
## $p_iv
## [1] 3
##
## $N
## [1] 1900
##
## $N_cl
## NULL
##
## $rho
## [1] 0.9039
```

This paper argues that mass education for dispute resolution can reduce violence and promote dispute resolution. Our replication finds that the OLS estimate is 0.001 (with a standard error of 0.000), and 2SLS estimate is 0.001 (with a standard error of 0.001). The replicated first-stage partial F statistic is 2542.57. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Croke et al. (2016)

Replication Summary	
Unit of analysis	individual
Treatment	education attainment
Instrument	access to the secondary education
Outcome	political participation

```r
df <-readRDS("./data/apsr_Croke_etal_2016.rds")
D <- "edu"
Y <- "part_scale"
Z <- "treatment"
controls <-NULL
cl<- "district"
FE<- "year_survey"
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 20.105 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.0204     0.0051     0.0079    -0.0320    -0.0020
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.0980     0.0263     0.0273    -0.1520    -0.0450
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    79.7552    78.2588    71.1356    65.4134
##
## $p_iv
## [1] 1
##
## $N
## [1] 1842
##
## $N_cl
## [1] 61
##
## $rho
## [1] 0.2041
```

This paper argues that education decreases political participation. Our replication finds that the OLS estimate is -0.020 (with a standard error of 0.005), and 2SLS estimate is -0.098 (with a standard error of 0.026). The replicated first-stage partial F statistic is 65.41. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Laitin and Ramachandran (2016)

Replication Summary	
Unit of analysis	country
Treatment	language choice

Replication Summary	
Instrument	geographic distance from the origins of writing
Outcome	human development index

```
df <-readRDS("./data/apsr_Laitin_2016.rds")
D <-"avgdistance_delta50"
Y <- "zhdi_2010"
Z <- "DIST_BGNC"
controls <- c("cdf2003","ln_GDP_Indp", "edes1975",
              "America","xconst")
cl<- NULL
FE<- NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 8.808 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -1.3676     0.1707     0.1931    -1.7450    -0.9820
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -1.3815     0.3127     0.3012    -1.9430    -0.7790
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    55.1871    32.4040         NA    33.3881
##
## $p_iv
## [1] 1
##
## $N
## [1] 137
##
## $N_cl
## NULL
##
## $rho
## [1] 0.5459
```

This article argues that an official language distant from the local indigenous languages reduces proxies for human capital and health. Our replication finds that the OLS estimate is -1.368 (with a standard error of 0.171), and 2SLS estimate is -1.381 (with a standard error of 0.313). The replicated first-stage partial F statistic is 33.39. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Ritter and Conrad (2016)

Replication Summary	
Unit of analysis	province in 54 African countries*day
Treatment	mobilized dissent
Instrument	rainfall
Outcome	repression

```
df <- readRDS("./data/apsr_Ritter_etal_2016.rds")
D <- "dissentcount"
Y <- "represscount"
Z <- c("lograin", "rainannualpct")
controls <-"urban_mean"
cl<- NULL
FE<- NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 23.112 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.1885     0.0007     0.0068     0.1760     0.2020
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.2708     0.0732     0.0667     0.1380     0.4030
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    58.3505    73.6819         NA    74.0894
##
## $p_iv
## [1] 2
##
## $N
## [1] 1258733
##
## $N_cl
## NULL
##
## $rho
## [1] 0.0096
```

This article argues that mobilized dissent leads to state repression. Our replication finds that the OLS estimate is 0.189 (with a standard error of 0.001), and 2SLS estimate is 0.271 (with a standard

error of 0.073). The replicated first-stage partial F statistic is 74.09. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Colantone and Stanig (2018)

Replication Summary	
Unit of analysis	region
Treatment	regional-level import shock from China
Instrument	imports from China to the United States * local industrial structure
Outcome	leave share

```
df<-readRDS("./data/apsr_Colantone_etal_2018.rds")
D <-'import_shock'
Y <- "leave_share"
Z <- "instrument_for_shock"
controls <- c("immigrant_share", "immigrant_arrivals")
cl <- "fix"
FE <- "nuts1"
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 19.026 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    12.0854     4.1846     4.3347     4.2610    21.5070
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    12.2993     4.3304     4.4960     3.5710    21.5300
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##  2158.0662   792.4682   613.9804   597.3922
##
## $p_iv
## [1] 1
##
## $N
## [1] 167
##
## $N_cl
## [1] 39
##
```

```
## $rho
## [1] 0.9663
```

This paper argues that economic globalization increases that support for the Leave option in the Brexit referendum. Our replication finds that the OLS estimate is 12.085 (with a standard error of 4.185), and 2SLS estimate is 12.299 (with a standard error of 4.330). The replicated first-stage partial F statistic is 597.39. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Dower et al. (2018) (a)

Replication Summary	
Unit of analysis	district*year
Treatment	frequency of unrest
Instrument	religious polarization
Outcome	peasant representation

```
df <- readRDS("./data/apsr_Dower_etal_2018.rds")
D <-"afreq"
Y <-"peasantrepresentation_1864"
Z <-"religpolarf4_1870"
controls <- c("distance_moscow", "goodsoil", "lnurban", "lnpopn", "province_capital")
cl <- NULL
FE <- NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 7.246 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -3.8696     1.9651     1.8376    -7.4820    -0.2460
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##   -32.7701    13.7615    52.6156   -85.3310    -1.7610
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    12.0237    14.0828         NA    14.3606
##
## $p_iv
## [1] 1
##
## $N
## [1] 361
```

```
##
## $N_cl
## NULL
##
## $rho
## [1] 0.1812
```

This paper argues that unrest reduces Russian peasant representation in the 1850s. Our replication finds that the OLS estimate is -3.870 (with a standard error of 1.965), and 2SLS estimate is -32.770 (with a standard error of 13.761). The replicated first-stage partial F statistic is 14.36. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Dower et al. (2018) (b)

Replication Summary	
Unit of analysis	district*year
Treatment	frequency of unrest
Instrument	religious polarization
Outcome	peasant representation

```
df <- readRDS("./data/apsr_Dower_etal_2018.rds")
D <-"afreq"
Y <-"peasantrepresentation_1864"
Z <-"serfperc1"
controls <- c("distance_moscow", "goodsoil", "lnurban", "lnpopn", "province_capital")
cl <- NULL
FE <- NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 7.047 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -4.2492     1.9738     1.9100    -7.8470    -0.6250
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##   -42.4545     8.2406     8.8979   -64.6660   -28.4090
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    47.6256    51.0176         NA    46.7031
##
## $p_iv
## [1] 1
```

```
##
## $N
## [1] 365
##
## $N_cl
## NULL
##
## $rho
## [1] 0.3427
```

This paper argues that unrest reduces Russian peasant representation in the 1850s. Our replication finds that the OLS estimate is -4.249 (with a standard error of 1.974), and 2SLS estimate is -42.455 (with a standard error of 8.241). The replicated first-stage partial F statistic is 46.70. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Kapoor and Magesan (2018)

Replication Summary	
Unit of analysis	constituency*election
Treatment	number of independent candidates
Instrument	changes in entry costs
Outcome	voter turnout

```
df<-readRDS("./data/apsr_Kapoor_etal_2018.rds")
D <-'CitCand_s'
Y <- "Turnout"
Z <- "UnScheduledDepChange"
controls <- c("CitCandBaseTrend", "CitCandBaseTrendSq", "CitCandBaseTrendCu",
              "CitCandBaseTrendQu",  "TurnoutBaseTrend", "TurnoutBaseTrendSq",
              "TurnoutBaseTrendCu", "TurnoutBaseTrendQu", "LnElectors",
              "LagWinDist", "LagWinDistSq", "LagWinDistCu",
              "LagWinDistQu", "LagTightElection")
cl<- "constituency"
FE <- c("year","constituency")
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 21.532 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.3067     0.1006     0.2615    -1.1500    -0.1620
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     5.8312     2.5955     3.0776     1.4560    14.0960
```

```
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    11.2301    23.7168    19.1635    18.6247
##
## $p_iv
## [1] 1
##
## $N
## [1] 4297
##
## $N_cl
## [1] 543
##
## $rho
## [1] 0.0548
```

This paper argues that the participation of independent candidates increases voter turnout in elections. Our replication finds that the OLS estimate is -0.307 (with a standard error of 0.101), and 2SLS estimate is 5.831 (with a standard error of 2.595). The replicated first-stage partial F statistic is 18.62. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Nellis and Siddiqui (2018)

Replication Summary	
Unit of analysis	district*election
Treatment	the proportion of MNA seats in a district won by secularist candidates
Instrument	narrow victory by secular parties in a district
Outcome	religious violence

```
df<-readRDS("./data/apsr_Nellis_etal_2018.rds")
D <-'secular_win'
Y <- "any_violence"
Z <- "secular_close_win"
controls <-"secular_close_race"
cl <- "cluster_var"
FE <- "pro"
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 19.605 sec.

## $est_ols
```

```
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.0150     0.0372     0.0367    -0.0820     0.0620
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.6603     0.2199     0.2636    -1.1320    -0.1400
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    22.0208    60.0400    53.9103    40.3022
##
## $p_iv
## [1] 1
##
## $N
## [1] 437
##
## $N_cl
## [1] 54
##
## $rho
## [1] 0.2207
```

This paper argues that secularist candidates can reduce religious violence in Pakistan. Our replication finds that the OLS estimate is -0.015 (with a standard error of 0.037), and 2SLS estimate is -0.660 (with a standard error of 0.220). The replicated first-stage partial F statistic is 40.30. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Dorsch and Maarek (2019)

Replication Summary	
Unit of analysis	country*year
Treatment	democratization events
Instrument	regional share of democracies
Outcome	Gini coefficient

```
df<-readRDS("./data/apsr_Dorsch_etal_2019.rds")
D <-'Lacemoglu_demo'
Y <- "solt_ginet"
Z <- c("Lneighbour_demo", "L6neighbour_demo")
controls <- c("Llog_gdp","Lsolt_ginet" )
cl<- "ccode"
FE <- c("year","ccode")
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
```

```
## Parallelising 1000 reps on 15 cores
## Bootstrap took 22.255 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.0602     0.0652     0.0973    -0.2540     0.1360
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.2193     0.2610     0.3889    -0.4790     1.0310
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##   123.2485    83.1868     8.8103     8.5158
##
## $p_iv
## [1] 2
##
## $N
## [1] 3905
##
## $N_cl
## [1] 164
##
## $rho
## [1] 0.2503
```

This paper argues that democratization may drive extreme income distributions to a "middle ground"; for countries with low initial level of income inequality, democracy may increase economic inequality. Our replication finds that the OLS estimate is -0.060 (with a standard error of 0.065), and 2SLS estimate is 0.219 (with a standard error of 0.261). The replicated first-stage partial F statistic is 8.52. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Hager et al. (2019)

Replication Summary	
Unit of analysis	individual
Treatment	ethnic riots (destruction)
Instrument	distance to the nearest location where armored military vehicles were stolen
Outcome	prosocial behavior

```
df <- readRDS("./data/apsr_Hager_etal_2019.rds")
D <-"affected"
Y <- "pd_in_scale"
Z <-  "apc_min_distance"
controls <- NULL
```

```r
cl <- NULL
FE <- NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 8.847 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.2335     0.0672     0.0679    -0.3670    -0.0980
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.5200     0.1396     0.1422    -0.8160    -0.2370
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##   271.8565   637.5699         NA   636.9464
##
## $p_iv
## [1] 1
##
## $N
## [1] 878
##
## $N_cl
## NULL
##
## $rho
## [1] 0.4867
```

This paper argues that exposure to ethnic riots reduces people's prosocial behaviors. Our replication finds that the OLS estimate is -0.234 (with a standard error of 0.067), and 2SLS estimate is -0.520 (with a standard error of 0.140). The replicated first-stage partial F statistic is 636.95. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

AJPS

Kocher et al. (2011)

Replication Summary	
Unit of analysis	hamlet (smallest population unit)
Treatment	aerial bombing
Instrument	past insurgent control
Outcome	changes in local control

```r
df<-readRDS("./data/ajps_Kocher_etal_2011.rds")
D <-"bombed_969"
Y<- "mod2a_1adec"
Z <- c("mod2a_1ajul", "mod2a_1aaug")
controls <- c("mod2a_1asep", "score", "ln_dist", "std", "lnhpop")
cl<- NULL
FE <-NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 11.432 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0249     0.0035     0.0044     0.0170     0.0340
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     1.4640     0.1582     0.1388     1.2080     1.7640
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    44.1703    59.8861         NA    59.0432
##
## $p_iv
## [1] 2
##
## $N
## [1] 9707
##
## $N_cl
## NULL
##
## $rho
## [1] 0.095
```

This paper argues that bombing was counterproductive as a counterinsurgency practice: US bombing increases local control by the Viet Cong during the Vietnam War. Our replication finds that the OLS estimate is 0.025 (with a standard error of 0.004), and 2SLS estimate is 1.464 (with a standard error of 0.158). The replicated first-stage partial F statistic is 59.04. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

De La O (2013)

Replication Summary	
Unit of analysis	village

Replication Summary	
Treatment	early coverage of Conditional Cash Transfer
Instrument	random assignment to early coverage
Outcome	incumbent party's vote share

```
df <- readRDS("./data/ajps_De_La_O_2013.rds")
D <-"early_progresa_p"
Y <- "t2000"
Z <- "treatment"
controls <- c("avgpoverty","pobtot1994", "votos_totales1994",
              "pri1994", "pan1994", "prd1994")
cl <- NULL
FE <- "villages"
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 19.111 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0222     0.0517     0.0461    -0.0650     0.1110
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.1563     0.0939     0.0905    -0.0060     0.3580
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##   177.1916   153.2854         NA   151.3373
##
## $p_iv
## [1] 1
##
## $N
## [1] 417
##
## $N_cl
## NULL
##
## $rho
## [1] 0.556
```

This paper argues that the conditional cash transfer program in Mexico mobilizes voter turnout. Our replication finds that the OLS estimate is 0.022 (with a standard error of 0.052), and 2SLS estimate is 0.156 (with a standard error of 0.094). The replicated first-stage partial F statistic is 151.34. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Tajima (2013)

Replication Summary	
Unit of analysis	village and urban neighborhood
Treatment	distance to police posts (as a proxy for exposure to military intervention)
Instrument	distance to health station
Outcome	incidence of communal violence

```
df<-readRDS("./data/ajps_Tajima_2013.rds")
D <-"z2_distpospol"
Y <- "horiz2"
Z <- "z2_dispuskes"
controls <- c("flat", "z2_altitude","urban", "natres", "z2_logvillpop", "z2_logdensvil",
              "z2_povrateksvil", "z2_fgtksvild", "z2_covyredvil", "z2_npwperhh",
              "z2_ethfractvil","z2_ethfractsd", "z2_ethfractd", "z2_relfractvil",
              "z2_relfractsd", "z2_relfractd", "z2_ethclustsd", "z2_ethclustvd",
              "z2_relclustsd", "z2_relclustvd", "z2_wgcovegvil", "z2_wgcovegsd",
              "z2_wgcovegd", "z2_wgcovrgvil", "z2_wgcovrgsd", "z2_wgcovrgd",
              "natdis","javanese_off_java", "islam", "split_kab03", "split_vil03")
cl <- 'kabid03'
FE <- 'prop'
weights<-"probit_touse_wts03"
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 2.977 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.0024     0.0005     0.0006    -0.0040    -0.0010
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.0041     0.0011     0.0015    -0.0070    -0.0010
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
## 13363.7649  1529.0807   202.6374   227.7397
##
## $p_iv
## [1] 1
##
## $N
## [1] 51913
##
## $N_cl
```

```
## [1] 326
##
## $rho
## [1] 0.4527
```

This paper argues that prior exposure to military intervention led to Indonesia's spike in violence during its recent democratic transition. Our replication finds that the OLS estimate is -0.002 (with a standard error of 0.001), and 2SLS estimate is -0.004 (with a standard error of 0.001). The replicated first-stage partial F statistic is 227.74. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Vernby (2013)

Replication Summary	
Unit of analysis	municipality*term
Treatment	share of noncitizens in the electorate
Instrument	immigration Inflow 1940–1950; Immigration Inflow 1960–1967
Outcome	municipal education and social spending

```
df<-readRDS("./data/ajps_Vernby_2013.rds")
D <-"noncitvotsh"
Y <- "Y"
Z <- c("inv1950", "inv1967")
controls <- c("Taxbase2", "L_Taxbase2", "manu", "L_manu", "pop", "L_pop")
cl <- "lan"
FE <- NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 7.865 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     8.9328     1.8008     2.4312     3.0510    12.4730
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    10.5903     2.7459     4.2226     1.6400    18.6350
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    66.2203    49.5670    79.6400    26.8207
##
## $p_iv
## [1] 2
##
```

```
## $N
## [1] 183
##
## $N_cl
## [1] 25
##
## $rho
## [1] 0.6574
```

This paper argues that noncitizens in the electorate increase municipal spending on education. Our replication finds that the OLS estimate is 8.933 (with a standard error of 1.801), and 2SLS estimate is 10.590 (with a standard error of 2.746). The replicated first-stage partial F statistic is 26.82. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

McClendon (2014)

Replication Summary	
Unit of analysis	individual
Treatment	reading social esteem promising email
Instrument	assignment to treatment
Outcome	participation in LGBTQ events

```
df <- readRDS("./data/ajps_McClendon_2014.rds")
D<-"openedesteem"
Y<- "intended"
Z <- "esteem"
controls <- NULL
cl<- NULL
FE <- NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 8.465 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.2823     0.0158     0.0345     0.2140     0.3500
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.3149     0.0950     0.0898     0.1350     0.4910
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##   103.7604   207.1798         NA   207.9915
##
```

```
## $p_iv
## [1] 1
##
## $N
## [1] 3647
##
## $N_cl
## NULL
##
## $rho
## [1] 0.1664
```

This paper argues that the promise of social esteem from an ingroup can encourage participation in contentious politics. Our replication finds that the OLS estimate is 0.282 (with a standard error of 0.016), and 2SLS estimate is 0.315 (with a standard error of 0.095). The replicated first-stage partial F statistic is 207.99. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Barth et al. (2015)

Replication Summary	
Unit of analysis	country*year
Treatment	wage inequality
Instrument	adjusted bargaining coverage; effective number of union confederations
Outcome	welfare support

```
df<- readRDS("./data/ajps_Barth_2015.rds")
D <-"ld9d1"
Y <- "welfareleft"
Z <- c("l2ip_adjcov5", "l2ip_enucfs")
controls <- c("lgdpgr", "lelderly", "llntexp", "lud", "ludsq",
              "lechp", "lnet", "lannual", "ltrend", "ltrendsq")
cl <- FE <- "countrynumber"
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 20.062 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.7755     0.3045     0.3115    -1.4130    -0.1090
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -1.4265     0.7194     2.0106    -4.7970     1.8640
##
```

```
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##     9.7741    15.0268    11.5754     2.7146
##
## $p_iv
## [1] 2
##
## $N
## [1] 117
##
## $N_cl
## [1] 21
##
## $rho
## [1] 0.4345
```

This paper argues that irising inequality reduces political parties' support for welfare. Our replication finds that the OLS estimate is -0.775 (with a standard error of 0.304), and 2SLS estimate is -1.427 (with a standard error of 0.719). The replicated first-stage partial F statistic is 2.71. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Gerber et al. (2015)

Replication Summary	
Unit of analysis	individual
Treatment	nonreturned mail
Instrument	assignment to the treatment
Outcome	voting turnout

```
df<-readRDS("./data/ajps_Gerber_etal_2015.rds")
D <- "treat_combinedXnreturn"
Y <- "registered"
Z <- "treat_combined"
controls <- c("log_daysserved", "ageonelecday", "ageonelecday2",
              "timesincerelease", "timesincerelease2", "v08yes",
              "v08no")
cl <- NULL
FE <- "fix_effect"
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 21.679 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0313     0.0069     0.0077     0.0160     0.0460
```

```
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0298     0.0105     0.0107     0.0100     0.0520
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##   4720.836   4735.295         NA   4892.777
##
## $p_iv
## [1] 1
##
## $N
## [1] 6280
##
## $N_cl
## NULL
##
## $rho
## [1] 0.6564
```

This paper shows that an informational outreach can increase ex-felons' political participation. Our replication finds that the OLS estimate is 0.031 (with a standard error of 0.007), and 2SLS estimate is 0.030 (with a standard error of 0.011). The replicated first-stage partial F statistic is 4892.78. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Coppock and Green (2016)

Replication Summary	
Unit of analysis	individual
Treatment	voting in November 2007 municipal elections
Instrument	mailing showing 2005 Vote
Outcome	voting in the 2008 presidential primary

```
df<-readRDS("./data/ajps_Coppock_etal_2016.rds")
D <-"og2007"
Y <- "JAN2008"
Z <- "treat2"
controls <- NULL
cl <- "hh"
FE <- NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 38.224 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.3126     0.0011     0.0013     0.3100     0.3150
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.3728     0.0782     0.0943     0.1850     0.5470
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##   165.8659   151.8337   113.3680   113.0345
##
## $p_iv
## [1] 1
##
## $N
## [1] 773556
##
## $N_cl
## [1] 562460
##
## $rho
## [1] 0.0146
```

This paper argues that social pressure messages can increase voter turnout. Our replication finds that the OLS estimate is 0.313 (with a standard error of 0.001), and 2SLS estimate is 0.373 (with a standard error of 0.078). The replicated first-stage partial F statistic is 113.03. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Stokes (2016)

Replication Summary	
Unit of analysis	precinct
Treatment	turbine location
Instrument	wind speed
Outcome	vote turnout

```
df<-readRDS("./data/ajps_Stokes_2016.rds")
D <-"prop_3km"
Y <- "chng_lib"
Z <- "avg_pwr_log"
controls <- c("mindistlake", "mindistlake_sq", "longitude",
              "long_sq", "latitude", "lat_sq", "long_lat")
cl <- NULL
FE <- "ed_id"
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 20.361 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.0203     0.0076     0.0071    -0.0350    -0.0070
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.0770     0.0261     0.0302    -0.1450    -0.0230
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    67.9032    65.7306         NA    66.0544
##
## $p_iv
## [1] 1
##
## $N
## [1] 708
##
## $N_cl
## NULL
##
## $rho
## [1] 0.3025
```

This paper argues that turbines decrease incumbent party votes. Our replication finds that the OLS estimate is -0.020 (with a standard error of 0.008), and 2SLS estimate is -0.077 (with a standard error of 0.026). The replicated first-stage partial F statistic is 66.05. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Trounstine (2016)

Replication Summary	
Unit of analysis	city*year
Treatment	racial segregation
Instrument	the number of waterways in a city; logged population
Outcome	direct general expenditures

```
df<-readRDS("./data/ajps_Trounstine_2016.rds")
D <-"H_citytract_NHW_i"
Y <- "dgepercap_cpi"
Z <- c("total_rivs_all", "logpop")
controls <- c("dgepercap_cpilag","diversityinterp","pctblkpopinterp",
              "pctasianpopinterp","pctlatinopopinterp","medincinterp",
              "pctlocalgovworker_100","pctrentersinterp","pctover65",
```

```
              "pctcollegegradinterp","northeast","south","midwest",
              "y5", "y6", "y7", "y8", "y9")
cl <- NULL
FE <- NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 49.462 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.9265     0.4847     0.9015    -2.6930     0.5690
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -2.6757     0.9352     1.7360    -5.7880     0.7940
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##   3883.651   2506.495         NA   2656.165
##
## $p_iv
## [1] 2
##
## $N
## [1] 21145
##
## $N_cl
## NULL
##
## $rho
## [1] 0.5185
```

This paper argues that segregation along racial lines contributes to public goods inequalities. Our replication finds that the OLS estimate is -0.926 (with a standard error of 0.485), and 2SLS estimate is -2.676 (with a standard error of 0.935). The replicated first-stage partial F statistic is 2656.17. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Carnegie and Marinov (2017)

Replication Summary	
Unit of analysis	country*year
Treatment	foreign aid
Instrument	being a former colony of one of the Council members
Outcome	CIRI Human Empowerment index

```r
df<-readRDS("./data/ajps_Carnegie_etal_2017.rds")
D <-"EV"
Y <- "new_empinxavg"
Z <- "l2CPcol2"
controls <- c( "covloggdp", "covloggdpCF", "covloggdpC",
               "covdemregionF", "covdemregion", "coviNY_GDP_PETR_RT_ZSF",
               "coviNY_GDP_PETR_RT_ZS", "covwvs_relF", "covwvs_rel",
               "covwdi_imp", "covwdi_fdiF", "covwdi_fdi",
               "covwdi_expF", "covwdi_exp", "covihme_ayemF", "covihme_ayem")
cl<-"ccode"
FE <- c("year","ccode")
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 21.995 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.1903     0.0421     0.1234    -0.0370     0.4230
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     1.7054     1.0765    12.5991     0.0110     9.0150
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##     4.5101     4.5766     5.8243     5.3607
##
## $p_iv
## [1] 1
##
## $N
## [1] 1792
##
## $N_cl
## [1] 115
##
## $rho
## [1] 0.0523
```

This paper argues that foreign aid has positive effects on human rights and democracy. Our replication finds that the OLS estimate is 0.190 (with a standard error of 0.042), and 2SLS estimate is 1.705 (with a standard error of 1.077). The replicated first-stage partial F statistic is 5.36. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Goldstein and You (2017)

Replication Summary	
Unit of analysis	city
Treatment	lobbying spending
Instrument	direct flight to Washington, DC
Outcome	total earmarks or grants awarded

```
df <- readRDS("./data/ajps_Goldstein_etal_2017.rds")
D <-"ln_citylob"
Y <- "ln_earmark"
Z<- c("direct_flight_dc", "diverge2_e")
controls <- c("pop_e", "land_e", "water_e", "senior_e",
              "student_e", "ethnic_e", "mincome_e", "unemp_e",
              "poverty_e", "gini_e", "city_propertytaxshare_e",
              "city_intgovrevenueshare_e", "city_airexp_e",
              "houdem_e", "ln_countylob")
cl <- "state2"
FE <- 'state2'
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 22.133 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.3198     0.0358     0.0423     0.2150     0.3900
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     1.0278     0.2609     0.2964     0.4300     1.6370
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    15.2856    12.8054    15.3128    15.3990
##
## $p_iv
## [1] 2
##
## $N
## [1] 1262
##
## $N_cl
## [1] 50
##
## $rho
## [1] 0.1579
```

This paper argues that the lobbying spending of a city generates substantial returns in terms of earmarks and grants. Our replication finds that the OLS estimate is 0.320 (with a standard error of 0.036), and 2SLS estimate is 1.028 (with a standard error of 0.261). The replicated first-stage partial F statistic is 15.40. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Lelkes et al. (2017)

Replication Summary	
Unit of analysis	state*year
Treatment	number of broadband Internet providers
Instrument	state-level ROW index
Outcome	affective polarization

```
#nodemos$outcome<-zero1(nodemos$infeels-nodemos$outfeels)
#nodemos$D<-log(nodemos$providers)
#nodemos$HHINC_log<-log(nodemos$HHINC)
#nodemos$Total_log<-log(nodemos$Total)
df<-readRDS("./data/ajps_Lelkes_2017.rds")
D <-"D"
Y <- "outcome"
Z <- "Total_log"
controls <- c("region", "percent_black", "percent_white",
              "percent_male", "lowed", "unemploymentrate",
              "density", "HHINC_log")
cl<- "state"
FE <- "year"
weights=NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 3.159 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0041     0.0018     0.0036    -0.0030     0.0110
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0316     0.0065     0.2085    -0.0090     0.1200
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##  9525.8467  8161.7346    11.1632     6.7960
##
## $p_iv
```

```
## [1] 1
##
## $N
## [1] 114803
##
## $N_cl
## [1] 48
##
## $rho
## [1] 0.2768
```

This paper shows that access to broadband Internet increases partisan hostility. Our replication finds that the OLS estimate is 0.004 (with a standard error of 0.002), and 2SLS estimate is 0.032 (with a standard error of 0.006). The replicated first-stage partial F statistic is 6.80. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Rueda (2017)

Replication Summary	
Unit of analysis	city
Treatment	actual polling place size.
Instrument	the size of the polling station
Outcome	citizens' reports of electoral manipulation

```
df <- readRDS("./data/ajps_Rueda_2017.rds")
D <-"lm_pob_mesa"
Y <- "e_vote_buying"
Z <- "lz_pob_mesa_f"
controls <- c("lpopulation", "lpotencial")
cl <- "muni_code"
FE <- NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 10.640 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.6750     0.0893     0.1002    -0.8780    -0.4870
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.9835     0.1385     0.1414    -1.2800    -0.7280
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
```

```
##   3106.387   3108.591   8598.326   8526.043
##
## $p_iv
## [1] 1
##
## $N
## [1] 4352
##
## $N_cl
## [1] 1098
##
## $rho
## [1] 0.6455
```

This paper argues that polling place sizes reduce vote buying. Our replication finds that the OLS estimate is -0.675 (with a standard error of 0.089), and 2SLS estimate is -0.984 (with a standard error of 0.139). The replicated first-stage partial F statistic is 8526.04. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Zhu (2017)

Replication Summary	
Unit of analysis	province*period
Treatment	MNC activity
Instrument	weighted geographic closeness
Outcome	corruption

```
df <- readRDS("./data/ajps_Zhu_2017.rds")
D <-"MNC"
Y <- "corruption1"
Z <- "lwdist"
controls <- c("lgdpcap6978", "gdp6978", "population", "lgovtexp9302",
              "pubempratio", "leduc", "pwratio", "female", "time")
cl <- NULL
FE <- NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 8.065 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.3531     0.1355     0.1255     0.0570     0.5440
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
```

```
##     0.4855     0.1977     0.2231     0.1430     0.9220
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    45.9155    45.5515         NA    24.4434
##
## $p_iv
## [1] 1
##
## $N
## [1] 61
##
## $N_cl
## NULL
##
## $rho
## [1] 0.6919
```

This paper argues that the entry and presence of MNCs may contribute to rent creation in developing countries, thereby leading to a high level of corruption. Our replication finds that the OLS estimate is 0.353 (with a standard error of 0.136), and 2SLS estimate is 0.485 (with a standard error of 0.198). The replicated first-stage partial F statistic is 24.44. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Colantone and Stanig (2018)

Replication Summary	
Unit of analysis	region*year
Treatment	regional import shock from China
Instrument	Chinese imports to the United States
Outcome	Economic nationalism

```
df <-readRDS("./data/ajps_Colantone_etal_2018.rds")
D <-"import_shock"
Y <- "median_nationalism"
Z <- "instrument_for_shock"
controls <- NULL
cl <- "nuts2_year"
FE <- "fix_effect"
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 23.325 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
```

```
##     0.6442     0.1337     0.3840     0.1970     1.6770
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     1.3096     0.3073     0.5857     0.5200     2.7750
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##  1810.3678    42.8350    19.1709    12.4818
##
## $p_iv
## [1] 1
##
## $N
## [1] 7782
##
## $N_cl
## [1] 739
##
## $rho
## [1] 0.4358
```

This paper argues that Chinese import shocks lead to economic nationalism in the UK. Our replication finds that the OLS estimate is 0.644 (with a standard error of 0.134), and 2SLS estimate is 1.310 (with a standard error of 0.307). The replicated first-stage partial F statistic is 12.48. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Escriba-Folch et al. (2018)

Replication Summary	
Unit of analysis	country*period
Treatment	remittances
Instrument	time trend for received remittances in high-income OECD countries and a country's average distance from the coast.
Outcome	protests

```
df<-readRDS("./data/ajps_Escriba_etal_2018.rds")
D <-"remit"
Y <- "Protest"
Z <-  "distwremit"
controls <- c("l1gdp", "l1pop", "l1nbr5", "l12gr", "l1migr", "elec3","dict")
cl<- "caseid"
FE <- c("cowcode","period")
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 21.867 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0357     0.0088     0.0215    -0.0060     0.0770
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0354     0.0456     0.0776    -0.1170     0.1810
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    89.9012    83.0913    45.1487    41.3127
##
## $p_iv
## [1] 1
##
## $N
## [1] 2428
##
## $N_cl
## [1] 208
##
## $rho
## [1] 0.1934
```

This paper argues that remittances increase political protest in nondemocracies. Our replication finds that the OLS estimate is 0.036 (with a standard error of 0.009), and 2SLS estimate is 0.035 (with a standard error of 0.046). The replicated first-stage partial F statistic is 41.31. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Spenkuch and Tillmann (2018)

Replication Summary	
Unit of analysis	electoral district
Treatment	religion of voters living in the same areas more than three and a half centuries later
Instrument	individual princes' decisions concerning whether to adopt Protestantism
Outcome	Nazi vote share

```
df <-readRDS("./data/ajps_Spenkuch_etal_2018.rds")
D <-"r_1925C_kath"
Y <- "r_NSDAP_NOV1932_p"
Z <- c("r_kath1624", "r_gem1624")
controls <- c("r_1925C_juden", "r_1925C_others",
```

```
              "r_M1925C_juden","r_M1925C_others")
cl <- 'WKNR'
FE <- NULL
weights="r_wahlberechtigte_NOV1932"
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 11.815 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.2504     0.0069     0.0186    -0.2860    -0.2130
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.2544     0.0081     0.0182    -0.2880    -0.2180
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##  1215.3547   726.7058   212.7390   198.5380
##
## $p_iv
## [1] 2
##
## $N
## [1] 982
##
## $N_cl
## [1] 35
##
## $rho
## [1] 0.8446
```

This paper argues that Catholicism decreased Nazi vote shares. Our replication finds that the OLS estimate is -0.250 (with a standard error of 0.007), and 2SLS estimate is -0.254 (with a standard error of 0.008). The replicated first-stage partial F statistic is 198.54. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Chong et al. (2019)

Replication Summary	
Unit of analysis	household
Treatment	actual proportion of households treated in the locality
Instrument	treatment assignment in get-out-to-vote campaigns
Outcome	voted in 2013 presidential election

```
df <-readRDS("./data/ajps_Chong_etal_2019.rds")
D <-"ratio_treat"
Y <- "elecc_presid2013"
Z <- c("D2D30", "D2D40", "D2D50")
controls <-c("age", "married", "children", "num_children",
             "employed", "languag", "yrseduc", "bornloc",
             "hh_asset_index", "log_pop", "mujeres_perc",
             "pob_0_14_perc", "pob_15_64_perc", "pob_65mas_perc",
             "analfabetos_perc", "asiste_escuela_perc",
             "TASA_women", "TASA_men", "electricidad_perc",
             "agua_perc", "desague_perc", "basura_perc",
             "fono_fijo_perc", "fono_cel_perc", "ocupantes",
             "Rural",  "distancia2_final", "db_age",
             "db_married", "db_children", "db_num_children",
             "db_employed", "db_languag", "db_yrseduc",
             "db_bornloc", "db_hh_asset_index", "db_log_pop",
             "db_mujeres_perc", "db_pob_0_14_perc",
             "db_pob_15_64_perc", "db_pob_65mas_perc",
             "db_analfabetos_perc", "db_asiste_escuela_perc",
             "db_TASA_women", "db_TASA_men", "db_electricidad_perc",
             "db_agua_perc", "db_desague_perc", "db_basura_perc",
             "db_fono_fijo_perc", "db_fono_cel_perc",
             "db_ocupantes", "db_Rural", "db_distancia2_final",
             "dpto1", "elecc_presid2008", "db_elecc_presid2008")
cl <- "loc"
FE <- NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 24.886 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0715     0.0393     0.0452    -0.0200     0.1570
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.1242     0.0549     0.0555     0.0140     0.2300
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##  1163.8658   270.5690    37.7653    32.2500
##
## $p_iv
## [1] 3
##
## $N
```

```
## [1] 3350
##
## $N_cl
## [1] 282
##
## $rho
## [1] 0.7163
```

This paper argues that door-to-door get-out-to-vote campaigns increased voter turnout. Our replication finds that the OLS estimate is 0.071 (with a standard error of 0.039), and 2SLS estimate is 0.124 (with a standard error of 0.055). The replicated first-stage partial F statistic is 32.25. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Hager and Hilbig (2019) a

Replication Summary	
Unit of analysis	city
Treatment	equitable inheritance customs
Instrument	mean elevation
Outcome	female representation

```
df<-readRDS("./data/ajps_Hager_etal_2019.rds")
D <-"fair_dic"
Y <- "gem_women_share"
Z <- "elev_mean"
controls <- c("lon", "lat", "childlabor_mean_1898",
              "support_expenses_total_capita","gem_council",
              "gem_pop_density","pop_tot")
cl<- NULL
FE<- c("state2","law_cat2")
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 20.423 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0072     0.0041     0.0043    -0.0010     0.0160
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.1363     0.0263     0.0270     0.0920     0.1970
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##   122.1930    79.2985         NA    75.9368
```

```
##
## $p_iv
## [1] 1
##
## $N
## [1] 3850
##
## $N_cl
## NULL
##
## $rho
## [1] 0.1758
```

This paper argues that equitable inheritance increased female political representation. Our replication finds that the OLS estimate is 0.007 (with a standard error of 0.004), and 2SLS estimate is 0.136 (with a standard error of 0.026). The replicated first-stage partial F statistic is 75.94. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Hager and Hilbig (2019) b

Replication Summary	
Unit of analysis	city
Treatment	equitable inheritance customs
Instrument	Roman rule
Outcome	female representation

```
df<-readRDS("./data/ajps_Hager_etal_2019.rds")
D <-"fair_dic"
Y <- "gem_women_share"
Z <-"roman"
controls <- c("lon", "lat", "childlabor_mean_1898",
              "support_expenses_total_capita","gem_council",
              "gem_pop_density","pop_tot")
cl<- NULL
FE<- c("state2","law_cat2")
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 20.516 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0072     0.0041     0.0043    -0.0010     0.0160
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
```

```
##     0.0390     0.0381     0.0392    -0.0440     0.1140
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    46.0764    41.5303         NA    43.4084
##
## $p_iv
## [1] 1
##
## $N
## [1] 3851
##
## $N_cl
## NULL
##
## $rho
## [1] 0.109
```

This paper argues that equitable inheritance increased female political representation. Our replication finds that the OLS estimate is 0.007 (with a standard error of 0.004), and 2SLS estimate is 0.039 (with a standard error of 0.038). The replicated first-stage partial F statistic is 43.41. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Kim (2019)

Replication Summary	
Unit of analysis	municipality*year
Treatment	Democratic institutions
Instrument	population threshold
Outcome	women political engagement

```
df<- readRDS("./data/ajps_Kim_2019.rds")
D <-"direct"
Y <- "wm_turnout"
Z <-  "new"
controls <- c("left", "wm_voters", "enep")
cl <- NULL
FE <- "year"
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 20.643 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0170     0.4912     0.4663    -0.9450     0.9430
```

```
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     3.9287     0.9580     1.1123     2.0470     6.3620
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##  1007.3382   914.6461         NA   894.5646
##
## $p_iv
## [1] 1
##
## $N
## [1] 2749
##
## $N_cl
## NULL
##
## $rho
## [1] 0.5186
```

This paper argues that the presence of direct democracy expands gender equality in political participation. Our replication finds that the OLS estimate is 0.017 (with a standard error of 0.491), and 2SLS estimate is 3.929 (with a standard error of 0.958). The replicated first-stage partial F statistic is 894.56. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Sexton et al. (2019)

Replication Summary	
Unit of analysis	department*year
Treatment	health budget
Instrument	soldier fatalities
Outcome	health social service

```
df <-readRDS("./data/ajps_Sexton_etal_2019.rds")
D<-"socialservice_b"
Y <- "Finfant_mortality"
Z <- "Lgk_budget"
controls <- c("Lgk_prebudget", "ln_pbi_pc", "execution_nohealth")
cl <- "deptcode"
FE <- c("year","deptcode")
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 19.780 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -1.3472     1.1557     1.1148    -3.2080     1.2900
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##   -15.0645    15.6822    39.0728   -46.4970    10.1230
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##     1.0172     2.5692     7.4923     2.7772
##
## $p_iv
## [1] 1
##
## $N
## [1] 72
##
## $N_cl
## [1] 24
##
## $rho
## [1] 0.1538
```

This paper argues that attacks on soldiers during the budget negotiation period worsened social welfare. Our replication finds that the OLS estimate is -1.347 (with a standard error of 1.156), and 2SLS estimate is -15.065 (with a standard error of 15.682). The replicated first-stage partial F statistic is 2.78. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

López-Moctezuma et al. (2020)

Replication Summary	
Unit of analysis	individual
Treatment	town-hall meetings
Instrument	assignment to treatment
Outcome	voting behavior

```
df <-readRDS("./data/ajps_Moctezuma_etal_2020.rds")
df<-as.data.frame(df)
D<-"treatment"
Y <- "vote"
Z <- "assignment"
  controls <- NULL
cl <- "barangay"
FE <- "city"
weights<-"weight.att"
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
```

```
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 20.206 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    16.1643     3.1714     4.3437     6.6790    23.5660
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    17.6531     3.9211   215.4995    -7.3990    73.9860
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##  1663.9064   521.4034    25.2694     5.6935
##
## $p_iv
## [1] 1
##
## $N
## [1] 890
##
## $N_cl
## [1] 30
##
## $rho
## [1] 0.8089
```

This paper argues that town-hall meetings have a positive effect on parties' vote shares. Our replication finds that the OLS estimate is 16.164 (with a standard error of 3.171), and 2SLS estimate is 17.653 (with a standard error of 3.921). The replicated first-stage partial F statistic is 5.69. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

JOP

Gehlbach and Keefer (2012)

Replication Summary	
Unit of analysis	nondemocratic episode
Treatment	age of ruling party less leader years in office
Instrument	whether the first ruler in a nondemocratic episode is a military leader
Outcome	private invest

```
df<- readRDS("./data/jop_Gelbach_etal_2012.rds")
D <- "gov1_yrs"
```

```r
Y <- "gfcf_priv_gdp"
Z <- "military_first_alt"
controls <- c("tenure", "stabs", "fuelex_gdp", "oresex_gdp",
              "frac_ethn", "frac_relig", "frac_ling", "pop_yng_pct",
              "pop_tot", "pop_ru_pct", "land_km", "gdppc_ppp_2005_us")
cl <- "ifs_code"
FE <-NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 9.451 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.1304     0.0469     0.0425     0.0530     0.2230
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.3956     0.2084     0.3307     0.1100     1.0860
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##     6.3713     9.2042     9.5714     8.9259
##
## $p_iv
## [1] 1
##
## $N
## [1] 99
##
## $N_cl
## [1] 86
##
## $rho
## [1] 0.2641
```

This paper argues that ruling-party institutionalization increases private investments in autocracies. Our replication finds that the OLS estimate is 0.130 (with a standard error of 0.047), and 2SLS estimate is 0.396 (with a standard error of 0.208). The replicated first-stage partial F statistic is 8.93. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Charron and Lapuente (2013)

Replication Summary	
Unit of analysis	region

Replication Summary	
Treatment	clientelism
Instrument	consolidation of clientelistic networks in regions where rulers have historically less constraints to their decisions
Outcome	quality of governments

```
df<-readRDS("./data/jop_Charron_etal_2013.rds")
D <- "pc_all4_tol"
Y <- "eqi"
Z <- c("pc_institutions","literacy1880")
controls <- c("logpop", "capitalregion", "ger", "it", "uk","urb_1860_1850_30")
cl <- NULL
FE <- NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 8.947 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0176     0.0030     0.0034     0.0110     0.0240
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0233     0.0039     0.0041     0.0150     0.0310
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    37.2005    31.2712         NA    31.8018
##
## $p_iv
## [1] 2
##
## $N
## [1] 56
##
## $N_cl
## NULL
##
## $rho
## [1] 0.7828
```

This paper argues that clientelism reduces quality of government. Our replication finds that the OLS estimate is 0.018 (with a standard error of 0.003), and 2SLS estimate is 0.023 (with a standard error of 0.004). The replicated first-stage partial F statistic is 31.80. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Flores-Macias and Kreps (2013)

Replication Summary	
Unit of analysis	country*year
Treatment	trade volume
Instrument	lagged energy production
Outcome	foreign policy convergence

```
df<- readRDS("./data/jop_Flores_etal_2013.rds")
D <- "log_tot_trade"
Y <- "log_HRVOTE"
Z <- "lag_log_energ_prod"
controls <- c("log_cinc", "us_aid100", "log_tot_ustrade",
              "Joint_Dem_Dum", "pts_score", "dummy2004")
cl <- NULL
FE <- 'statea'
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 19.680 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0191     0.0043     0.0045     0.0120     0.0280
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0456     0.0135     0.0141     0.0210     0.0760
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    66.1143    53.6345         NA    47.3118
##
## $p_iv
## [1] 1
##
## $N
## [1] 592
##
## $N_cl
## NULL
##
## $rho
## [1] 0.3295
```

This paper argues that trade flows with China affected states to converge with China on issues of foreign policy. Our replication finds that the OLS estimate is 0.019 (with a standard error of 0.004),

and 2SLS estimate is 0.046 (with a standard error of 0.013). The replicated first-stage partial F statistic is 47.31. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Healy and Malhotra (2013)

Replication Summary	
Unit of analysis	individual
Treatment	the share of a respondent's siblings who are female
Instrument	whether the younger sibling is a sister
Outcome	gender-role attitude in 1973

```
df <- readRDS("./data/jop_Healy_etal_2013.rds")
D <-"share_sis"
Y <- "womens_rights73"
Z <- "closest"
controls <- "num_sib"
cl <- "PSU"
FE <- NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 8.894 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0451     0.0541     0.0515    -0.0540     0.1420
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.1706     0.0788     0.0874     0.0090     0.3480
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##   255.3329   252.1198   244.4704   255.4208
##
## $p_iv
## [1] 1
##
## $N
## [1] 279
##
## $N_cl
## [1] 89
##
```

```
## $rho
## [1] 0.6932
```

This paper argues that having sisters causes young men to be more likely to express conservative viewpoints with regards to gender roles and to identify as Republicans. Our replication finds that the OLS estimate is 0.045 (with a standard error of 0.054), and 2SLS estimate is 0.171 (with a standard error of 0.079). The replicated first-stage partial F statistic is 255.42. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Kriner and Schickler (2014)

Replication Summary	
Unit of analysis	month
Treatment	committee investigations
Instrument	number of days that Congress was in session in a given month
Outcome	presidential approval

```
df<-readRDS("./data/jop_Kriner_etal_2014.rds")
D <- "misconductdays"
Y <- "approval"
Z <- "alldaysinsession"
controls <- c("icst1", "positive", "negative", "vcaslast6mos",
              "iraqcaslast6mos", "honeymoon", "approvalt1", "ike","jfk",
              "lbj","rmn","ford","carter","reagan","bush","clinton","wbush")
cl <- NULL
FE <- NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 9.849 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.0314     0.0158     0.0157    -0.0630    -0.0010
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.1262     0.0426     0.0463    -0.2250    -0.0420
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##   105.5872   121.5394         NA   123.8176
##
## $p_iv
## [1] 1
##
```

```
## $N
## [1] 636
##
## $N_cl
## NULL
##
## $rho
## [1] 0.382
```

This paper argues that congressional investigations of the executive branch damage the president's support among the public. Our replication finds that the OLS estimate is -0.031 (with a standard error of 0.016), and 2SLS estimate is -0.126 (with a standard error of 0.043). The replicated first-stage partial F statistic is 123.82. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Lorentzen et al. (2014)

Replication Summary	
Unit of analysis	city
Treatment	large firm dominance in 2007
Instrument	same variable measured in 1999
Outcome	pollution information transparency index

```
df<-readRDS("./data/jop_Lorentzen_2014.rds")
D <- "lfd2007"
Y <- "pitiave3"
Z <- "lfd99"
controls <- c("lbudgetrev", "lexpratio", "tertratio", "sat_air_pca")
cl <- NULL
FE <- NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 8.921 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -2.4789     1.0109     1.0670    -4.5160    -0.3920
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -6.3664     1.8619     1.7285    -9.8920    -3.3070
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    53.6182    53.4100         NA    50.4088
```

```
##
## $p_iv
## [1] 1
##
## $N
## [1] 112
##
## $N_cl
## NULL
##
## $rho
## [1] 0.5796
```

This paper argues that Chinese cities dominated by large industrial firms lagged in implementing environmental transparency. Our replication finds that the OLS estimate is -2.479 (with a standard error of 1.011), and 2SLS estimate is -6.366 (with a standard error of 1.862). The replicated first-stage partial F statistic is 50.41. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Alt et al. (2015)

Replication Summary	
Unit of analysis	individual
Treatment	unemployment expectations
Instrument	assignment to receiving an aggregate unemployment forecast
Outcome	vote intention

```
df<- readRDS("./data/jop_Alt_etal_2015.rds")
D <- "urate_fut"
Y <- "gov"
Z <- "treatment"
controls <- "urate_now"
cl <- NULL
FE <- NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 10.525 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.0131     0.0027     0.0025    -0.0180    -0.0080
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.0347     0.0139     0.0136    -0.0620    -0.0080
```

```
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    60.1863    68.9098         NA    58.6136
##
## $p_iv
## [1] 1
##
## $N
## [1] 5705
##
## $N_cl
## NULL
##
## $rho
## [1] 0.0801
```

This paper argues that unemployment expectations among unsophisticated voters decrease their vote intention. Our replication finds that the OLS estimate is -0.013 (with a standard error of 0.003), and 2SLS estimate is -0.035 (with a standard error of 0.014). The replicated first-stage partial F statistic is 58.61. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Dietrich and Wright (2015)

Replication Summary	
Unit of analysis	transition
Treatment	economic aid
Instrument	constructed Z
Outcome	transitions to multipartyism

```
df <- readRDS("./data/jop_Dietrich_2015.rds")
D <- "econaid"
Y <- "mp"
Z <- c("Iinfl3","econaid_lgdp_g", "econaid_lpop_g",
       "econaid_cwar_g", "econaid_dnmp_g",
       "econaid_dnmp2_g", "econaid_dnmp3_g")
controls <- c('lgdp', 'lpop', 'cwar', 'dmp',
              'dmp2', 'dmp3', "dnmp", "dnmp2", "dnmp3")
cl<- "cowcode"
FE <- NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 9.730 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0576     0.0259     0.0292    -0.0140     0.1000
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.1075     0.0431     0.0464     0.0030     0.1910
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    28.9900    47.6878    22.5931     2.2195
##
## $p_iv
## [1] 7
##
## $N
## [1] 370
##
## $N_cl
## [1] 44
##
## $rho
## [1] 0.6026
```

This paper argues that economic aid increases the likelihood of transition to multiparty politics in sub-Saharan Africa. Our replication finds that the OLS estimate is 0.058 (with a standard error of 0.026), and 2SLS estimate is 0.107 (with a standard error of 0.043). The replicated first-stage partial F statistic is 2.22. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Dube and Naidu (2015)

Replication Summary	
Unit of analysis	municipality*year
Treatment	changes in US funding to Colombia
Instrument	US funding in countries outside of Latin America
Outcome	the number of paramilitary attacks

```
df<-readRDS("./data/jop_Dube_etal_2015.rds")
D <- "bases6xlrmilnar_col"
Y <- "paratt"
Z <- "bases6xlrmilwnl"
controls <-"lnnewpop"
cl <- "municipality"
FE <- c("year","municipality")
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 30.190 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.1503     0.0173     0.0602     0.0480     0.2830
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.3149     0.0312     0.1255     0.0870     0.5890
##
## $F_stat
##  F.standard    F.robust   F.cluster      F.boot
##   7003.8727    810.8395 185092.5288 181429.0379
##
## $p_iv
## [1] 1
##
## $N
## [1] 16606
##
## $N_cl
## [1] 936
##
## $rho
## [1] 0.556
```

This paper argues that US military assistance leads to differential increases in attacks by paramilitaries but has no effect on guerilla attacks. Our replication finds that the OLS estimate is 0.150 (with a standard error of 0.017), and 2SLS estimate is 0.315 (with a standard error of 0.031). The replicated first-stage partial F statistic is 181429.04. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Feigenbaum and Hall (2015)

Replication Summary	
Unit of analysis	congressional district*decade
Treatment	localized trade shocks in congressional districts
Instrument	Chinese exports to other economies*local exposure
Outcome	trade score based on congressional voting

```
df<-readRDS("./data/jop_Feigenbaum_etal_2015.rds")
D <-"x"
Y <- "tradescore"
Z <- "z"
controls <- c("dem_share")
cl <- "state_cluster"
```

```r
FE <- "decade"
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 19.710 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.1080     0.2542     0.3182    -0.7190     0.5390
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.6976     0.3346     0.4051    -1.5050     0.1300
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##  1189.3393   204.4798    75.5233    75.1928
##
## $p_iv
## [1] 1
##
## $N
## [1] 862
##
## $N_cl
## [1] 94
##
## $rho
## [1] 0.7622
```

This paper argues that Chinese import shocks causes legislators to vote in a more protectionist direction on trade bills but cause no change in their voting on all other bills. Our replication finds that the OLS estimate is -0.108 (with a standard error of 0.254), and 2SLS estimate is -0.698 (with a standard error of 0.335). The replicated first-stage partial F statistic is 75.19. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Acharya et al. (2016)

Replication Summary	
Unit of analysis	county
Treatment	slave proportion in 1860
Instrument	measures of the environmental suitability for growing cotton
Outcome	proportion Democrat

```r
df<-readRDS("./data/jop_Acharya_etal_2016.rds")
Y <- "dem"
D <-"pslave1860"
Z <- "cottonsuit"
controls <- c("x2", "rugged", "latitude", "x2", "longitude", "x3","x4", "water1860")
cl <- NULL
FE <- 'code'
weights<-"sample.size"
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 19.720 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.0318     0.0301     0.0501    -0.1290     0.0650
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.2766     0.1044     0.1483    -0.5900     0.0030
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##   106.4957    37.6527         NA    35.1812
##
## $p_iv
## [1] 1
##
## $N
## [1] 1120
##
## $N_cl
## NULL
##
## $rho
## [1] 0.2973
```

This paper argues that slavery decreased support for Democrats in the American South. Our replication finds that the OLS estimate is -0.032 (with a standard error of 0.030), and 2SLS estimate is -0.277 (with a standard error of 0.104). The replicated first-stage partial F statistic is 35.18. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Henderson and Brooks (2016)

Replication Summary	
Unit of analysis	district*year
Treatment	Democratic vote margins

Replication Summary	
Instrument	rain around Election
Outcome	incumbent roll call positioning

```
df<- readRDS("./data/jop_Henderson_etal_2016.rds")
D <- "dose_iv"
Y <- "vote_iv"
Z <- c("rain_day", "rain_day_prev")
controls <- c("d_inc", "dist_prev", "midterm", "pres_party", "black",
              "construction", "educ", "minc", "farmer", "forborn",
              "gvtwkr", "manuf", "pop", "unempld", "urban", "retail",
              "sos", "gov", "comp_cq", "redistricted", "dose_prv", "vote_prv")
cl <- "fe_id_num"
FE <- NULL; weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 19.579 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.0048     0.0303     0.0350    -0.0690     0.0680
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -1.3782     0.4287     0.5220    -2.5540    -0.4760
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    20.8100    18.9358    19.1716    19.1236
##
## $p_iv
## [1] 2
##
## $N
## [1] 6234
##
## $N_cl
## [1] 2639
##
## $rho
## [1] 0.0816
```

This paper argues that losses in vote support for Democratic candidates shifts incumbents rightward in their roll call positions in subsequent Congresses. Our replication finds that the OLS estimate is -0.005 (with a standard error of 0.030), and 2SLS estimate is -1.378 (with a standard error of 0.429). The replicated first-stage partial F statistic is 19.12. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Johns and Pelc (2016)

Replication Summary	
Unit of analysis	WTO dispute
Treatment	the number third parties
Instrument	trade stake of the rest of the world
Outcome	becoming a third party

```
df<-readRDS("./data/jop_Johns_etal_2016.rds")
D='third_num_excl'
Y='thirdparty'
Z='ln_ROW_before_disp'
controls=c("ln_gdpk_partner", "ln_history_third", "ln_history_C",
    "Multilateral", "trade_before_dispute",  "ARTICLEXXII")
cl <- NULL
FE <- NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 9.742 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0190     0.0014     0.0016     0.0160     0.0220
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.0809     0.0299     0.0400    -0.1740    -0.0380
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    16.9224    18.1200         NA    18.6068
##
## $p_iv
## [1] 1
##
## $N
## [1] 2462
##
## $N_cl
## NULL
##
## $rho
## [1] 0.0828
```

This paper argues that the number for existing third parties in WTO disputes lowers a country's likelihood to join as a third party. Our replication finds that the OLS estimate is 0.019 (with a standard

error of 0.001), and 2SLS estimate is -0.081 (with a standard error of 0.030). The replicated first-stage partial F statistic is 18.61. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Rozenas (2016)

Replication Summary	
Unit of analysis	country*year
Treatment	economic crises
Instrument	proximity-weighted economic shock
Outcome	measure of office insecurity

```
df<-readRDS("./data/jop_Rozenas_2016.rds")
D <- "crisis"
Y <- "mirt"
Z <- "growthcw_lag"
controls <- c("lambda", "i1", "gdp.gle_t1", "agedem_t1", "nelda19",
              "legelec", "polityw_lag","t1","t2","t3")
cl <- FE <- "country"
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 20.816 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.2381     0.0665     0.0825    -0.3950    -0.0630
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.4353     0.2907     0.4356    -1.3880     0.3270
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    45.0029    31.5596    23.8940    22.1600
##
## $p_iv
## [1] 1
##
## $N
## [1] 930
##
## $N_cl
## [1] 114
##
```

```
## $rho
## [1] 0.2301
```

This paper argues that political insecurity (e.g. due to economic crisis) leads to more electoral manipulation. Our replication finds that the OLS estimate is -0.238 (with a standard error of 0.067), and 2SLS estimate is -0.435 (with a standard error of 0.291). The replicated first-stage partial F statistic is 22.16. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Schleiter and Tavits (2016)

Replication Summary	
Unit of analysis	election
Treatment	opportunistic election calling
Instrument	prime Minister dissolution power
Outcome	vote share of Prime Minister's party

```
df<- readRDS("./data/jop_Schleiter_etal_2016.rds")
D <- "term2"
Y <- "pm_voteshare_next"
Z <- "disspm"
controls <- c("pm_voteshare", "gdp_chg1yr", "cpi1yr",  "dumcpi1yr")
cl <- "countryn"
FE <- "decade"
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 20.600 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     3.0828     1.3214     1.1934     1.4610     6.1740
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     5.0282     2.1732   108.7793     0.7140    21.2970
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##   107.0322    75.6881    57.1949    24.2783
##
## $p_iv
## [1] 1
##
## $N
## [1] 191
```

```
##
## $N_cl
## [1] 25
##
## $rho
## [1] 0.6117
```

This paper argues that opportunistic elections will increase PM party vote share in the next one. Our replication finds that the OLS estimate is 3.083 (with a standard error of 1.321), and 2SLS estimate is 5.028 (with a standard error of 2.173). The replicated first-stage partial F statistic is 24.28. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Charron et al. (2017)

Replication Summary	
Unit of analysis	region
Treatment	more developed bureaucracy
Instrument	proportion of Protestant residents in a region; aggregate literacy in 1880
Outcome	percent of single bidders in procurement contracts

```
df <- readRDS("./data/jop_Charron_etal_2017.rds")
D <- "pubmerit"
Y <- "lcri_euc1_r"
Z <- c("litrate_1880", 'pctprot')
controls <- c("logpopdens", "logppp11", "trust", "pctwomenparl")
cl <- "country"
FE <- NULL
weights<-"eu_popweights"
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 10.128 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.0900     0.0127     0.0233    -0.1110    -0.0170
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.1472     0.0270     0.1125    -0.3500     0.0520
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    27.8775    23.2292    36.2651     6.5219
##
## $p_iv
## [1] 2
```

```
##
## $N
## [1] 175
##
## $N_cl
## [1] 20
##
## $rho
## [1] 0.4992
```

This paper argues that more developed bureaucracy reduces corruption risks. Our replication finds that the OLS estimate is -0.090 (with a standard error of 0.013), and 2SLS estimate is -0.147 (with a standard error of 0.027). The replicated first-stage partial F statistic is 6.52. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Grossman et al. (2017)

Replication Summary	
Unit of analysis	region * year
Treatment	government fragmentation
Instrument	the number of distinct landmasses; length of medium and small streams; over-time variation in the number of regional governments
Outcome	public goods provision

```
df<-readRDS("./data/jop_Grossman_2017.rds")
Y <- "ServicesCA"
D <- "ladminpc_l5"
Z <- c("lmeanMINUSi_adminpc_l6", "lmeanMINUSi_adminpc2_l6",
       "herf", "herf2", "llength", "llength2")
controls <- c("lpop_l", "wdi_urban_l", "lgdppc_l", "conflict_l",
              "dpi_state_l", "p_polity2_l",
              "loilpc_l", "aid_pc_l","al_ethnic")
cl <- "ccodecow"
FE <- "year"
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 21.652 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0364     0.0376     0.1292    -0.1900     0.3110
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
```

```
##     0.4164     0.0713     0.2063    -0.1190     0.6920
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    39.9978    40.9874    11.9593     1.1970
##
## $p_iv
## [1] 6
##
## $N
## [1] 518
##
## $N_cl
## [1] 31
##
## $rho
## [1] 0.581
```

This paper argues that government fragmentation leads to more public good provision. Our replication finds that the OLS estimate is 0.036 (with a standard error of 0.038), and 2SLS estimate is 0.416 (with a standard error of 0.071). The replicated first-stage partial F statistic is 1.20. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Lerman et al. (2017)

Replication Summary	
Unit of analysis	individual
Treatment	public versus only private health insurance
Instrument	born 1946 or 1947
Outcome	support ACA

```
df<-readRDS("./data/jop_Lerman_2017.rds")
Y <-'suppafford'
D <-'privpubins3r'
Z <-'byr4647'
controls<-c( 'rep', 'ind', 'con', 'mod',
              'ideostrength', 'hcsocial', 'fininsur',
             'healthcaresupport', 'child18', 'male',
             'married', 'labor', 'mobility', 'homeowner',
             'religimp','employed', 'votereg', 'vote08',
             'black', 'hispanic2', 'military', 'educ',
              'fincome', 'newsint', 'publicemp', 'bornagain')
cl<-NULL
FE<-NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
```

```
## Parallelising 1000 reps on 15 cores
## Bootstrap took 17.186 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0093     0.0109     0.0109    -0.0120     0.0310
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0459     0.0229     0.0227     0.0030     0.0900
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##   1272.162   1194.659         NA   1294.247
##
## $p_iv
## [1] 1
##
## $N
## [1] 4389
##
## $N_cl
## NULL
##
## $rho
## [1] 0.4752
```

This paper argues that personal experience with public health insurance programs exerts a positive influence on attitudes toward both Medicare and the Affordable Care Act. Our replication finds that the OLS estimate is 0.009 (with a standard error of 0.011), and 2SLS estimate is 0.046 (with a standard error of 0.023). The replicated first-stage partial F statistic is 1294.25. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Stewart and Liou (2017)

Replication Summary	
Unit of analysis	insurgency*year
Treatment	foreign territory
Instrument	log total border length and the total number of that state's neighbors
Outcome	civilian casualties

```
df <- readRDS("./data/jop_Stewart_2017.rds")
D <- "exterrdum_low"
Y <- "oneside_best_log"
Z <- "total_border_ln"
controls <- c("bd_log", "terrdum", "strengthcent_ord", "rebstrength_ord",
              'nonmilsupport', 'rebestsize', 'l1popdensity',
              'l1gdppc_log','l1gdppc_change')
```

```
cl <- NULL
FE <- c("year","countrynum")
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 20.990 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.8030     0.2845     0.3270     0.1310     1.4370
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     1.1929     1.0236     7.6582    -0.1360     2.6760
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    33.9859    99.3150         NA    52.3256
##
## $p_iv
## [1] 1
##
## $N
## [1] 466
##
## $N_cl
## NULL
##
## $rho
## [1] 0.2786
```

This paper argues that when rebel groups control foreign territory, they tend to cause more civilian casualties. Our replication finds that the OLS estimate is 0.803 (with a standard error of 0.284), and 2SLS estimate is 1.193 (with a standard error of 1.024). The replicated first-stage partial F statistic is 52.33. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

West (2017)

Replication Summary	
Unit of analysis	individual
Treatment	Obama win
Instrument	IEM (prediction market) price
Outcome	political efficacy

```
df<- readRDS("./data/jop_West_2017.rds")
D <- "obama"
Y <- "newindex"
Z <- "avgprice"
controls <- c("partyd1", "partyd2", "partyd3",
              "partyd4", "partyd5", "wa01_a", "wa02_a",
              "wa03_a", "wa04_a", "wa05_a",   "wfc02_a",
              "ra01_b",  "rd01", "wd02_b", "rkey",
              "wave_1", "dt_w12", "dt_w12_2")
cl <- c("state")
FE <- c("state","religion")
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 22.638 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0358     0.0110     0.0119     0.0120     0.0610
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.2073     0.0853     0.0793     0.0580     0.3730
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    41.7917    37.8652    49.2358    46.4260
##
## $p_iv
## [1] 1
##
## $N
## [1] 2283
##
## $N_cl
## [1] 43
##
## $rho
## [1] 0.1362
```

This paper finds that African American efficacy increases with Obama's perceived probability of success, while white Democrats who prefer Obama are unaffected. Our replication finds that the OLS estimate is 0.036 (with a standard error of 0.011), and 2SLS estimate is 0.207 (with a standard error of 0.085). The replicated first-stage partial F statistic is 46.43. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Bhavnani and Lee (2018)

Replication Summary	
Unit of analysis	district*period
Treatment	bureaucrats' embeddedness
Instrument	early-career job assignment
Outcome	proportion of villages with high schools

```
df <-readRDS("./data/jop_Bhavnani_etal_2018.rds")
D <- "ALLlocal"
Y <- "Phigh"
Z <- "EXALLlocal"
controls <- c("ALLbachdivi", "lnnewpop", "lnnvill", "p_rural", "p_work",
              "p_aglab", "p_sc", "p_st", "lnmurderpc", "stategov", "natgov")
cl <- NULL
FE<- c('distcode71',"year")
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 22.014 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0195     0.0085     0.0103    -0.0020     0.0380
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0220     0.0121     0.0140    -0.0080     0.0470
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##   243.2947   215.8574         NA   119.1893
##
## $p_iv
## [1] 1
##
## $N
## [1] 569
##
## $N_cl
## NULL
##
## $rho
## [1] 0.7002
```

This paper argues that Indian locally embedded bureaucrats enhance public goods provisioning when they can be held accountable by the public. Our replication finds that the OLS estimate is 0.019 (with a standard error of 0.009), and 2SLS estimate is 0.022 (with a standard error of 0.012). The replicated first-stage partial F statistic is 119.19. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Cirone and Van Coppenolle (2018)

Replication Summary	
Unit of analysis	deputy*year
Treatment	budget committee service
Instrument	random assignment of budget incumbents to bureaux
Outcome	legislator sponsorship on a budget bill

```
df<- readRDS("./data/jop_Cirone_etal_2018.rds")
D <- "budget"
Y <- "F1to5billbudgetdummy"
Z <- "bureauotherbudgetincumbent"
controls <- c("budgetincumbent", "cummyears", "cummyears2",
              "age", "age2", "permargin", "permargin2",
              "inscrits", "inscrits2", "proprietaire",
              "lib_all", "civil", "paris")
cl <- "id"
FE <- "year"
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 29.708 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0305     0.0154     0.0189    -0.0040     0.0710
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.6341     0.2679     0.2825     0.1340     1.2550
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    32.1302    34.2557    33.1259    32.6949
##
## $p_iv
## [1] 1
##
## $N
## [1] 8147
##
## $N_cl
## [1] 1330
##
## $rho
## [1] 0.0628
```

This paper argues that in the French Third Republic, appointments to budget committee service increased legislative entrepreneurship concerning budget legislation but not other types. Our replication finds that the OLS estimate is 0.030 (with a standard error of 0.015), and 2SLS estimate is 0.634 (with a standard error of 0.268). The replicated first-stage partial F statistic is 32.69. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Arias and Stasavage (2019)

Replication Summary	
Unit of analysis	country*year
Treatment	government expenditures
Instrument	trade shock × UK bond yield
Outcome	regular leader turnover

```
# Variables are already residualized against controls, fixed effects, and unit-specific trends
df<-readRDS("./data/jop_Arias_etal_2019.rds")
Y <- "regular_res"
D <- "dexpenditures_res"
Z <- "interact_res"
controls <- NULL
cl<-"ccode"
FE<-NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 11.180 sec.

## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##    -0.0215     0.0437     0.0429    -0.1050     0.0700
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.8282     1.4001     6.4034    -1.4410    10.6310
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##     3.0429     3.4739    10.6429     7.6183
##
## $p_iv
## [1] 1
##
## $N
## [1] 2745
##
```

```
## $N_cl
## [1] 31
##
## $rho
## [1] 0.0333
```

This paper argues that fiscal austerity led to more political turnovers. Our replication finds that the OLS estimate is -0.021 (with a standard error of 0.044), and 2SLS estimate is 0.828 (with a standard error of 1.400). The replicated first-stage partial F statistic is 7.62. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

Pianzola et al. (2019)

Replication Summary	
Unit of analysis	individual
Treatment	smartvote use
Instrument	random assignment of the e-mail treatment
Outcome	vote intentions

```
df <- readRDS("./data/jop_Pianzola_etal_2019.rds")
D <- "smartvote"
Y <- "diff_top_ptv"
Z <- "email"
controls <- NULL
cl <- NULL
FE <- NULL
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 10.696 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.0805     0.0586     0.0681    -0.0640     0.2080
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.7550     0.3789     0.3854     0.0900     1.5620
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##    46.7293    46.7612         NA    44.2650
##
## $p_iv
## [1] 1
##
```

```
## $N
## [1] 1775
##
## $N_cl
## NULL
##
## $rho
## [1] 0.1602
```

This paper argues that usage of the Swiss VAA smartvote strengthened the vote intention for the most preferred party and increased the number of parties considered as potential vote options. Our replication finds that the OLS estimate is 0.081 (with a standard error of 0.059), and 2SLS estimate is 0.755 (with a standard error of 0.379). The replicated first-stage partial F statistic is 44.27. The F statistic and standard errors are estimated using bootstrap of 1,000 replications at the 5% level.

Ziaja (2020)

Replication Summary	
Unit of analysis	country*year
Treatment	number of democracy donors
Instrument	constructed instrument
Outcome	democracy scores

```
df <-readRDS("./data/jop_Ziaja_2020.rds")
D <- "l.CMgnh"
Y <- "v2x.polyarchy.n"
Z <- "l.ZwvCMgwh94"
controls <-c("l.pop.log.r", "l.gdpcap.log.r", "l.war25")
cl<- "cnamef"
FE<- c("cnamef", "periodf")
weights<-NULL
(bootres=boot_run(data=df, Y=Y, D=D, Z=Z, controls=controls, FE =FE,
                  cl =cl,weights=weights))
```

```
## Bootstrapping:
## Parallelising 1000 reps on 15 cores
## Bootstrap took 21.001 sec.
```

```
## $est_ols
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.8746     0.0768     0.2016     0.4300     1.2350
##
## $est_2sls
##       Coef       SE.t       SE.b  CI.b 2.5% CI.b 97.5%
##     0.8726     0.1311     0.4143    -0.1290     1.5290
##
## $F_stat
## F.standard   F.robust  F.cluster     F.boot
##  1158.1467   775.0850   199.9166   206.0612
```

```
##
## $p_iv
## [1] 1
##
## $N
## [1] 2367
##
## $N_cl
## [1] 130
##
## $rho
## [1] 0.586
```

This paper argues that more democracy donors increase a country's democracy scores. Our replication finds that the OLS estimate is 0.875 (with a standard error of 0.077), and 2SLS estimate is 0.873 (with a standard error of 0.131). The replicated first-stage partial F statistic is 206.06. The F statistic and standard errors are estimated usingblock bootstrap of 1,000 replications at the 5% level.

References

Acharya, A., Blackwell, M., and Sen, M. (2016). The political legacy of american slavery. *The Journal of Politics*, 78(3):621–641. Publisher: University of Chicago Press Chicago, IL. Cited on pages 2 and 57.

Alt, J., Marshall, J., and Lassen, D. (2015). Credible sources and sophisticated voters: when does new information induce economic voting? *The Journal of Politics*, 78(2):327–342. Publisher: University of Chicago Press Chicago, IL. Cited on pages 2 and 53.

Arias, E. and Stasavage, D. (2019). How large are the political costs of fiscal austerity? *The Journal of Politics*, 81(4):1517–1522. Cited on pages 2 and 71.

Barth, E., Finseraas, H., and Moene, K. (2015). Political reinforcement: how rising inequality curbs manifested welfare generosity. *American Journal of Political Science*, 59(3):565–577. Publisher: Wiley Online Library. Cited on pages 1 and 25.

Bhavnani, R. and Lee, A. (2018). Local embeddedness and bureaucratic performance: evidence from india. *The Journal of Politics*, 80(1):71–87. Publisher: University of Chicago Press Chicago, IL. Cited on pages 2 and 68.

Blattman, C., Hartman, A., and Blair, R. (2014). How to promote order and property rights under weak rule of law? an experiment in changing dispute resolution behavior through community education. *American Political Science Review*, page 100–120. Publisher: JSTOR. Cited on pages 1 and 7.

Carnegie, A. and Marinov, N. (2017). Foreign aid, human rights, and democracy promotion: Evidence from a natural experiment. *American Journal of Political Science*, 61(3):671–683. Publisher: Wiley Online Library. Cited on pages 2 and 30.

Charron, N., Dahlström, C., Fazekas, M., and Lapuente, V. (2017). Careers, connections, and corruption risks: Investigating the impact of bureaucratic meritocracy on public procurement processes. *The Journal of Politics*, 79(1):89–104. Publisher: University of Chicago Press Chicago, IL. Cited on pages 2 and 63.

Charron, N. and Lapuente, V. (2013). Why do some regions in europe have a higher quality of government? *The Journal of Politics*, 75(3):567–582. Publisher: Cambridge University Press New York, USA. Cited on pages 2 and 47.

Chong, A., León-Ciliotta, G., Roza, V., Valdivia, M., and Vega, G. (2019). Urbanization patterns, information diffusion, and female voting in rural paraguay. *American Journal of Political Science*, 63(2):323–341. Publisher: Wiley Online Library. Cited on pages 2 and 39.

Cirone, A. and Van Coppenolle, B. (2018). Cabinets, committees, and careers: the causal effect of committee service. *The Journal of Politics*, 80(3):948–963. Publisher: University of Chicago Press Chicago, IL. Cited on pages 2 and 70.

Colantone, I. and Stanig, P. (2018). Global competition and brexit. *American political science review*, 112(2):201–218. Cited on pages 1 and 12.

Colantone, I. and Stanig, P. (2018). The trade origins of economic nationalism: Import competition and voting behavior in western europe. *American Journal of Political Science*, 62(4):936–953. Publisher: Wiley Online Library. Cited on pages 2 and 36.

Coppock, A. and Green, D. (2016). Is voting habit forming? new evidence from experiments and regression discontinuities. *American Journal of Political Science*, 60(4):1044–1062. Publisher: Wiley Online Library. Cited on pages 2 and 27.

Croke, K., Grossman, G., Larreguy, H., and Marshall, J. (2016). Deliberate disengagement: How education can decrease political participation in electoral authoritarian regimes. *American Political Science Review*, 110(3):579–600. Publisher: Cambridge University Press. Cited on pages 1 and 8.

De La O, A. (2013). Do conditional cash transfers affect electoral behavior? evidence from a randomized experiment in mexico. *American Journal of Political Science*, 57(1):1–14. Publisher: Wiley Online Library. Cited on pages 1 and 20.

Dietrich, S. and Wright, J. (2015). Foreign aid allocation tactics and democratic change in africa. *The Journal of Politics*, 77(1):216–234. Publisher: University of Chicago Press Chicago, IL. Cited on pages 2 and 54.

Dorsch, M. and Maarek, P. (2019). Democratization and the conditional dynamics of income distribution. *American Political Science Review*, 113(2):385–404. Publisher: Cambridge University Press. Cited on pages 1 and 17.

Dower, P., Finkel, E., Gehlbach, S., and Nafziger, S. (2018). Collective action and representation in autocracies: Evidence from russia's great reforms. *American Political Science Review*, 112(1):125–147. Publisher: Cambridge University Press. Cited on pages 1, 13 and 14.

Dube, O. and Naidu, S. (2015). Bases, bullets, and ballots: The effect of us military aid on political conflict in colombia. *The Journal of Politics*, 77(1):249–267. Publisher: University of Chicago Press Chicago, IL. Cited on pages 2 and 55.

Escriba-Folch, A., Meseguer, C., and Wright, J. (2018). Remittances and protest in dictatorships. *American Journal of Political Science*, 62(4):889–904. Publisher: Wiley Online Library. Cited on pages 2 and 37.

Feigenbaum, J. and Hall, A. (2015). How legislators respond to localized economic shocks: Evidence from chinese import competition. *The Journal of Politics*, 77(4):1012–1030. Publisher: University of Chicago Press Chicago, IL. Cited on pages 2 and 56.

Flores-Macias, G. and Kreps, S. (2013). The foreign policy consequences of trade: China's commercial relations with africa and latin america, 1992–2006. *The Journal of Politics*, 75(2):357–371. Publisher: Cambridge University Press New York, USA. Cited on pages 2 and 49.

Gehlbach, S. and Keefer, P. (2012). Private investment and the institutionalization of collective action in autocracies: ruling parties and legislatures. *The Journal of Politics*, 74(2):621–635. Publisher: Cambridge University Press New York, USA. Cited on pages 2 and 46.

Gerber, A., Gimpel, J., Green, D., and Shaw, D. (2011). How large and long-lasting are the persuasive effects of televised campaign ads? results from a randomized field experiment. *American Political Science Review*, page 135–150. Publisher: JSTOR. Cited on pages 1, 4 and 5.

Gerber, A., Huber, G., Meredith, M., Biggers, D., and Hendry, D. (2015). Can incarcerated felons be (re) integrated into the political system? results from a field experiment. *American Journal of Political Science*, 59(4):912–926. Publisher: Wiley Online Library. Cited on pages 1 and 26.

Gerber, A., Huber, G., and Washington, E. (2010). Party affiliation, partisanship, and political beliefs: A field experiment. *American Political Science Review*, 104(4):720–744. Publisher: Cambridge University Press. Cited on pages 1 and 3.

Goldstein, R. and You, H. (2017). Cities as lobbyists. *American Journal of Political Science*, 61(4):864–876. Publisher: Wiley Online Library. Cited on pages 2 and 31.

Grossman, G., Pierskalla, J., and Boswell Dean, E. (2017). Government fragmentation and public goods provision. *The Journal of Politics*, 79(3):823–840. Publisher: University of Chicago Press Chicago, IL. Cited on pages 2 and 64.

Hager, A. and Hilbig, H. (2019). Do inheritance customs affect political and social inequality? *American Journal of Political Science*, 63(4):758–773. Publisher: Wiley Online Library. Cited on pages 2, 41 and 42.

Hager, A., Krakowski, K., and Schaub, M. (2019). Ethnic riots and prosocial behavior: Evidence from kyrgyzstan. *American Political Science Review*, 113(4):1029–1044. Cited on pages 1 and 18.

Healy, A. and Malhotra, N. (2013). Childhood socialization and political attitudes: Evidence from a natural experiment. *The Journal of Politics*, 75(4):1023–1037. Publisher: Cambridge University Press New York, USA. Cited on pages 2 and 50.

Henderson, J. and Brooks, J. (2016). Mediating the electoral connection: The information effects of voter signals on legislative behavior. *The Journal of Politics*, 78(3):653–669. Cited on pages 2 and 58.

Johns, L. and Pelc, K. (2016). Fear of crowds in world trade organization disputes: Why don't more countries participate? *The Journal of Politics*, 78(1):88–104. Publisher: University of Chicago Press Chicago, IL. Cited on pages 2 and 60.

Kapoor, S. and Magesan, A. (2018). Independent candidates and political representation in india. *American Political Science Review*, 112(3):678–697. Publisher: Cambridge University Press. Cited on pages 1 and 15.

Kim, J. (2019). Direct democracy and women's political engagement. *American Journal of Political Science*, 63(3):594–610. Publisher: Wiley Online Library. Cited on pages 2 and 43.

Kocher, M., Pepinsky, T., and Kalyvas, S. (2011). Aerial bombing and counterinsurgency in the vietnam war. *American Journal of Political Science*, 55(2):201–218. Publisher: Wiley Online Library. Cited on pages 1 and 19.

Kriner, D. and Schickler, E. (2014). Investigating the president: Committee probes and presidential approval, 1953–2006. *The Journal of Politics*, 76(2):521–534. Publisher: Cambridge University Press New York, USA. Cited on pages 2 and 51.

Laitin, D. and Ramachandran, R. (2016). Language policy and human development. *American Political Science Review*, 110(3):457–480. Publisher: Cambridge University Press. Cited on pages 1 and 9.

Lelkes, Y., Sood, G., and Iyengar, S. (2017). The hostile audience: The effect of access to broadband internet on partisan affect. *American Journal of Political Science*, 61(1):5–20. Publisher: Wiley Online Library. Cited on pages 2 and 33.

Lerman, A., Sadin, M., and Trachtman, S. (2017). Policy uptake as political behavior: evidence from the affordable care act. *The American Political Science Review*, 111(4):755. Publisher: Cambridge University Press. Cited on pages 2 and 65.

López-Moctezuma, G., Wantchekon, L., Rubenson, D., Fujiwara, T., and Pe Lero, C. (2020). Policy deliberation and voter persuasion: Experimental evidence from an election in the philippines. *American Journal of Political Science*. Cited on pages 2 and 45.

Lorentzen, P., Landry, P., and Yasuda, J. (2014). Undermining authoritarian innovation: the power of china's industrial giants. *The Journal of Politics*, 76(1):182–194. Publisher: Cambridge University Press New York, USA. Cited on pages 2 and 52.

McClendon, G. (2014). Social esteem and participation in contentious politics: A field experiment at an lgbt pride rally. *American Journal of Political Science*, 58(2):279–290. Publisher: Wiley Online Library. Cited on pages 1 and 24.

Meredith, M. (2013). Exploiting friends-and-neighbors to estimate coattail effects. *American Political Science Review*, page 742–765. Publisher: JSTOR. Cited on pages 1 and 6.

Nellis, G. and Siddiqui, N. (2018). Secular party rule and religious violence in pakistan. *The American Political Science Review*, 112(1):49. Publisher: Cambridge University Press. Cited on pages 1 and 16.

Pianzola, J., Trechsel, A., Vassil, K., Schwerdt, G., and Alvarez, R. (2019). The impact of personalized information on vote intention: Evidence from a randomized field experiment. *The Journal of Politics*, 81(3):833–847. Publisher: The University of Chicago Press Chicago, IL. Cited on pages 2 and 72.

Ritter, E. and Conrad, C. (2016). Preventing and responding to dissent: The observational challenges of explaining strategic repression. *American Political Science Review*, 110(1):85–99. Publisher: Cambridge University Press. Cited on pages 1 and 11.

Rozenas, A. (2016). Office insecurity and electoral manipulation. *The Journal of Politics*, 78(1):232–248. Cited on pages 2 and 61.

Rueda, M. (2017). Small aggregates, big manipulation: Vote buying enforcement and collective monitoring. *American Journal of Political Science*, 61(1):163–177. Publisher: Wiley Online Library. Cited on pages 2 and 34.

Schleiter, P. and Tavits, M. (2016). The electoral benefits of opportunistic election timing. *The Journal of Politics*, 78(3):836–850. Publisher: University of Chicago Press Chicago, IL. Cited on pages 2 and 62.

Sexton, R., Wellhausen, R., and Findley, M. (2019). How government reactions to violence worsen social welfare: evidence from peru. *American Journal of Political Science*, 63(2):353–367. Publisher: Wiley Online Library. Cited on pages 2 and 44.

Spenkuch, J. and Tillmann, P. (2018). Elite influence? religion and the electoral success of the nazis. *American Journal of Political Science*, 62(1):19–36. Publisher: Wiley Online Library. Cited on pages 2 and 38.

Stewart, M. and Liou, Y. (2017). Do good borders make good rebels? territorial control and civilian casualties. *The Journal of Politics*, 79(1):284–301. Publisher: University of Chicago Press Chicago, IL. Cited on pages 2 and 66.

Stokes, L. (2016). Electoral backlash against climate policy: A natural experiment on retrospective voting and local resistance to public policy. *American Journal of Political Science*, 60(4):958–974. Publisher: Wiley Online Library. Cited on pages 2 and 28.

Tajima, Y. (2013). The institutional basis of intercommunal order: Evidence from indonesia's democratic transition. *American Journal of Political Science*, 57(1):104–119. Publisher: Wiley Online Library. Cited on pages 1 and 22.

Trounstine, J. (2016). Segregation and inequality in public goods. *American Journal of Political Science*, 60(3):709–725. Publisher: Wiley Online Library. Cited on pages 2 and 29.

Vernby, K. (2013). Inclusion and public policy: Evidence from sweden's introduction of noncitizen suffrage. *American Journal of Political Science*, 57(1):15–29. Publisher: Wiley Online Library. Cited on pages 1 and 23.

West, E. (2017). Descriptive representation and political efficacy: Evidence from obama and clinton. *The Journal of Politics*, 79(1):351–355. Publisher: University of Chicago Press Chicago, IL. Cited on pages 2 and 67.

Zhu, B. (2017). Mncs, rents, and corruption: Evidence from china. *American Journal of Political Science*, 61(1):84–99. Publisher: Wiley Online Library. Cited on pages 2 and 35.

Ziaja, S. (2020). More donors, more democracy. *The Journal of Politics*, 82(2):433–447. Publisher: The University of Chicago Press Chicago, IL. Cited on pages 2 and 73.