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

Political Analysis 32(4): 521–540 (2024)

Apoorva Lal Mackenzie Lockhart Yiqing Xu Ziwen Zu*

November 2026

Abstract

We correct Lal et al. (2024) and its Supplementary Materials (SM). Three issues affect how readers should interpret the article's results. First, for an instrument predicted from a rule with cutoffs, the first-stage F statistic reflects all variation left after adjusting for the controls. This includes variation beyond the cutoff jumps on which a fuzzy regression discontinuity (RD) design relies. We replace a major application affected by this issue with McClendon (2014), a randomized encouragement design. Second, we computed the effective F incorrectly for designs with more than one instrument. The number of designs with an effective F below 10 is unchanged. The corrected median effective F is similar in experimental and observational designs. The published median was higher in experimental ones. Third, the grid searches missed parts of the Anderson–Rubin (AR) confidence sets printed in the SM for two designs. The remaining imprecisions change no finding.

1. Overview

Table 1 lists every number printed in the article that changes beyond rounding. Sections 2–4 discuss the three issues in order of importance. Section 5 lists the remaining imprecisions in the article's descriptions and in the SM. The 67 replicated articles contribute 70 designs, the units we count unless stated otherwise.

*Corresponding author: Yiqing Xu, Associate Professor, Department of Political Science, Stanford University. Email: yiqingxu@stanford.edu. We used AI tools (Anthropic's Claude and OpenAI's Codex) to help us find and check the problems documented here, fix the software, rerun the computations, and draft the text and proofread. We checked their findings and take full responsibility for the content of this note. We thank Dan de Kadt and Andy Harris for drawing our attention to concerns about instrument strength in the example application and for sharing their early work reanalyzing the application.

Verification. We recomputed every number in this note from the replication package (Dataverse, doi:10.7910/DVN/MM5THZ) with ivDiag 1.0.8, on CRAN since 2 November 2026. The article used version 1.0.6. Version 1.0.8 fixes the computational problems in Sections 3 and 4 and several smaller software issues that change no count. With ivDiag 1.0.8, we reproduce every other analytic number in the article and the SM exactly, while bootstrap results differ only by simulation error. Each fix is described, with examples, in the changelog of the online ivDiag manual (`https://yiqingxu.org/packages/ivDiag/`).

This note's replication files contain the code, corrected design-level data and figures, and smaller items not listed here. $F_{\texttt{Eff}}$ denotes the effective F statistic (Montiel Olea and Pflueger, 2013). Design labels in `typewriter` follow the SM, and studies named without a reference are designs in the article's replication sample.

Table 1. Numbers printed in the article that change

Where	Printed	Should read
Section 1	"17%–35%" of the designs cannot reject the null hypothesis	16%–36% (Sections 3 and 4)
Sections 1 and 3, Table 1, Figure 1	114 IV articles, 38 with incomplete data; "76 (62%)" with materials; "67 out of the remaining 71"; "30 articles in the APSR, 33 articles in the AJPS"; in Table 1, AJPS 30 and 25 replicable, JOP 19 incomplete and 28 (55%) replicable	113 and 37; 76 (67%) with data; 72 with data and code, of which 67 replicated; 33 in the APSR, 29 in the AJPS; AJPS 29 and 24, JOP 18 and 29 (57%). Figure 1 should count 150 IV articles, and two of its bars change by one
Section 4.2 (Finding 1)	"9 (16%) use classic analytic SEs"	10 (17%): `Digiuseppe2022` reports the classic F but is coded as robust
Section 4.2 (Finding 1)	median $F_{\texttt{Eff}}$ "higher in experimental studies compared to nonexperimental ones (67.7 vs. 53.5)"	similar in experimental and nonexperimental studies (52.3 vs. 52.2); SM p. A-1 likewise (Section 3)
Section 4.3 (Finding 2)	"the median ratio of the analytic SEs equal to 3.8"	2.8, as in Figure 3(a)
Section 4.3 (Finding 2)	the AR test fails to reject "in 12 studies (17%)"	11 (16%) (Section 4)
Section 4.3 (Finding 2)	bootstrap-c fails to reject in "19 (27%)" designs	21 (30%), as in Table 3, which matches the stored estimates; two designs sit at $p \approx 0.05$, so the count is 19 or 21 depending on the bootstrap run
Section 4.3 (Finding 2)	tF procedure: "54 studies that use single IVs"; "19 studies (35%)" not significant; "7 studies (13%)" switch to insignificant	53 (Section 3); 19 (36%); 9 (17%), since 7 counts only designs with $F_{\texttt{Eff}} \geq 10$
Section 4.4 (Finding 3)	"40% of the studies consider the OLS results to be their main findings"	51% (36 of 70 designs)
Table 3	experimental, observational and all designs: median effective F 67.7, 53.5, 53.5; tF $p > 0.05$ .400, .341, .352; AR $p > 0.05$ .417, .121, .171	52.3, 52.2, 52.2; .444, .341, .358; .417, .103, .157 (Sections 3 and 4)

2. The first-stage F of rule-based instruments

Some instruments are predicted from a rule with cutoffs. In Rueda (2017), used in the article's Figure 6, Colombian law caps the number of voters per polling station at 400 (2006 and 2007), 500 (2010) or 350 (2011). With a cap of 400, a municipality with 790 registered voters gets two stations of 395 voters, and one with 810 gets three stations of 270. Rueda instruments the log of actual average size with the log of this predicted size, controlling for log registered voters and log population. Rueda estimates an effect of -0.98 (SE 0.14) on citizens' reports of vote buying. In our replication, $F_{\texttt{Eff}}$ is 8,598, and we read it as a strong instrument. Yet in a working paper, de Kadt and Harris (2026) show that the actual number of stations barely changes at the cutoffs, raising concerns about the instrument's strength.[1]

What the first-stage F measures. The first-stage F statistic measures how strongly the instrument predicts the treatment using all variation left after adjusting for the controls. A fuzzy RD design relies only on the cutoff jumps, which its continuity assumption makes as good as random. Let $\mathbf{y}$, $\mathbf{d}$ and $\mathbf{z}$ denote the outcome, treatment and instrument. Let $\mathbf{M_W}$ residualize on the researcher's controls $\mathbf{W}$, and $\mathbf{M}_{\text{RD}}$ on the larger set $\mathbf{W}_{\text{RD}} \supseteq \mathbf{W}$ used in a fuzzy RD design. The RD control set adds a polynomial in the distance to the nearest cutoff on each side and cutoff fixed effects (Hahn et al., 2001; Imbens and Lemieux, 2008).

We assume one common sample and common weights for every regression. With weights, read each vector as premultiplied by their square roots. The instrument left after $\mathbf{W}$ splits into two orthogonal parts,

$$\tilde{\mathbf{z}} = \mathbf{M_W z} = \underbrace{\mathbf{M}_{\text{RD}}\mathbf{z}}_{\mathbf{z}^J\text{: the jumps}} + \underbrace{(\mathbf{M_W} - \mathbf{M}_{\text{RD}})\mathbf{z}}_{\mathbf{z}^V\text{: everything else}}, \tag{1}$$

where $\mathbf{z}^J$ contains only the jumps when $\mathbf{W}_{\text{RD}}$ spans the smooth part of the instrument. The 2SLS estimate then combines the fuzzy RD estimate and an estimate from the remaining variation,

$$\hat{\beta}_{\mathbf{W}} = \frac{\mathbf{y}'\tilde{\mathbf{z}}}{\mathbf{d}'\tilde{\mathbf{z}}} = \omega\,\hat{\beta}_{\text{RD}} + (1-\omega)\,\hat{\beta}_V, \qquad \omega = \frac{\mathbf{d}'\mathbf{z}^J}{\mathbf{d}'\tilde{\mathbf{z}}}, \tag{2}$$

where $\hat{\beta}_{\text{RD}} = \mathbf{y}'\mathbf{z}^J/\mathbf{d}'\mathbf{z}^J$ is the 2SLS estimate with the RD controls and $\hat{\beta}_V = \mathbf{y}'\mathbf{z}^V/\mathbf{d}'\mathbf{z}^V$. The weights sum to one but need not lie between 0 and 1. With a common first-stage error variance (the residual variance of the article's first stage), the first-stage F of $\tilde{\mathbf{z}}$ is at most the sum of the F statistics of its two parts. This gives no lower bound on the F of the jumps. The first-stage F can be in the thousands while the jumps alone have an F near zero.

When the jumps are weak and carry little of the instrument's variance, ω is near zero. The remaining variation then dominates the strong first stage and the 2SLS estimate. That variation identifies the effect only if there is a separate reason to treat it as unrelated to other causes of the outcome. Continuity at the cutoffs does not provide one.

[1] They shared a draft of their paper with us and raised concerns, which prompted our reanalysis.

Rueda's instrument. The predicted size varies with three features of the rule. The cap changes between elections and shifts every municipality at once. Predicted size also varies with the position of registered voters between two cutoffs and jumps at a cutoff. The jump is 0.41 in logs when registered voters cross twice the cap but 0.05 when they cross twenty times the cap. Rueda's controls absorb none of the three on their own. With a linear term on each side of the nearest cutoff and cutoff fixed effects, $\mathbf{W}_{\text{RD}}$ absorbs the cap and the position exactly, so $\mathbf{z}^J$ is the residualized jump.

After Rueda's controls, the residualized log cap explains 94% of the variance of $\tilde{\mathbf{z}}$ and the jumps carry 2.8%. We obtain $\omega = 0.004$, so (2) reads $-0.98 = 0.004 \times (-3.58) + 0.996 \times (-0.97)$. Almost all of the published estimate comes from variation outside the cutoff jumps (Figure 1).

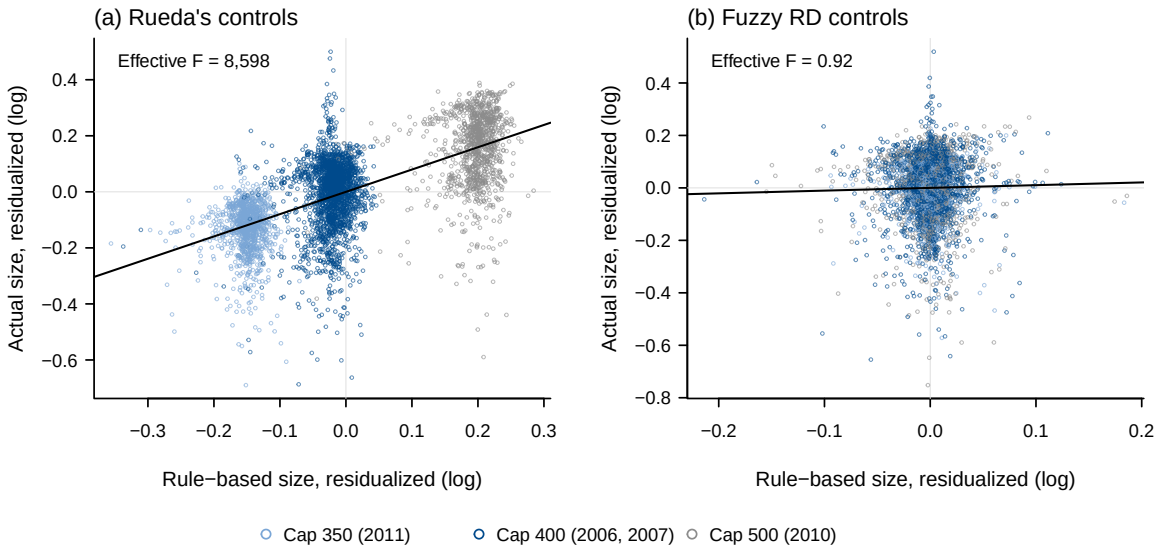


FIGURE 1. FIRST STAGE OF RUEDA (2017, TABLE 5, COLUMN 1) UNDER TWO SETS OF CONTROLS

Note: Each point is a municipality-election (N = 4,352). Both axes are residualized: (a) on log registered voters and log population, as in the article; (b) on these controls plus a linear term in the log distance to the nearest cutoff on each side and cutoff fixed effects. Lines are least squares fits. $F_{\texttt{Eff}}$ with standard errors clustered by municipality.

The jumps are weak because the rule binds loosely. Only 16% of municipality-elections have the number of stations the rule requires. Another 79% have more. Rueda (2017) notes that isolated rural areas tend to have smaller polling stations, and commissions often add stations to cut voters' travel and wait times (de Kadt and Harris, 2026). Half have more than 20 times the cap in registered voters, where crossing a cutoff lowers predicted size by less than 5%. The cap cannot substitute for the jumps because it changes only between elections. Using it compares elections that differ in more than the cap.

What changes. With only the jumps, $F_{\texttt{Eff}}$ is below 1 and the AR set is the whole real line (Table 2). The AR set is also the whole real line in every other specification we tried. With bandwidths of 0.05 to 0.2 in log units (triangular kernel, linear or quadratic terms), $F_{\texttt{Eff}}$ is below

Table 2. Polling station size and vote buying: 2SLS under two sets of controls

Variation left in the instrument	N	$F_{\texttt{Eff}}$	2SLS (SE)	AR 95% set
Cap, position, jumps (article)	4,352	8,598	−0.98 (0.14)	$[-1.26, -0.71]$
Jumps (fuzzy RD)	4,352	0.92	−3.58 (4.37)	whole line

Note: Instrument: log size predicted by the rule. Both rows control for log registered voters and log population. The fuzzy RD row adds a linear term in the log distance to the nearest cutoff on each side and cutoff fixed effects, and uses every observation, each compared with its nearest cutoff. In Rueda's discontinuity sample (`discont_f`, ±50 voters, $N = 1{,}073$), the fuzzy RD gives $F_{\texttt{Eff}} = 0.48$, −0.77 (3.96) and a whole-line AR set; with the article's controls, its $F_{\texttt{Eff}}$ is 889. Standard errors clustered by municipality.

1. Below 10.5 times the cap, where the jumps are largest, it is 2.50. In our replication, the cutoff jumps carry too little information to bound the effect of polling station size. This does not show that the effect is zero, and the paper's other analyses stand on their own.

No number in the article changes, since it reports $F_{\texttt{Eff}}$ in each author's specification. We clarify how to read $F_{\texttt{Eff}}$ for rule-based instruments and replace the example in Figure 6, which compares OLS and 2SLS estimates. We use McClendon (2014, Table 2, column 1), where random assignment to an email promising public recognition instruments for opening it (Figure 2). The 2SLS estimate of 0.31 (SE 0.09) is slightly larger than the OLS estimate of 0.28 (SE 0.03). The $F_{\texttt{Eff}}$ is 207, and all inferential methods agree.

The other three fuzzy RD designs in Table 2 use an indicator for crossing a threshold as the instrument instead of a predicted treatment. For Kim (2019), adding a linear slope on each side lowers $F_{\texttt{Eff}}$ from 915 to 68, and the instrument remains strong. Lei and Zhou (2022) already include slopes on each side.

What to report. Our checklist already advises plotting the treatment and its first-stage prediction against the covariates behind the rule (Section 5). For fuzzy RD designs, we add five steps.

- Estimate the fuzzy RD as 2SLS with the RD's own controls, a low-order polynomial in the distance to the nearest cutoff on each side and cutoff fixed effects (by period when the rule changes). Apply kernel weights within a bandwidth.
- Plot the treatment against the instrument after residualizing both on the controls. Do this once with the controls of the reported 2SLS and once with the RD's own controls, as in Figure 1. The plot shows how much of the first stage comes from the cutoff jumps.
- Report $F_{\texttt{Eff}}$ from that regression, since a weak jump distorts conventional fuzzy RD inference (Feir et al., 2016). If a 2SLS regression with fewer controls is also reported, report ω in (2), computed on the same sample and with the same weights.
- Base inference on the AR set or the tF procedure (Lee et al., 2022) with a bandwidth small enough that smoothing bias is negligible. Alternatively, use bias-aware fuzzy RD intervals (Noack and

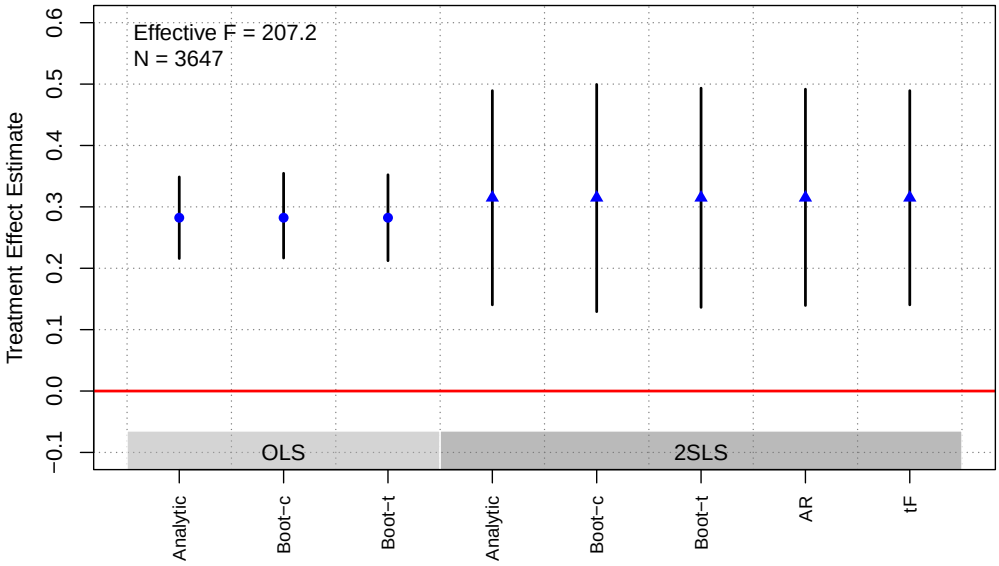


FIGURE 2. REPLACEMENT FOR THE ARTICLE'S FIGURE 6: OLS AND 2SLS ESTIMATES, MCCLENDON (2014, TABLE 2, COLUMN 1)

Note: Outcome: stated intention to attend the rally. Treatment: opening an email that promises public recognition. Instrument: random assignment to such an email. Estimates with 95% confidence intervals from each inferential method; robust standard errors; $N = 3{,}647$; ivDiag 1.0.8.

Rothe, 2024).

- If the remaining variation is meant to identify the effect, defend it as a separate design. A change in the rule between periods is a comparison of periods and requires ruling out period-specific shocks.

3. The effective F with more than one instrument

The effective F is $F_{\texttt{Eff}} = \hat{\pi}'\hat{Q}_{ZZ}\hat{\pi}/\mathrm{tr}(\hat{\Sigma}\hat{Q}_{ZZ})$. Here, $\hat{\pi}$ holds the first-stage coefficients on the instruments and $\hat{\Sigma}$ is their robust or cluster-robust analytic variance. SM Section A.1.3 calls this a cluster bootstrap. We compute $\hat{Q}_{ZZ} = \tilde{Z}'\Omega\tilde{Z}$ using $\tilde{Z}$, the instruments after partialling out the controls and fixed effects, and Ω, the regression weights. ivDiag 1.0.6 used instruments without partialling out or weighting. With one instrument, $\hat{Q}_{ZZ}$ is a scalar and cancels, so $F_{\texttt{Eff}}$ is correct. With two or more instruments, it does not cancel, and $F_{\texttt{Eff}}$ is wrong.

Seventeen designs are affected. Sixteen have two to seven instruments, and `Alt2015` has an eight-level factor instrument that enters as seven dummies. Table 3 lists the corrected values, which differ from the published ones by −61% to +84%. No design crosses $F_{\texttt{Eff}} = 10$, so the count of designs with $F_{\texttt{Eff}} < 10$ (8, or 11%) and the statement that at least 11% of designs rely on weak instruments stand. The median $F_{\texttt{Eff}}$ is now similar in experimental and observational designs (Table 1). The 17 points move in Figure 2(a) (Figure 3) and in SM Figures A1, A4(b) and A4(c).

The factor instrument of `Alt2015`. ivDiag 1.0.6 entered this factor as seven dummies in the 2SLS fit and in $F_{\texttt{Eff}}$. It used one number (1 to 8) for the other first-stage F statistics and the AR test. It also applied the tF procedure, which is defined for one instrument. With the dummies, the AR p-value is 0.175 (was 0.967). Neither the AR test nor the F statistics change any count. We remove `Alt2015`, a design with seven instruments, from the tF analysis, leaving 53 designs (Table 1). Its point moves in Figures 2(a), 2(b) and 4(c) and leaves Figure 4(d).

TABLE 3. EFFECTIVE F STATISTICS THAT CHANGE

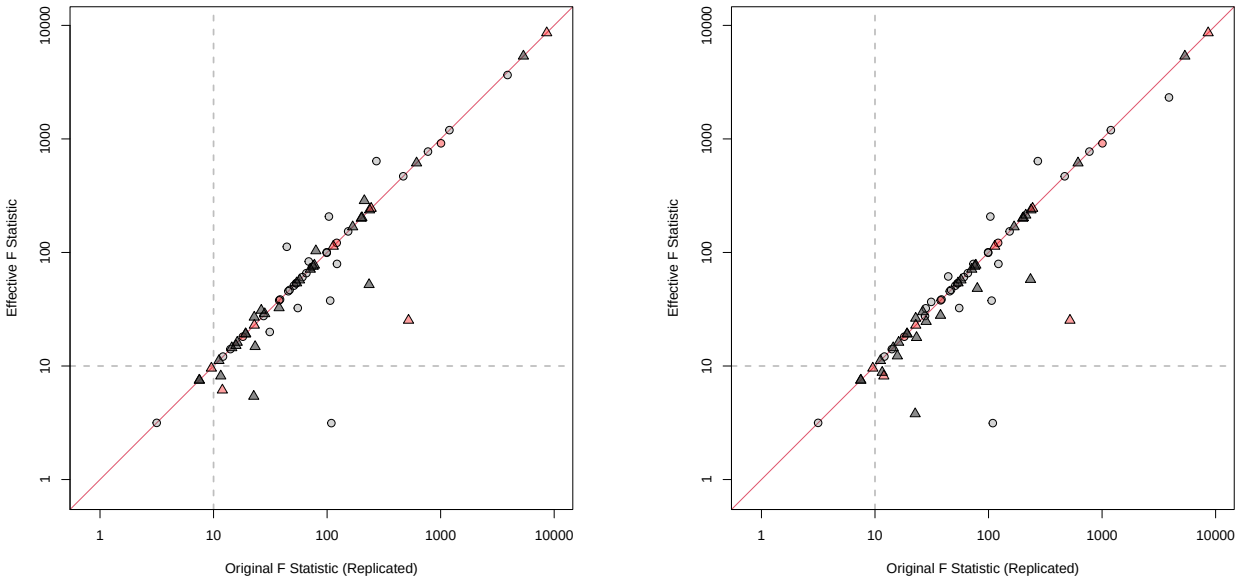
Design	Instruments	Type	Published $F_{\texttt{Eff}}$	Corrected $F_{\texttt{Eff}}$
`Alt2015`	7 (factor)	Exp.	83.3	32.2
`Barth2015`	2	Obs.	8.2	8.8
`Blattman2014`	3	Exp.	52.1	57.8
`Charron2013`	2	Obs.	20.0	36.7
`Charron2017`	2	Obs.	14.8	17.8
`Chong2019`	3	Exp.	32.6	27.9
`Dietrich2015`	7	Obs.	5.4	3.8
`Goldstein2017`	2	Obs.	15.2	12.3
`Grossman2017`	6	Obs.	6.1	8.2
`Henderson2016_a`	2	Obs.	26.9	26.3
`Henderson2016_b`	2	Obs.	30.9	30.0
`Hong2022`	2	Obs.	28.9	24.7
`Kocher2011`	2	Obs.	112.2	61.5
`Ritter2016`	2	Obs.	74.4	79.3
`Spenkuch2018`	2	Obs.	286.0	213.2
`Trounstine2016`	2	Obs.	3,654.7	2,318.7
`Vernby2013`	2	Obs.	103.4	48.2

Note: Exp.: experimental; Obs.: observational.

4. AR confidence sets

ivDiag 1.0.6 searched for the AR confidence set (Anderson and Rubin, 1949) on a grid within 10 standard errors of the 2SLS estimate. It extended the set beyond the grid according to whether the test rejected at its two end points. In two designs, the search missed a ray of the set, an interval running to $-\infty$ or $+\infty$.

Figure 4 shows the problem for Hager and Krakowski (2022). Its 2SLS estimate is 0.18 (SE 0.095), so the search covered $[-0.78, 1.13]$. As we move below -0.78, the test rejects until -1.16 and accepts again below that. The printed set, $[0.0642, \infty)$, misses the left ray of the correct set, $(-\infty, -1.1605] \cup [0.0640, \infty)$. For Gerber et al. (2010), the printed set $[0.0227, \infty)$ should read $(-\infty, -36.24] \cup [-0.0031, \infty)$. The corrected set includes zero, consistent with its AR p-value of 0.0501. In the other 68 designs, the grid's spacing moved the finite end points by less than 0.07 standard errors, and no set gains or loses zero. The AR p-values never depended on the grid because they were computed directly at zero.



(a) Published (b) Corrected

FIGURE 3. ARTICLE FIGURE 2(A): $F_{\texttt{Eff}}$ AGAINST THE REPLICATED ORIGINAL F

Note: Panel (b) uses the corrected $F_{\texttt{Eff}}$ of Table 3 and the corrected first-stage F of `Alt2015`, and is drawn with the article's code. Markers and axes as in the article.

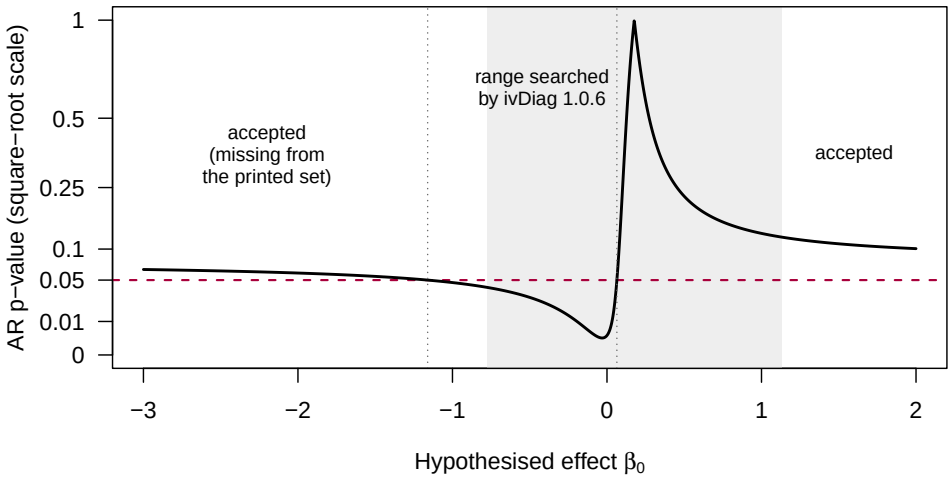
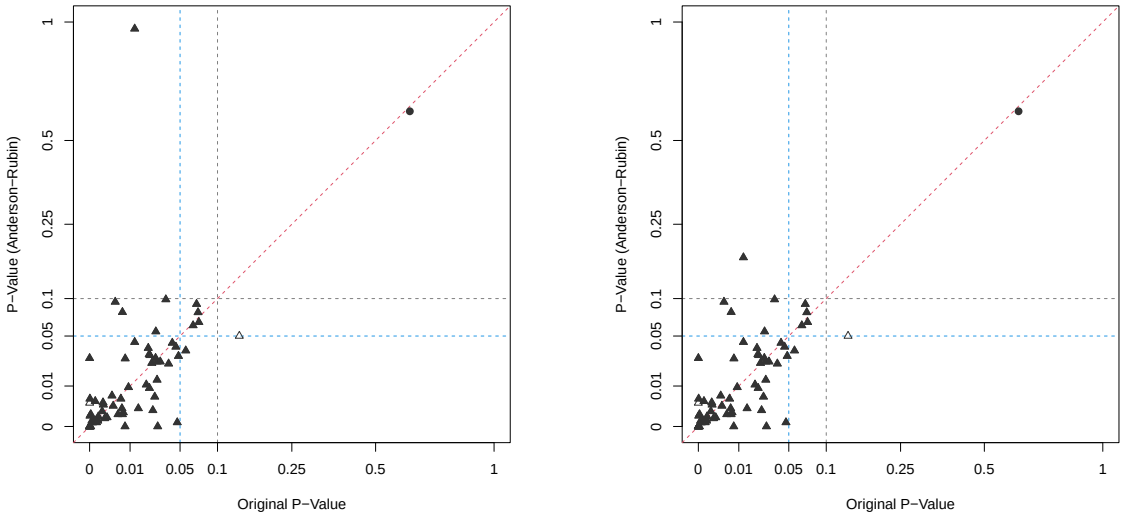


FIGURE 4. AR p-VALUE AGAINST THE HYPOTHESISED EFFECT, HAGER AND KRAKOWSKI (2022, TABLE 3, COLUMN 1)

Note: The dashed line is 0.05, and values above it are in the 95% AR set. The shaded band is the range searched by ivDiag 1.0.6. Dotted lines mark the ends of the exact set, −1.1605 and 0.0640.

The AR count. Our replication code classified a design as "AR $p > 0.05$" if its AR p-value exceeded 0.05 *or* its AR set was unbounded. An unbounded set can still exclude zero, as in Hager and Krakowski (2022), whose test rejects ($p = 0.0034$). Counting by the p-value gives 11 designs (16%) in which the AR test fails to reject, compared with the reported 12 (17%) (Table 1). Figure 5 compares the article's Figure 4(c) with the corrected version. Correcting the AR sets changes no

AR p-value. Only `Alt2015` moves, from 0.967 to 0.175, because of its factor instrument (Section 3).



(a) Published (b) Corrected

FIGURE 5. ARTICLE FIGURE 4(C): AR p-VALUE AGAINST THE ORIGINAL p-VALUE

Note: Drawn with the article's code; markers and axes as in the article.

5. Imprecisions

Two statements behind the article's findings need qualifying. In Section 4.2 (Finding 1), the original F tends to exceed the bootstrapped F under clustering. It does not tend to exceed the effective F. In Section 4.3 (Finding 2), replicated p-values are usually larger than the original ones only under bootstrap-c. All original p-values use a normal approximation.

We list the remaining imprecisions in the article's descriptions and the SM below. None of these items changes a finding of the article. Smaller items, such as labels, typos and catalog entries in the SM, are in the replication files.

- *Section 2.2, the AR test.* The article says the critical value is a percentile of χ^2_1. It is asymptotically χ^2_k with k instruments. ivDiag produced the article's AR results by comparing the statistic divided by k with $F(k, n-k-1)$. The confidence set is one of $\tilde{\tau}$, not $\tilde{\gamma}$.
- *Section 2.2 and footnote 8, the bootstrap-t.* The article describes an equal-tailed interval from an uncentred t^* and a cluster bootstrap. The code centres t^* and reports a symmetric interval. With two clustering variables, it resamples only the first.
- *Footnote 8, F_{boot}.* The formula is written with $\hat{\tau}_{2SLS}$. It should read $\hat{\pi}'\hat{\mathbb{V}}_{boot}(\hat{\pi})^{-1}\hat{\pi}/p_z$, with first-stage coefficients $\hat{\pi}$. "Section A2 of the SM" should read Section A.1.3.
- *Section 3, Dower et al. (2018).* The article says they "argue that religious polarization could only impact collective action through its impact on representation in local institutions." It should read "could only impact representation in local institutions through collective action."
- *Table 2.* The column "#Articles" counts designs, and its 17.1% is a share of designs (17.9% of articles).
- *SM Figure A3.* The reported F of `Hager2022` should read 114.8. The printed 297 is its sample size.
- *SM Section A.3.* The SM calls Nunn's placebo regression a first stage. It is a reduced form.

References

Anderson, T. W. and Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics*, 20(1):46–63.

de Kadt, D. and Harris, J. A. (2026). Does increasing electoral access cause clientelism? Working paper.

Feir, D., Lemieux, T., and Marmer, V. (2016). Weak identification in fuzzy regression discontinuity designs. *Journal of Business & Economic Statistics*, 34(2):185–196.

Gerber, A. S., Huber, G. A., and Washington, E. (2010). Party affiliation, partisanship, and political beliefs: A field experiment. *American Political Science Review*, 104(4):720–744.

Hager, A. and Krakowski, K. (2022). Does state repression spark protests? Evidence from secret police surveillance in communist Poland. *American Political Science Review*, 116(2):564–579.

Hahn, J., Todd, P., and van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209.

Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635.

Kim, J. H. (2019). Direct democracy and women's political engagement. *American Journal of Political Science*, 63(3):594–610.

Lal, A., Lockhart, M., Xu, Y., and Zu, Z. (2024). How much should we trust instrumental variable estimates in political science? Practical advice based on 67 replicated studies. *Political Analysis*, 32(4):521–540.

Lee, D. S., McCrary, J., Moreira, M. J., and Porter, J. (2022). Valid t-ratio inference for IV. *American Economic Review*, 112(10):3260–3290.

McClendon, G. 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.

Montiel Olea, J. L. and Pflueger, C. (2013). A robust test for weak instruments. *Journal of Business & Economic Statistics*, 31(3):358–369.

Noack, C. and Rothe, C. (2024). Bias-aware inference in fuzzy regression discontinuity designs. *Econometrica*, 92(3):687–711.

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