ONLINE APPENDIX
"p-Hacking and Publication Bias in Design-Based Causal Studies in Information Systems"

Appendix A. Data Collection and Description
A.1. Data Collection

Table EC.1 The Top Seven IS Journals

Journal	Abbreviation
MIS Quarterly	MISQ
Information Systems Research	ISR
Journal of the Association for Information Systems	JAIS
Management Science (IS Department only)	MS
Journal of Management Information Systems	JMIS
Journal of Information Technology	JIT
Information Systems Journal	ISJ

Table EC.2 Methods and Their Corresponding List of Keywords

Method	**Keywords**
RCT	"random*"
DID	"difference-in-difference*"
	"differences-in-difference*"
	"difference in difference*"
	"differences in difference*"
IV	"instrumental variable*"
SC	"synthetic control"
RD	"regression discontinuity"

Notes: We follow Brodeur et al. (2020) for DID, IV, and RD regarding the search process. The asterisk (*) acts as a wildcard, allowing for the inclusion of plurals and other variations of the terms.

A.2. Data Description

Figure EC.1 Annual Growth of Causal Methods in IS Research

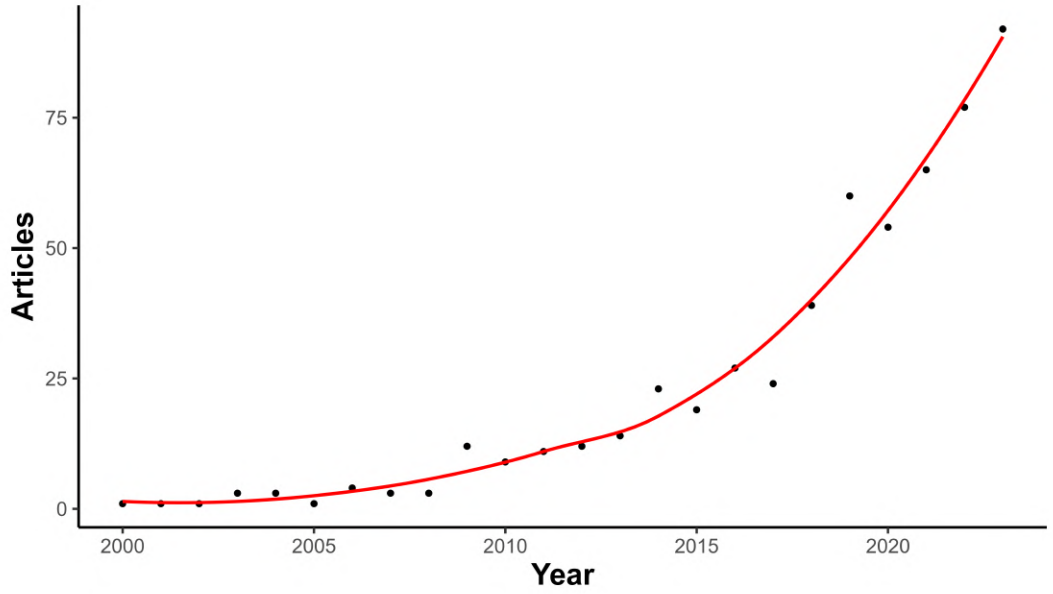
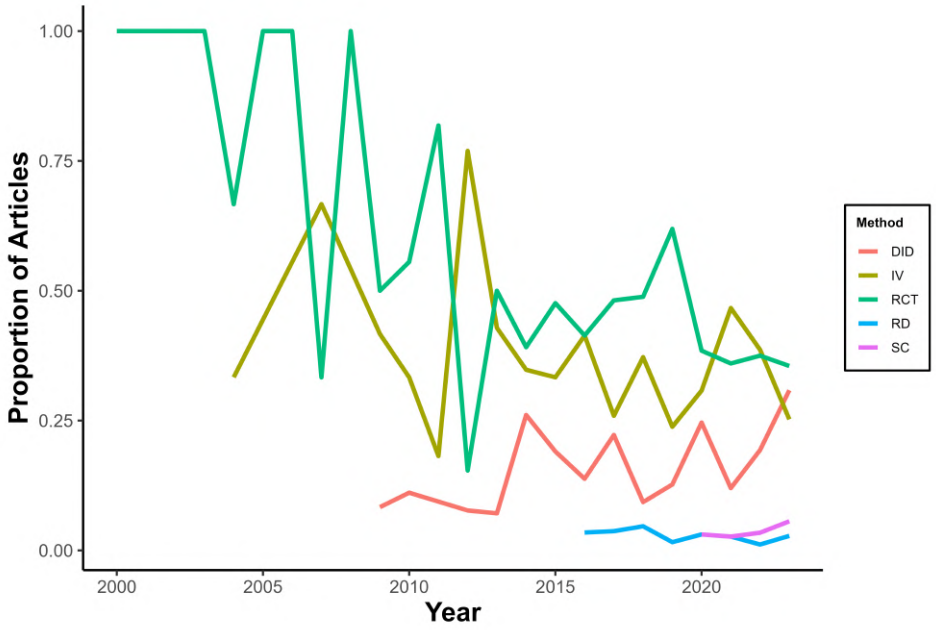


Figure EC.2 Proportions of Methods over Time



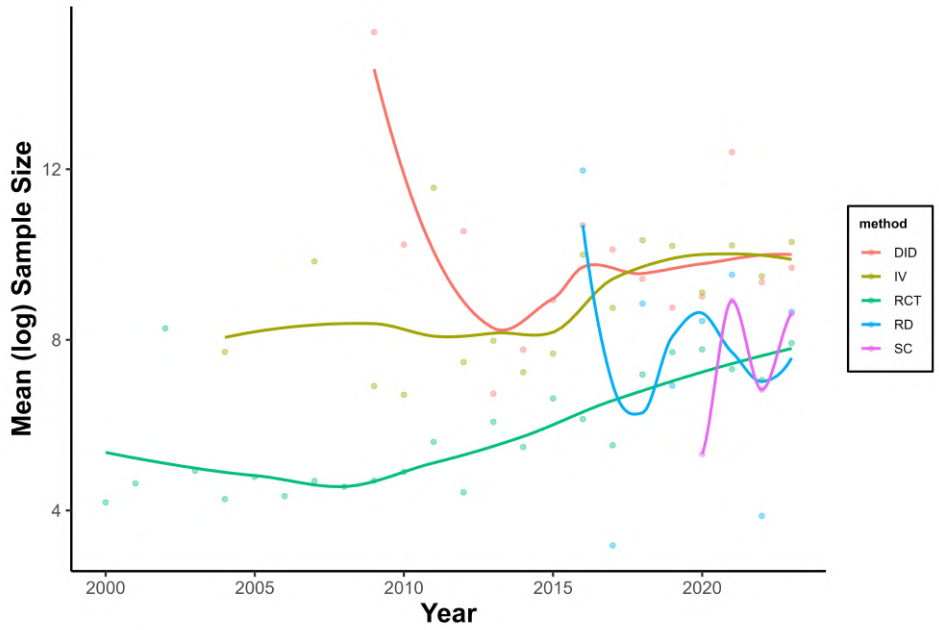


Figure EC.3 Smoothed Mean Sample Size by Year and Method Family

Author Information

For each article, we collect information about the authors and their affiliations. A manual search of curriculum vitae allowed us to gather the following data: gender, year of graduation, institution of Ph.D., institution at the time of publication, and whether the author was an editor of an IS journal at the time of publication. Thus, for each article, we can determine the proportion of female authors, the average years of experience, the percentage of authors from top Ph.D. schools, the percentage of authors currently at top institutions, and whether any of the authors have editorship experience. Table EC.3 provides descriptive statistics for author characteristics. In the sample, almost 90% of authors were at a top institution at the time of publication and/or gained their PhD from a top institution[16]. Additionally, 20% of the authors are female, and on average, they have 16 years of experience in the field after graduation.

Table EC.3 Author Characteristics

Variable	Min	Median	Mean	Max
Editor	0.00	0.00	0.33	1.00
Solo	0.00	0.00	0.01	1.00
Experience	3.00	14.00	16.08	65.00
Female	0.00	0.00	0.23	1.00
Top Current School	0.00	1.00	0.86	1.00
Top PhD School	0.00	1.00	0.86	1.00

[16] Top schools are defined in the UTD ranking. See more details in the Department Ranking section in Appendix A.

A decomposition by method reveals several key differences in author characteristics (Table EC.4). Authors using the RD method earned their Ph.D. relatively recently, as indicated by their lower average years of experience, and are more likely to be from top institutions. In contrast, authors with editorial experience are more likely to use the IV method and less likely to use the RD method. Female authors are more likely to use the SC method and less likely to use the IV and RD methods, suggesting that the SC method is more prevalent among female authors compared to other methods.

Journal articles published in top IS journals are significantly more likely to have been written by authors affiliated with and graduated from top institutions. This trend is consistent across all methods but is most pronounced for articles using the SC method. Additionally, solo-authored articles are relatively rare across all methods but are most common in the IV method. The average years of experience for authors also varies by method, with those using the SC method having the highest average experience (18 years) and those using the RD method having the least (14 years).

Table EC.4 Author Characteristics per Method

	DID	**IV**	**RCT**	**RD**	**SC**	**Total**
Editor	0.309 (0.006)	0.478 (0.007)	0.35 (0.004)	0.242 (0.019)	0.374 (0.02)	0.368 (0.003)
Solo	0.04 (0.006)	0.065 (0.006)	0.03 (0.003)	0.034 (0.015)	0 (0)	0.038 (0.002)
Experience	14.95 (0.122)	16.339 (0.118)	16.104 (0.074)	14.974 (0.342)	18.149 (0.371)	15.988 (0.055)
Female	0.235 (0.006)	0.189 (0.006)	0.272 (0.004)	0.175 (0.012)	0.361 (0.021)	0.249 (0.003)
Top Current School	0.845 (0.007)	0.898 (0.005)	0.832 (0.004)	0.829 (0.016)	0.975 (0.01)	0.849 (0.003)
Top PhD School	0.911 (0.005)	0.932 (0.004)	0.825 (0.004)	0.992 (0.004)	0.981 (0.008)	0.865 (0.003)

Department Ranking

We define the top 100 IS departments globally using the UTD Top Business School Research Rankings, specifically focusing on publications in the leading IS journals, MISQ and ISR, from 2000 to 2023[17]. This ranking employs a five-year window to evaluate publications; for instance, the 2000 ranking is based on publications from 1996 through 2000. This method is consistently applied to subsequent years up to 2023. Based on this ranking, we introduce two variables to assess the impact of these rankings on author affiliations. The variable top school graduation is assigned a value of 1 if an article's author graduated from a top-ranked school at the time of graduation, and 0 otherwise. Similarly, top school current is set to 1 if the author was affiliated with a top IS department at the time of the article's publication, and 0 otherwise.

Descriptive Statistics

There has been exponential growth in the use of design-based causal methods in the IS domain since 2000, with particularly rapid growth in the last seven years (Figure EC.1). The University of

[17] See `https://jsom.utdallas.edu/the-utd-top-100-business-school-research-rankings/`.

Table EC.5 Top 100 Schools/Institutions By Productivity

School	N	country
University of Minnesota at Twin Cities	49	USA
Temple University	45	USA
Carnegie Mellon University	35	USA
University of Maryland at College Park	33	USA
Arizona State University	28	USA
Indiana University at Bloomington	25	USA
New York University	25	USA
University of Houston	24	USA
City University of Hong Kong	23	China
University of Arizona	22	USA
University of Florida	22	USA
University of Texas at Dallas	22	USA
Georgia Institute of Technology	20	USA
University of Pennsylvania	20	USA
Georgia State University	19	USA
McGill University	19	Canada
National University of Singapore	19	Singapore
Hong Kong University of Science and Technology	18	China
University of British Columbia	17	Canada
University of Connecticut	17	USA
Boston University	16	USA
University of Texas at Austin	16	USA
University of Southern California	14	USA
Purdue University	13	USA
Tsinghua University	13	China
Brigham Young University	12	USA
University of Notre Dame	12	USA
University of Oklahoma	12	USA
University of South Florida	12	USA
Chinese University of Hong Kong	11	China
Fudan University	11	China
George Mason University	11	USA
HEC Paris	11	France
Korea Advanced Institute of Science and Technology	11	Korea
Emory University	10	USA
University of Georgia	10	USA
University of Michigan at Ann Arbor	10	USA
University of Virginia	10	USA
Virginia Tech	10	USA
Peking University	9	China
University of Washington at Seattle	9	USA
City University of New York	8	USA
Michigan State University	8	USA
University of Wisconsin at Madison	8	USA
Erasmus University Rotterdam	7	The Netherlands
Hong Kong Polytechnic University	7	China
Massachusetts Institute of Technology	7	USA
Nanjing University	7	China
Pennsylvania State University at State College	7	USA
Technische Universitat Darmstadt	7	Germany
University of Arkansas at Fayetteville	7	USA
University of Hong Kong	7	China
University of Illinois at Chicago	7	USA
University of Mannheim	7	Germany
University of Melbourne	7	Australia
University of Rochester	7	USA
University of Texas at Arlington	7	USA
Florida State University	6	USA
Goethe University of Frankfurt	6	Germany
Harbin Institute of Technology	6	China
Lehigh University	6	USA
Nanyang Technological University	6	Singapore
University of Illinois at Urbana Champaign	6	USA
University of Massachusetts at Amherst	6	USA
University of North Carolina at Charlotte	6	USA
University of North Carolina at Greensboro	6	USA
University of St. Gallen	6	Switzerland
Zhejiang University	6	China
Copenhagen Business School	5	Denmark
Harvard University	5	USA
IE University	5	Spain
Renmin University of China	5	China
Sichuan University	5	China
Singapore Management University	5	Singapore
Southern University of Science and Technology	5	China
Tel Aviv University	5	Israel
Texas A&M University at College Station	5	USA
Tulane University	5	USA
Universidade Catolica Portuguesa	5	Portugal
University of California at Irvine	5	USA
University of Cologne	5	Germany
University of Miami	5	USA
University of Pittsburgh	5	USA
University of Tennessee at Knoxville	5	USA
University of Wisconsin at Milwaukee	5	USA
Utah State University	5	USA
Baylor University	4	USA
Boston College	4	USA
Clemson University	4	USA
Collage.com	4	USA
ETH Zurich	4	Switzerland
Karlsruhe Institute of Technology	4	Germany
Miami University of Ohio	4	USA
Mississippi State University	4	USA
Northeastern University	4	USA
Santa Clara University	4	USA
Seoul National University	4	Korea
State University of New York at Albany	4	USA
Sun Yat Sen University	4	China
Texas Tech University	4	USA

Minnesota at Twin Cities leads in productivity, followed by Temple University, Carnegie Mellon University, and the University of Maryland (Table EC.5). On average, DID articles have 9 estimates, IV articles have 7, SC articles have 8, and RD articles have 10. Most papers in IS use DID and IV methods, consistent with trends in economics (Brodeur et al. 2020).

Figure EC.2 shows similar trends for each method over time. The RCT method generally dominated the other methods between 2000 and 2011. The use of DID has increased significantly, while the share of IV articles has decreased over the past two decades. In 2023, DID, IV, and RCT were the most popular methods for causal inference in IS research. The RCT and IV methods appeared earlier in IS literature, whereas the DID method was adopted later, often following Wooldridge (2010) guidelines. RD and SC methods are used less frequently, with RD being employed in the last seven years and SC in the last four years.

Figure EC.3 illustrates distinct trajectories in sample sizes across methods. DID studies initially relied on very large datasets, dipped in the early 2010s, and have since recovered, while IV applications remain stable at moderately large sample sizes. RCTs consistently use smaller samples, though their average size has gradually increased, likely due to the rise of online field experiments. RD exhibits substantial volatility, reflecting variation in access to administrative or policy data, and SC, as a relatively new method, shows inconsistent adoption with wide fluctuations in sample size. Together, these patterns highlight how methodological constraints and data availability shape the scale of empirical work in IS research.

A.3. Primer on Core Concepts

Estimate and standard error. Let $\hat{\theta}$ denote an estimate of a scalar quantity θ with estimated standard error $\widehat{\mathrm{se}}(\hat{\theta})$. The standard error summarizes sampling uncertainty. All statistics below are functions of $\hat{\theta}$ and $\widehat{\mathrm{se}}(\hat{\theta})$ only, independent of the specific model that produced them.

z-score and t-statistic. For a null value θ_0, the standardized statistic is

$$z \equiv \frac{\hat{\theta} - \theta_0}{\widehat{\mathrm{se}}(\hat{\theta})}.$$

Under regularity conditions, z is approximately standard normal. When a small sample or an additional finite sample structure is relevant, a t statistic is reported and compared to a Student t reference distribution with degrees of freedom ν. For large ν (i.e., a large sample), t and z coincide numerically.

p-value. For a two sided test of $H_0\colon \theta = \theta_0$ using a z statistic,

$$p = 2\{1 - \Phi(|z|)\},$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. For a one-sided test in the positive direction, $p = 1 - \Phi(z)$; for the negative direction, $p = \Phi(z)$. The mapping between $|z|$ and two-sided p is monotone.

Common two-sided thresholds.

Nominal level α	Critical $\lvert z\rvert$ for α
0.10	1.64
0.05	1.96
0.01	2.58

Confidence intervals. A $100(1-\alpha)\%$ confidence interval for θ is

$$\hat{\theta} \pm z_{1-\alpha/2}\,\widehat{\mathrm{se}}(\hat{\theta}),$$

or with t quantiles when a t reference is appropriate. Hypothesis tests and confidence intervals are equivalent summaries: $H_0\colon \theta = \theta_0$ is rejected at level α if and only if θ_0 lies outside the $100(1-\alpha)\%$ interval.

Effect size and standardization. The raw effect size is $\hat{\theta}$. A standardized effect is divided by a reference scale, for example $\hat{\theta}/\widehat{\mathrm{sd}}(Y)$ when available. Our distributional diagnostics do not require standardization beyond the z-score.

Type I and Type II errors, power, and minimum detectable effect. Type I error equals α. Power is $\Pr(\text{reject } H_0 \mid \theta = \theta_1)$ for an alternative θ_1. For a two sided z test with standard error $\widehat{\mathrm{se}}$, the minimum detectable effect at level α and power $1-\beta$ satisfies

$$\mathrm{MDE} \approx \left(z_{1-\alpha/2} + z_{1-\beta}\right)\widehat{\mathrm{se}}.$$

When $\widehat{\mathrm{se}} \propto n^{-1/2}$, the MDE scales as $n^{-1/2}$.

Scaling with sample size and a simple fragility metric. Holding design and variance fixed, $|z|$ scales approximately with $\sqrt{n}$. For a significant test with $|z| > 1.96$ at the 5 percent level, the smallest fraction of the current sample that would still retain $|z| \geq 1.96$ is

$$f_{\min} = \left(\frac{1.96}{|z|}\right)^2,$$

so a reduction of $100 \times (1 - f_{\min})$ percent would lose significance. For a non-significant test, the multiplicative increase in sample size needed to reach $|z| = 1.96$ is

$$m_{\min} = \left(\frac{1.96}{|z|} \right)^2 .$$

Multiple testing and error rates. When many hypotheses are reported, family-wise error rate controls the probability of at least one false rejection, while false discovery rate controls the expected share of false rejections among rejections. Adjusted p values or q values are common reporting tools.

Local false discovery rate (local FDR). Local FDR at a statistic value z is the posterior probability that the null is true given the observed z:

$$\mathrm{LFDR}(z) = \Pr(H_0 \mid z) = \frac{\pi_0 f_0(z)}{f(z)},$$

where f_0 is the null density, f is the marginal density of z, and π_0 is the prior proportion of nulls. In practice, f and often π_0 are estimated from the empirical distribution of z statistics. Local FDR is a pointwise measure of evidential reliability that complements p values.

Threshold adjacent diagnostics. To assess irregular mass near conventional cutoffs, define symmetric windows around $|z| = 1.96$:

$$\mathcal{W}_- = [1.46, 1.96), \qquad \mathcal{W}_+ = [1.96, 2.46].$$

A simple excess mass measure is the caliper excess ratio

$$\mathrm{CER} = \frac{\#\{i : |z_i| \in \mathcal{W}_+\}}{\#\{i : |z_i| \in \mathcal{W}_-\}},$$

with exact window choices varied in sensitivity checks. Values above one indicate more mass just above than just below the threshold.

Weighting and dependence aware uncertainty. When multiple statistics come from the same article, unit article weights prevent a single article from dominating pooled summaries. Dependence across statistics within an article can inflate precision if ignored; cluster robust or article-level bootstrap procedures address this by resampling at the article level.

Appendix B. Related Literature

Table EC.6 Related Papers on p-hacking and Publication Bias in Different Fields

Paper	Review	Survey	Simulation	Discussion	Field
De Long and Lang (1992)	✓				Economics
Stanley (2005)	✓				Economics
Brodeur et al. (2020)	✓				Economics
Brodeur et al. (2022)				✓	Economics
Brodeur et al. (2023)	✓				Economics
Hollenbeck and Wright (2017)				✓	Management
Harrison et al. (2017)	✓				Management
Fišar et al. (2024)	✓				Management
Friese and Frankenbach (2020)			✓		Psychology
Siegel et al. (2022)	✓				Psychology
Easterbrook et al. (1991)	✓				Medicine
Thornton and Lee (2000)			✓		Medicine
Tumber and Dickersin (2004)				✓	Medicine
Song et al. (2013)				✓	Medicine
Andrade (2021)				✓	Medicine
Weiß and Wagner (2011)				✓	Social sciences
Franco et al. (2014)	✓				Social sciences
Head et al. (2015)	✓				Metascience
Stefan and Schönbrodt (2023)			✓		Metascience
Fraser et al. (2018)	✓				Ecology
Elston (2021)		✓			Dermatology
Friedman and Wyatt (2001)				✓	Health informatics
Eysenbach (2004)				✓	Health informatics
Smith (2020)	✓				Health services
He and King (2008)	✓				IS
Kepes and Thomas (2018)				✓	IS
Weinhardt et al. (2019)				✓	IS
Jeyaraj and Dwivedi (2020)				✓	IS
Hu et al. (2023)		✓			IS
This study	✓				IS

Notes: IS stands for Information Systems.

Appendix C. Statistical Tests

C.1. Z-curve

The Z-curve 2.0 method is a statistical technique used to estimate the expected replication rate (ERR) and expected discovery rate (EDR) of published research findings. This method was developed by Bartoš and Schimmack (2022) as an extension of the original Z-curve method Brunner and Schimmack (2020). The primary aim of Z-curve 2.0 is to correct for selection bias that arises when only statistically significant results are published, thereby inflating the apparent effect sizes and discovery rates. The method achieves this by modeling the distribution of observed z-scores from published studies as a mixture of truncated normal distributions.

Expected Replication Rate (ERR) The ERR is the estimated probability that a significant result can be replicated in an exact replication study. It is calculated as the mean power of the statistically significant studies after adjusting for selection bias. In mathematical terms, the ERR is defined as:

$$\text{ERR} = \frac{\sum_{k=1}^{K} \epsilon_{2,k} \cdot \epsilon_{1,k}}{\sum_{k=1}^{K} \epsilon_{2,k}}, \tag{1}$$

where $\epsilon_{2,k}$ represents the power of the k-th study regardless of the outcome direction, and $\epsilon_{1,k}$ represents the power of the k-th study in the specified direction.

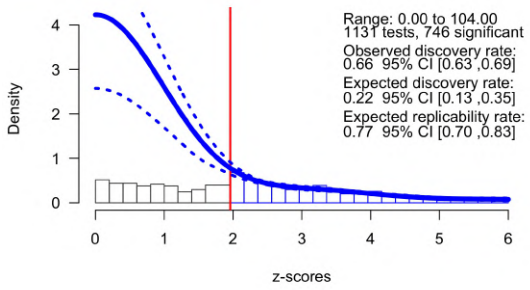
Expected Discovery Rate (EDR) The EDR estimates the proportion of statistically significant results among all conducted studies, including those that were not published. It provides an estimate of the true rate of discoveries in the research field and is calculated as:

$$\text{EDR} = \frac{\sum_{k=1}^{K} \epsilon_{2,k}}{K}, \tag{2}$$

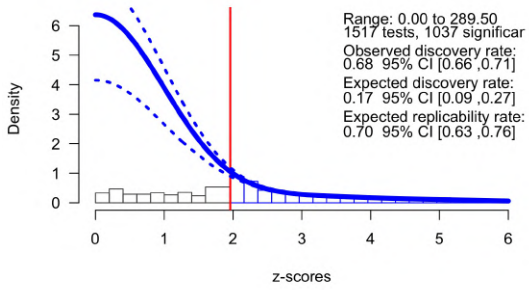
where K is the total number of studies considered.

For further details on the methodology and its implementation, refer to Bartoš and Schimmack (2022). Figure EC.4 presents Z-curves by method.

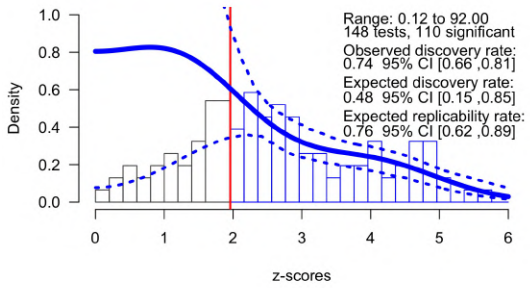
Figure EC.4 Combined Z-curve Plots



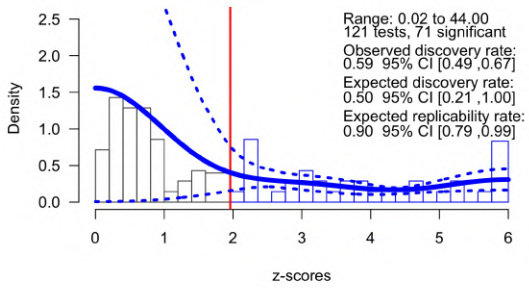
(a) Z-curve Plot DID



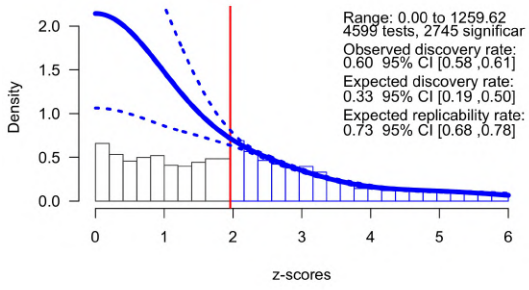
(b) Z-curve Plot IV



(c) Z-curve Plot RD



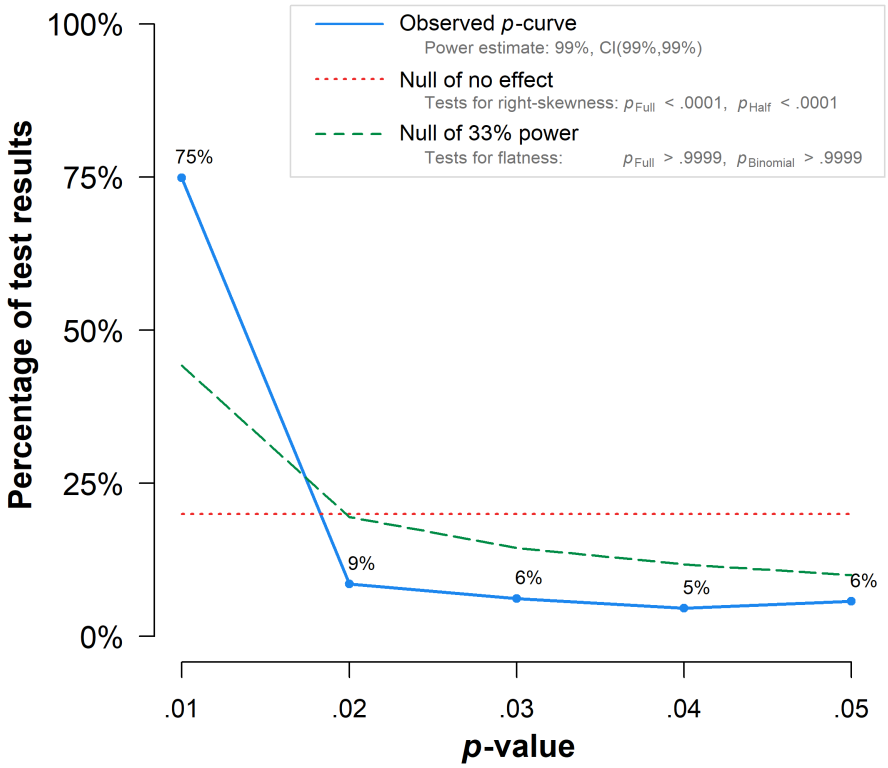
(d) Z-curve Plot SC



(e) Z-curve Plot RCT

The p-curve analysis indicates that the set of studies included in the analysis is likely to contain true effects rather than being driven by publication bias or p-hacking (Figure EC.5). The strong right-skewness and high estimated power suggest that the evidence is robust and that the findings are not due to random chance. The presence of many very low p-values further supports the notion that the results are reflecting some true effects.

Figure EC.5 p-curve analysis



Note: The observed *p*-curve includes 4709 statistically significant ($p < .05$) results, of which 4076 are $p < .025$. There were 2807 additional results entered but excluded from *p*-curve because they were $p > .05$.

C.2. Details on Caliper Test

The caliper test compares the number of estimates just above and below a statistical significance threshold, allowing us to control for author and article characteristics. This test is based on observed z-scores, where the sampling distribution $F(z)$ is continuous over an interval $[c, c'']$.

Under standard regularity conditions, the asymptotic sampling distribution of z is normal (Gerber and Malhotra 2008b,a). For any c, the probability that a draw from $F(z)$ is between $[c', c]$, where $c' > c$, is $F(c') - F(c)$. The conditional probability that a draw from F is in $[c, c']$, given it is drawn from $[c, c'']$ where $c'' > c' > c$, is $\frac{[F(c')-F(c)]}{[F(c'')-F(c)]}$.

Since F is continuous, $F(x+e) - F(x) = ef(x) + E$, where $f(x)$ is the probability density function and E is an approximation error. For small intervals or when $f'(x) = 0$, this ratio approximates to $\frac{[c'-c]f(c)}{[c''-c]f(c)}$. Canceling $f(c)$ terms, the conditional probability of observing an outcome in a subset of an interval equals the relative proportion of the subset to the interval. If $c' - c = c'' - c'$, then $[c', c'']$ is half the interval, and $p = 0.5$. For N independent draws, the number of occurrences in a subset of the interval is binomially distributed (N, p).

To test for publication bias, we set $c' = 1.96$, the critical value for the 5% significance level, and examine outcomes near 1.96. Under the null hypothesis that results are random draws, the number of occurrences above 1.96 is binomially distributed $(N, 0.5)$.

C.3. Caliper Tests

C.3.1. Alternative Thresholds

Table EC.7 Caliper Test (Logit) at the 10 Percent Significance Level

Dependent Variable:	Significant			
Model:	(1)	(2)	(3)	(4)
Variables				
DID	0.14	0.13	0.13	-0.10
	(0.74)	(0.68)	(0.68)	(0.67)
IV	0.18	0.25	0.25	0.32
	(0.71)	(0.64)	(0.64)	(0.62)
RCT	-0.66	-0.65	-0.65	-0.81
	(0.72)	(0.65)	(0.65)	(0.62)
SC	0.02	0.02	0.03	-0.63
	(1.1)	(0.97)	(0.97)	(0.78)
Log(Obs)	-0.09*	-0.10**	-0.10**	-0.07
	(0.04)	(0.04)	(0.04)	(0.05)
Experience		-0.06**	-0.06**	-0.07*
		(0.02)	(0.02)	(0.03)
Top Current School		0.42	0.42	0.72
		(0.47)	(0.47)	(0.55)
Top PhD School		0.18	0.18	0.25
		(0.44)	(0.44)	(0.49)
Solo		-0.13	-0.12	-0.60
		(0.59)	(0.60)	(0.63)
Female			-0.08	0.01
			(0.36)	(0.39)
Fixed-effects				
Editor	Yes	Yes	Yes	Yes
Journal	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes
Fit statistics				
Observations	1,364	1,352	1,352	897
Pseudo R^2	0.99	0.99	0.99	1.0
BIC	282.5	310.8	318.0	299.6
Window	1.64 ± 0.5	1.64 ± 0.5	1.64 ± 0.5	1.64 ± 0.35
RD Sig Rate	0.63	0.63	0.63	0.62

Clustered (Article) standard-errors in parentheses
*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05, .: 0.1*

Notes: This table reports log odds format from the logit regression. The outcome is 1 for significance at the 10% level. Experience is the average number of years since the authors started their PhDs to publication. Female is the percentage of female authors. Top current school is the percentage of authors at top universities. Top PhD school is the percentage of authors from top PhD programs. Editor indicates if the team has editorial experience. We use the RD method as the baseline since it does not show evidence of p-hacking or publication bias. Robust standard errors, clustered by article, are in parentheses. Observations are weighted by the inverse number of tests in each article.

Table EC.8 Caliper Test (Logit) at the 1 Percent Significance Level

Dependent Variable:	Significant			
Model:	(1)	(2)	(3)	(4)
Variables				
DID	-0.13	-0.25	-0.27	0.07
	(0.66)	(0.60)	(0.60)	(0.70)
IV	-0.36	-0.47	-0.49	-0.29
	(0.62)	(0.56)	(0.57)	(0.66)
RCT	-0.003	-0.06	-0.10	0.11
	(0.63)	(0.57)	(0.58)	(0.69)
SC	-3.0*	-3.2*	-3.2*	-1.8
	(1.4)	(1.4)	(1.4)	(1.6)
Log(Obs)	0.06	0.06	0.06	-0.02
	(0.04)	(0.05)	(0.05)	(0.05)
Experience		0.01	0.02	-0.008
		(0.02)	(0.02)	(0.02)
Top Current School		-0.45	-0.45	-0.41
		(0.48)	(0.48)	(0.52)
Top PhD School		0.20	0.20	0.39
		(0.51)	(0.52)	(0.50)
Solo		-0.43	-0.42	-0.75
		(0.70)	(0.70)	(0.70)
Female			0.16	0.65
			(0.44)	(0.48)
Fixed-effects				
Editor	Yes	Yes	Yes	Yes
Journal	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes
Fit statistics				
Observations	1,291	1,272	1,272	879
Pseudo R^2	0.96	0.96	0.95	0.98
BIC	308.3	336.4	343.6	306.8
Window	2.58 ± 0.5	2.58 ± 0.5	2.58 ± 0.5	2.58 ± 0.35
RD Sig Rate	0.49	0.49	0.49	0.48

Clustered (Article) standard-errors in parentheses
*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05, .: 0.1*

Notes: This table reports log odds format from the logit regression. The outcome is 1 for significance at the 1% level. Experience is the average number of years since the authors started their PhDs to publication. Female is the percentage of female authors. Top current school is the percentage of authors at top universities. Top PhD school is the percentage of authors from top PhD programs. Editor indicates if the team has editorial experience. We use the RD method as the baseline since it does not show evidence of p-hacking or publication bias. Robust standard errors, clustered by article, are in parentheses. Observations are weighted by the inverse number of tests in each article.

C.3.2. Logit Regression

Table EC.9 Caliper Test (Logit)

Dependent Variable:	Significant			
Model:	(1)	(2)	(3)	(4)
Variables				
DID	1.20*	1.53**	1.62**	1.42**
	(0.598)	(0.533)	(0.516)	(0.510)
IV	0.798	1.11*	1.21*	1.15*
	(0.581)	(0.517)	(0.504)	(0.500)
RCT	0.458	0.762	0.906·	1.01*
	(0.587)	(0.520)	(0.504)	(0.506)
SC	2.24*	2.66*	2.75**	3.26**
	(1.03)	(1.04)	(1.05)	(1.03)
Log(Obs)	-0.038	-0.031	-0.023	-0.015
	(0.038)	(0.038)	(0.037)	(0.044)
Experience		-0.0009	-0.003	-0.013
		(0.020)	(0.020)	(0.023)
Top Current School		0.432	0.414	0.194
		(0.391)	(0.387)	(0.469)
Top PhD School		-0.443	-0.416	-0.239
		(0.389)	(0.383)	(0.498)
Solo		1.26·	1.40*	1.28·
		(0.695)	(0.694)	(0.745)
Female			-0.982**	-0.897*
			(0.319)	(0.375)
Fixed-effects				
Editor	Yes	Yes	Yes	Yes
Journal	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes
Fit statistics				
Observations	1,713	1,709	1,709	1,281
Pseudo R^2	0.97	0.97	0.97	0.96
BIC	327.0	356.5	363.8	342.1
Window	1.96 ± 0.5	1.96 ± 0.5	1.96 ± 0.5	1.96 ± 0.35
RD Sig Rate	0.490	0.490	0.490	0.500

Clustered (Article) standard-errors in parentheses
*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05, .: 0.1*

Notes: This table reports log odds format from the logit regression. The outcome is 1 for significance at the 5% level. Experience is the average number of years since the authors started their PhDs to publication. Female is the percentage of female authors. Top current school is the percentage of authors at top 100 universities. Top PhD school is the percentage of authors from top 100 PhD programs. Editor indicates if the team has editorial experience. We use the RD method as the baseline since it does not show evidence of p-hacking or publication bias. Robust standard errors, clustered by article, are in parentheses. Observations are weighted by the inverse number of tests in each article.

Table EC.10 Caliper Test (Logit) - Top 20

Dependent Variable:	Significant			
Model:	(1)	(2)	(3)	(4)
Variables				
DID	1.20*	1.49**	1.57**	1.38**
	(0.598)	(0.534)	(0.521)	(0.536)
IV	0.798	1.09*	1.17*	1.11*
	(0.581)	(0.521)	(0.511)	(0.525)
RCT	0.458	0.768	0.894·	0.989·
	(0.587)	(0.519)	(0.508)	(0.529)
SC	2.24*	2.53*	2.62*	3.19**
	(1.03)	(1.01)	(1.03)	(1.03)
Log(Obs)	-0.038	-0.029	-0.022	-0.019
	(0.038)	(0.038)	(0.038)	(0.044)
Experience		0.003	-0.0005	-0.012
		(0.020)	(0.020)	(0.023)
Top Current School (20)		0.070	0.153	0.425
		(0.295)	(0.301)	(0.351)
Top PhD School (20)		-0.048	-0.133	-0.122
		(0.295)	(0.295)	(0.346)
Solo		1.29·	1.45*	1.33·
		(0.702)	(0.704)	(0.756)
Female			-1.01**	-0.940*
			(0.321)	(0.378)
Fixed-effects				
Editor	Yes	Yes	Yes	Yes
Journal	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes
Fit statistics				
Observations	1,713	1,709	1,709	1,281
Pseudo R^2	0.97	0.97	0.97	0.96
BIC	327.0	356.4	363.7	342.0
Window	1.96 ± 0.5	1.96 ± 0.5	1.96 ± 0.5	1.96 ± 0.35
RD Sig Rate	0.490	0.490	0.490	0.500

Clustered (Article) standard-errors in parentheses
*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05, .: 0.1*

Notes: This table reports log odds format from the logit regression. The outcome is 1 for significance at the 5% level. Experience is the average number of years since the authors started their PhDs to publication. Female is the percentage of female authors. Top current school is the percentage of authors at top 20 universities. Top PhD school is the percentage of authors from top 20 PhD programs. Editor indicates if the team has editorial experience. We use the RD method as the baseline since it does not show evidence of p-hacking or publication bias. Robust standard errors, clustered by article, are in parentheses. Observations are weighted by the inverse number of tests in each article.

C.3.3. Probit Regression

We run the following regression for the probit version

$$P(\text{Significant}_{ijt} = 1) = \Phi(\beta_j + \sigma_t + \text{Editor}_{ij} + X_{ij}\theta + \gamma_1 \text{DID}_{ij} + \gamma_2 \text{IV}_{ij} + \gamma_3 \text{RCT}_{ij} + \gamma_4 \text{SC}_{ij}). \quad (3)$$

Table EC.11 Caliper Test (Probit)

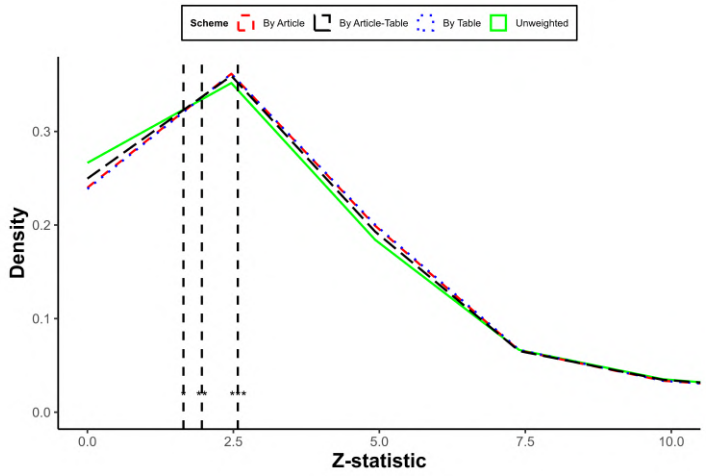
Dependent Variable:	Significant			
Model:	(1)	(2)	(3)	(4)
Variables				
DID	0.73*	0.92**	0.97**	0.86**
	(0.37)	(0.33)	(0.32)	(0.31)
IV	0.50	0.68*	0.73*	0.70*
	(0.36)	(0.32)	(0.31)	(0.31)
RCT	0.29	0.46	0.54·	0.60·
	(0.36)	(0.32)	(0.31)	(0.31)
SC	1.3*	1.5**	1.6**	1.8***
	(0.57)	(0.56)	(0.57)	(0.55)
Log(Obs)	-0.02	-0.02	-0.01	-0.01
	(0.02)	(0.02)	(0.02)	(0.03)
Experience		-0.0007	-0.002	-0.008
		(0.01)	(0.01)	(0.01)
Top Current School		0.26	0.25	0.12
		(0.24)	(0.23)	(0.29)
Top PhD School		-0.26	-0.24	-0.14
		(0.24)	(0.23)	(0.30)
Solo		0.72·	0.78·	0.72
		(0.40)	(0.41)	(0.45)
Female			-0.60**	-0.55*
			(0.20)	(0.23)
Fixed-effects				
Editor	Yes	Yes	Yes	Yes
Journal	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes
Fit statistics				
Observations	1,713	1,709	1,709	1,281
Pseudo R^2	0.27	0.29	0.30	0.25
BIC	327.0	356.5	363.8	342.1
Window	1.96 ± 0.5	1.96 ± 0.5	1.96 ± 0.5	1.96 ± 0.35
RD Sig Rate	0.49	0.49	0.49	0.50

Clustered (Article) standard-errors in parentheses
*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05, .: 0.1*

Notes: This table reports the latent variable format from the probit regression. The outcome is 1 for significance at the 5% level. Experience is the average number of years since the authors started their PhDs to publication. Female is the percentage of female authors. Top current school is the percentage of authors at top universities. Top PhD school is the percentage of authors from top PhD programs. Editor indicates if the team has editorial experience. We use the RD method as the baseline since it does not show evidence of p-hacking or publication bias. Robust standard errors, clustered by article, are in parentheses. Observations are weighted by the inverse number of tests in each article.

C.4. Test Statistics Plot

Figure EC.6 z**-statistics by Weight**



Notes: This figure shows histograms of test statistics for $z \in [0, 10]$, with distributions unweighted and weighted by article, table, or both, using the inverse of tests per table and tables per article.

Figure EC.7 z**-statistics Plots by Method**

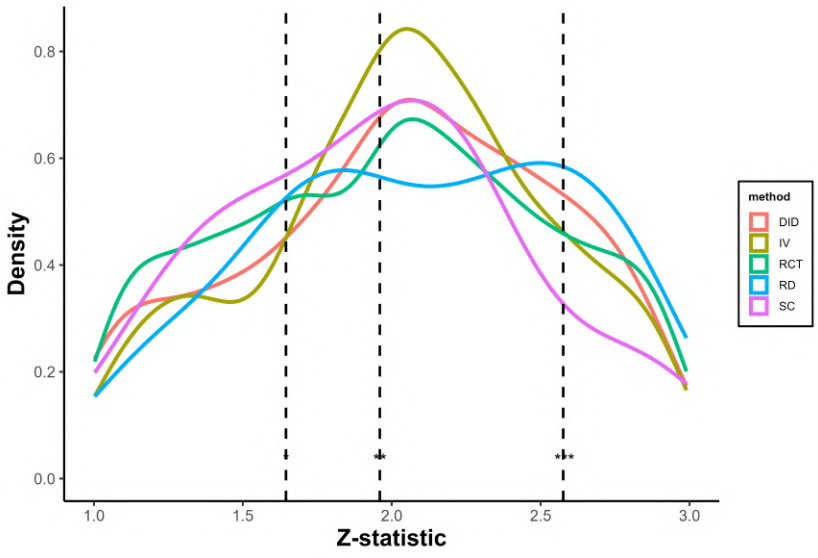
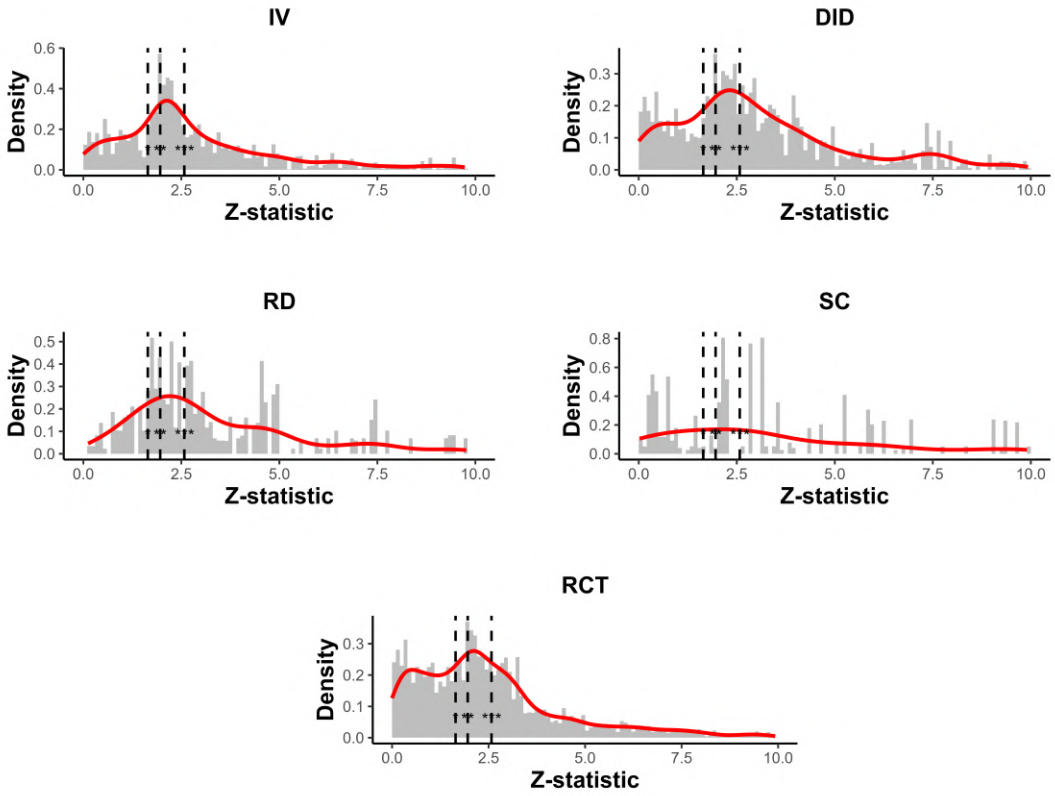


Figure EC.8 z**-statistics Plots by Method (Weighted)**



Notes. This figure presents histograms of test statistics for $z \in [0, 10]$, weighted by the inverse of tests per table and total tables. It categorizes statistics by method (DID, IV, RCT, RD, and SC) and marks conventional two-tailed significance levels.

C.5. Randomization Tests

Table EC.12 Randomization Tests (5 Percent Significance Threshold) - Weighted version

	IV	RCT	DID	RD	SC
Prop sig in 1.96 ± 0.5	0.654	0.585	0.714	0.474	0.907
One-sided p-value	0.0104	0.0808	0.0211	0.5412	0.1461
N of tests in 1.96 ± 0.5	396	1050	239	35	14
Prop sig in 1.96 ± 0.4	0.646	0.596	0.694	0.458	0.936
One-sided p-value	0.0195	0.0774	0.0433	0.5648	0.1516
N of tests in 1.96 ± 0.4	353	855	200	30	12
Prop sig in 1.96 ± 0.3	0.616	0.594	0.651	0.439	0.935
One-sided p-value	0.0675	0.1010	0.1176	0.5861	0.1704
N of tests in 1.96 ± 0.3	287	679	159	24	10
Prop sig in 1.96 ± 0.2	0.648	0.633	0.664	0.433	0.787
One-sided p-value	0.0610	0.0769	0.1542	0.5704	0.3350
N of tests in 1.96 ± 0.2	197	442	106	13	5
Prop sig in 1.96 ± 0.1	0.718	0.675	0.773	0.678	0.000
One-sided p-value	0.0552	0.0855	0.1156	0.3672	0.5002
N of tests in 1.96 ± 0.1	101	239	63	8	2
Prop sig in 1.96 ± 0.075	0.707	0.680	0.780	0.781	0.000
One-sided p-value	0.0816	0.1124	0.1442	0.3165	0.5002
N of tests in 1.96 ± 0.075	84	187	49	7	1
Prop sig in 1.96 ± 0.05	0.735	0.659	0.818	0.707	0.000
One-sided p-value	0.0930	0.1723	0.1332	0.3766	0.5002
N of tests in 1.96 ± 0.05	65	140	41	5	1

Table EC.13 Randomization Tests (5 Percent Significance Threshold - Random Sample)

	IV	RCT	DID	RD	SC
Prop sig in 1.96 ± 0.5	0.654	0.553	0.690	0.500	0.500
One-sided p-value	0.0002	0.0667	0.0009	0.6230	0.6562
N of tests in 1.96 ± 0.5	136	215	71	10	6
Prop sig in 1.96 ± 0.4	0.595	0.589	0.588	0.500	0.500
One-sided p-value	0.0178	0.0063	0.0909	0.6230	0.6562
N of tests in 1.96 ± 0.4	131	209	68	10	6
Prop sig in 1.96 ± 0.3	0.579	0.609	0.627	0.300	0.667
One-sided p-value	0.0555	0.0012	0.0337	0.9453	0.3437
N of tests in 1.96 ± 0.3	114	202	59	10	6
Prop sig in 1.96 ± 0.2	0.611	0.599	0.723	0.286	0.250
One-sided p-value	0.0198	0.0058	0.0015	0.9375	0.9375
N of tests in 1.96 ± 0.2	95	172	47	7	4
Prop sig in 1.96 ± 0.1	0.652	0.652	0.800	0.600	0.000
One-sided p-value	0.0093	0.0003	0.0003	0.5000	1.0000
N of tests in 1.96 ± 0.1	66	132	35	5	2
Prop sig in 1.96 ± 0.075	0.707	0.699	0.786	0.600	0.000
One-sided p-value	0.0011	0.0000	0.0019	0.5000	1.0000
N of tests in 1.96 ± 0.075	58	113	28	5	1
Prop sig in 1.96 ± 0.05	0.717	0.644	0.708	0.750	0.000
One-sided p-value	0.0023	0.0040	0.0320	0.3125	1.0000
N of tests in 1.96 ± 0.05	46	90	24	4	1

Table EC.14 Randomization Tests (5 Percent Significance Threshold)

	Without	With
Prop sig in 1.96 ± 0.5	0.62	0.64
One-sided p-value	0.00	0.00
N of tests in 1.96 ± 0.5	122	107
Prop sig in 1.96 ± 0.4	0.63	0.62
One-sided p-value	0.01	0.02
N of tests in 1.96 ± 0.4	102	89
Prop sig in 1.96 ± 0.3	0.57	0.63
One-sided p-value	0.11	0.02
N of tests in 1.96 ± 0.3	82	70
Prop sig in 1.96 ± 0.2	0.64	0.65
One-sided p-value	0.02	0.03
N of tests in 1.96 ± 0.2	56	46
Prop sig in 1.96 ± 0.1	0.71	0.70
One-sided p-value	0.01	0.03
N of tests in 1.96 ± 0.1	34	27
Prop sig in 1.96 ± 0.075	0.72	0.72
One-sided p-value	0.01	0.05
N of tests in 1.96 ± 0.075	29	18
Prop sig in 1.96 ± 0.05	0.78	0.61
One-sided p-value	0.00	0.29
N of tests in 1.96 ± 0.05	27	13

C.5.1. Alternative Thresholds

Table EC.15 Randomization Tests (1 Percent Significance Threshold)

	IV	RCT	DID	RD	SC
Prop sig in 2.58 ± 0.5	0.363	0.417	0.417	0.486	0.333
One-sided p-value	1.0000	1.0000	0.9927	0.6321	0.9102
N of tests in 2.58 ± 0.5	325	906	206	35	9
Prop sig in 2.58 ± 0.4	0.377	0.417	0.432	0.484	0.333
One-sided p-value	1.0000	1.0000	0.9677	0.6399	0.9102
N of tests in 2.58 ± 0.4	252	701	169	31	9
Prop sig in 2.58 ± 0.3	0.407	0.421	0.462	0.476	0.667
One-sided p-value	0.9954	0.9998	0.8327	0.6682	0.5000
N of tests in 2.58 ± 0.3	182	506	130	21	3
Prop sig in 2.58 ± 0.2	0.459	0.432	0.478	0.500	1.000
One-sided p-value	0.8287	0.9949	0.7008	0.5982	0.5000
N of tests in 2.58 ± 0.2	111	336	90	16	1
Prop sig in 2.58 ± 0.1	0.476	0.406	0.565	0.500	
One-sided p-value	0.6927	0.9930	0.2307	0.6230	
N of tests in 2.58 ± 0.1	63	160	46	10	
Prop sig in 2.58 ± 0.075	0.500	0.353	0.667	0.833	
One-sided p-value	0.5612	0.9995	0.0494	0.1094	
N of tests in 2.58 ± 0.075	42	116	30	6	
Prop sig in 2.58 ± 0.05	0.406	0.333	0.688	0.500	
One-sided p-value	0.8923	0.9995	0.1051	0.7500	
N of tests in 2.58 ± 0.05	32	90	16	2	

Table EC.16 Randomization Tests (10 Percent Significance Threshold)

	IV	RCT	DID	RD	SC
Prop sig in 1.64 ± 0.5	0.701	0.584	0.660	0.633	0.545
One-sided p-value	0.0000	0.0000	0.0000	0.1002	0.5000
N of tests in 1.64 ± 0.5	338	960	200	30	11
Prop sig in 1.64 ± 0.4	0.686	0.570	0.636	0.667	0.556
One-sided p-value	0.0000	0.0001	0.0003	0.0610	0.5000
N of tests in 1.64 ± 0.4	264	765	165	27	9
Prop sig in 1.64 ± 0.3	0.663	0.539	0.589	0.650	0.571
One-sided p-value	0.0000	0.0382	0.0407	0.1316	0.5000
N of tests in 1.64 ± 0.3	169	536	107	20	7
Prop sig in 1.64 ± 0.2	0.675	0.565	0.603	0.667	0.600
One-sided p-value	0.0001	0.0065	0.0503	0.1509	0.5000
N of tests in 1.64 ± 0.2	114	375	73	15	5
Prop sig in 1.64 ± 0.1	0.623	0.600	0.649	0.625	0.500
One-sided p-value	0.0492	0.0028	0.0494	0.3633	0.7500
N of tests in 1.64 ± 0.1	53	200	37	8	2
Prop sig in 1.64 ± 0.075	0.600	0.653	0.586	0.667	1.000
One-sided p-value	0.1553	0.0001	0.2291	0.3437	0.5000
N of tests in 1.64 ± 0.075	35	147	29	6	1
Prop sig in 1.64 ± 0.05	0.500	0.653	0.500	0.600	1.000
One-sided p-value	0.5747	0.0016	0.5927	0.5000	0.5000
N of tests in 1.64 ± 0.05	28	98	18	5	1

C.5.2. Weighted Estimates

Table EC.17 Randomization Tests (1 Percent Significance Threshold)

	IV	RCT	DID	RD	SC
Prop sig in 2.58 ± 0.5	0.356	0.431	0.435	0.441	0.367
One-sided p-value	0.9770	0.8633	0.7458	0.5849	0.6709
N of tests in 2.58 ± 0.5	325	906	206	35	9
Prop sig in 2.58 ± 0.4	0.381	0.448	0.424	0.383	0.367
One-sided p-value	0.9272	0.7720	0.7548	0.6452	0.6709
N of tests in 2.58 ± 0.4	252	701	169	31	9
Prop sig in 2.58 ± 0.3	0.399	0.441	0.419	0.424	0.754
One-sided p-value	0.8516	0.7572	0.7286	0.5757	0.2871
N of tests in 2.58 ± 0.3	182	506	130	21	3
Prop sig in 2.58 ± 0.2	0.435	0.432	0.469	0.444	1.000
One-sided p-value	0.6965	0.7480	0.5743	0.5529	0.4998
N of tests in 2.58 ± 0.2	111	336	90	16	1
Prop sig in 2.58 ± 0.1	0.472	0.402	0.607	0.442	
One-sided p-value	0.5633	0.7479	0.3274	0.5450	
N of tests in 2.58 ± 0.1	63	160	46	10	
Prop sig in 2.58 ± 0.075	0.445	0.318	0.690	0.920	
One-sided p-value	0.6000	0.8519	0.2495	0.3263	
N of tests in 2.58 ± 0.075	42	116	30	6	
Prop sig in 2.58 ± 0.05	0.334	0.294	0.743	0.818	
One-sided p-value	0.7387	0.8492	0.2663	0.3967	
N of tests in 2.58 ± 0.05	32	90	16	2	

Table EC.18 Randomization Tests (10 Percent Significance Threshold)

	IV	RCT	DID	RD	SC
Prop sig in 1.64 ± 0.5	0.747	0.631	0.733	0.753	0.793
One-sided p-value	0.0007	0.0224	0.0344	0.1880	0.3100
N of tests in 1.64 ± 0.5	338	960	200	30	11
Prop sig in 1.64 ± 0.4	0.729	0.601	0.699	0.772	0.538
One-sided p-value	0.0042	0.0836	0.0801	0.1821	0.4837
N of tests in 1.64 ± 0.4	264	765	165	27	9
Prop sig in 1.64 ± 0.3	0.711	0.571	0.695	0.764	0.474
One-sided p-value	0.0284	0.2069	0.1410	0.2230	0.5095
N of tests in 1.64 ± 0.3	169	536	107	20	7
Prop sig in 1.64 ± 0.2	0.751	0.604	0.753	0.800	0.508
One-sided p-value	0.0304	0.1602	0.1259	0.2257	0.4976
N of tests in 1.64 ± 0.2	114	375	73	15	5
Prop sig in 1.64 ± 0.1	0.670	0.677	0.758	0.690	0.286
One-sided p-value	0.1832	0.1160	0.2094	0.3592	0.5460
N of tests in 1.64 ± 0.1	53	200	37	8	2
Prop sig in 1.64 ± 0.075	0.648	0.738	0.705	0.612	1.000
One-sided p-value	0.2590	0.0864	0.2806	0.4286	0.4998
N of tests in 1.64 ± 0.075	35	147	29	6	1
Prop sig in 1.64 ± 0.05	0.559	0.679	0.651	0.521	1.000
One-sided p-value	0.4074	0.2145	0.3594	0.4879	0.4998
N of tests in 1.64 ± 0.05	28	98	18	5	1

C.5.3. Random Sample

Table EC.19 Randomization Tests (1 Percent Significance Threshold - Random Sample)

	IV	RCT	DID	RD	SC
Prop sig in 2.58 ± 0.5	0.412	0.439	0.395	0.600	0.400
One-sided p-value	0.9784	0.9697	0.9748	0.3770	0.8125
N of tests in 2.58 ± 0.5	119	223	76	10	5
Prop sig in 2.58 ± 0.4	0.440	0.471	0.437	0.500	0.400
One-sided p-value	0.9102	0.8151	0.8825	0.6367	0.8125
N of tests in 2.58 ± 0.4	109	210	71	8	5
Prop sig in 2.58 ± 0.3	0.456	0.403	0.484	0.375	0.667
One-sided p-value	0.8286	0.9971	0.6460	0.8555	0.5000
N of tests in 2.58 ± 0.3	90	191	64	8	3
Prop sig in 2.58 ± 0.2	0.500	0.417	0.463	0.500	1.000
One-sided p-value	0.5482	0.9860	0.7517	0.6367	0.5000
N of tests in 2.58 ± 0.2	68	163	54	8	1
Prop sig in 2.58 ± 0.1	0.490	0.472	0.567	0.571	
One-sided p-value	0.6101	0.7516	0.2923	0.5000	
N of tests in 2.58 ± 0.1	51	106	30	7	
Prop sig in 2.58 ± 0.075	0.441	0.397	0.609	1.000	
One-sided p-value	0.8042	0.9732	0.2024	0.0312	
N of tests in 2.58 ± 0.075	34	78	23	5	
Prop sig in 2.58 ± 0.05	0.360	0.359	0.643	0.500	
One-sided p-value	0.9461	0.9916	0.2120	0.7500	
N of tests in 2.58 ± 0.05	25	64	14	2	

Table EC.20 Randomization Tests (10 Percent Significance Threshold - Random Sample)

	IV	RCT	DID	RD	SC
Prop sig in 1.64 ± 0.5	0.697	0.596	0.746	0.700	0.500
One-sided p-value	0.0000	0.0030	0.0001	0.1719	0.6875
N of tests in 1.64 ± 0.5	122	213	59	10	4
Prop sig in 1.64 ± 0.4	0.724	0.558	0.618	0.800	0.667
One-sided p-value	0.0000	0.0593	0.0524	0.0547	0.5000
N of tests in 1.64 ± 0.4	105	199	55	10	3
Prop sig in 1.64 ± 0.3	0.704	0.517	0.619	0.750	0.333
One-sided p-value	0.0002	0.3524	0.0821	0.1445	0.8750
N of tests in 1.64 ± 0.3	81	174	42	8	3
Prop sig in 1.64 ± 0.2	0.727	0.563	0.622	0.571	0.667
One-sided p-value	0.0001	0.0714	0.0939	0.5000	0.5000
N of tests in 1.64 ± 0.2	66	151	37	7	3
Prop sig in 1.64 ± 0.1	0.667	0.644	0.696	0.714	0.500
One-sided p-value	0.0266	0.0021	0.0466	0.2266	0.7500
N of tests in 1.64 ± 0.1	39	104	23	7	2
Prop sig in 1.64 ± 0.075	0.600	0.619	0.667	0.600	1.000
One-sided p-value	0.1808	0.0188	0.1189	0.5000	0.5000
N of tests in 1.64 ± 0.075	30	84	18	5	1
Prop sig in 1.64 ± 0.05	0.458	0.661	0.538	0.500	1.000
One-sided p-value	0.7294	0.0076	0.5000	0.6875	0.5000
N of tests in 1.64 ± 0.05	24	62	13	4	1

C.5.4. Randomization Tests Figures

Figure EC.9 Share of Tests over Threshold (10%)

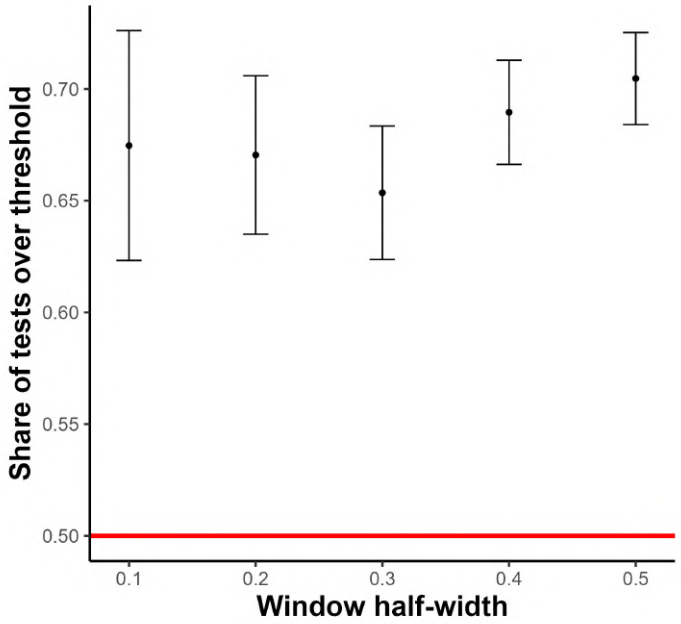
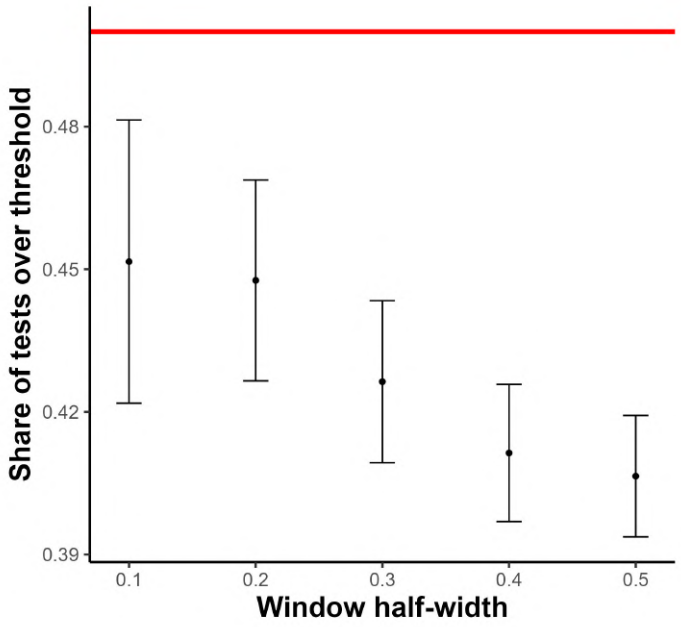


Figure EC.10 Share of Tests over Threshold (1%)

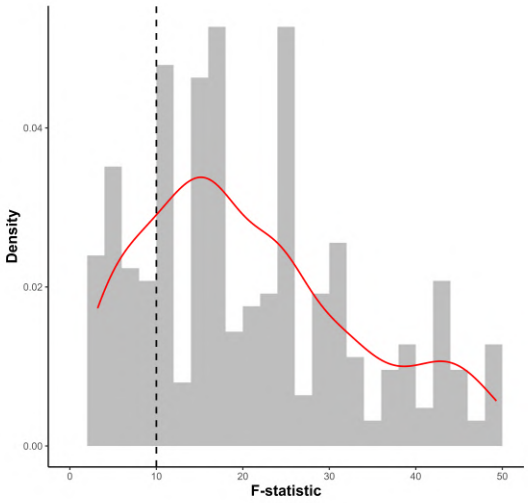


C.6. IV Inference

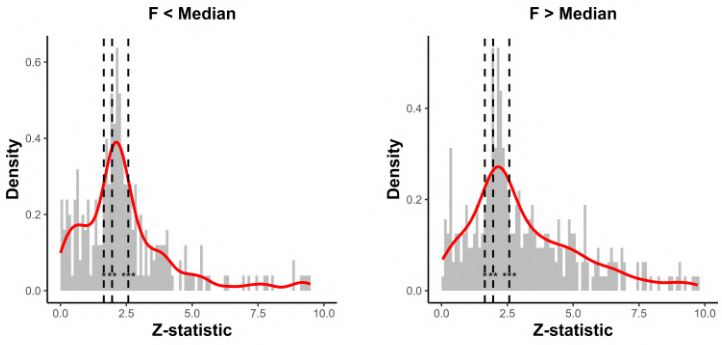
Instrumental Variables: F-statistics This analysis focuses on the instrumental variable (IV) method, where researchers have multiple stages to exercise discretion. First, we examine researchers' responsiveness to conventional cutoffs by analyzing first-stage estimates reported in IV studies, specifically the distribution of F-statistics in our sample. The first-stage F-statistic tests whether an instrumental variable is weak, indicating a low correlation with endogenous regressors.

Figure EC.11a shows the distribution of F-statistics within the interval $[0, 50]$. This sample includes 313 F-statistics, with approximately 20 percent being smaller than 10. This finding aligns with Andrews and Kasy (2019), who found that weak instruments are common, with nearly all published papers in their sample containing at least one such first-stage F-statistic. Similarly, Brodeur et al. (2020) found that 12 percent of 2,175 F-statistics reported values smaller than 10.

Figure EC.11 First-stage F-statistics



(a) F-statistic Distribution



(b) Associated z-statistic

Table EC.21 tests for discontinuities using randomization tests. The table presents binomial proportion tests, defining success as a first-stage F-stat value above 10. Reported p-values indicate the probability of observing the given (or greater) proportion, assuming an equal probability of being just above and below the threshold. Using windows as small as $0 < F < 35$, a statistically significant difference in the proportion around 10 is observed.[18]

The results indicate that both the first and second stages of IV studies display an excess of marginally significant test statistics. Further analysis reveals that second-stage results from compar-

[18] We understand that the proportion of tests between $0 < F < 10$ and $10 < F < 35$ can be different due to the width differences in the interval, we cannot reduce the sample further to a smaller window to have equal intervals due to sample size problem.

Table EC.21 Randomization Tests (F = 10)

	IV
Proportion above 10 in 10 ± 25	0.961
One-sided p-value	0.0000
N of tests in 10 ± 25	406
Proportion above 10 in 10 ± 20	0.956
One-sided p-value	0.0000
N of tests in 10 ± 20	367
Proportion above 10 in 10 ± 15	0.949
One-sided p-value	0.0000
N of tests in 10 ± 15	314

atively weak instruments show a higher proportion of z-statistics around conventional thresholds, suggesting greater p-hacking with weaker IVs (Figure EC.11b).

How Much Can We Trust IV Papers? Even though graphical inspection, randomization, and caliper tests show evidence of bunching of significant tests above the 5% significance threshold, we cannot disentangle whether this is due to p-hacking or publication bias. In this section, we explore the trustworthiness of evidence based on these hypothesis tests, assuming, in the best case, that the bunching behavior is solely due to publication bias and the tests are not manipulated.

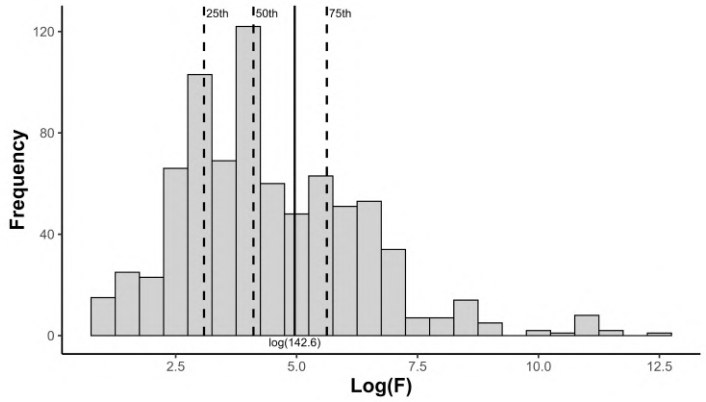
Due to the problem of invalid t-ratio inference using traditional approaches (Lee et al. 2022), we examine whether IS researchers are aware of this issue and whether their tests correct for this invalid inference approach. For IV models that are just-identified (i.e., the number of endogenous variables equals the number of instruments), we can recover valid t-values for inference based on the first-stage F-statistic. Out of the 209 IV articles analyzed, only 2% used Anderson and Rubin (1949) inference, and none used Lee et al. (2022).

Unfortunately, IS researchers often fail to implement the results of Staiger and Stock (1997) and Stock and Yogo (2005)'s papers. For example, researchers may pretest the F-statistic to be over the threshold of 10, instead of using the actual values provided in the Stock and Yogo (2005) tables. Hall et al. (1996) show that over-rejection can be even worse with pretesting for weak instruments. Andrews et al. (2019) discuss the practice of selectively dropping specifications when first-stage F-statistics do not meet a particular threshold, showing that severe distortion can result.

Despite the econometric literature affirming that Anderson and Rubin (1949) inference is valid and robust to weak instruments with several other attractive properties (Andrews et al. 2019, Lee et al. 2022), practitioners still prefer t-ratio inference. This preference might stem from a belief that any inferential approximation errors associated with the conventional t-ratio are negligible, or from the presumption that the inference has the intended significance or confidence level as long as the observed first-stage F-statistic is sufficiently large. For F that is sufficiently big (i.e., $F > 142.6$),

we can still have valid inferences for the causal coefficient of interest Lee et al. (2022). However, in our sample, at least 68 percent of papers have F values less than 142.6. Figure EC.12 displays the histogram of the F-statistics in our sample on a logarithmic scale. The twenty-fifth percentile, median, and seventy-fifth percentiles are 21.89, 60.8, and 268.72, respectively. In our sample[19], we found that 96.65% of the 209 IV articles analyzed (of the 210 IV articles in our sample) were just-identified models, indicating the importance and prevalence of the just-identified case in applied IS research. This is consistent with findings in economics, where Lee et al. (2022) and Andrews et al. (2019) reported a prevalence of 50%.

Figure EC.12 Distribution of First-stage F-statistics



Notes: Dashed lines correspond to the 25th (log(21.89)), 50th (log(60.8)), and 75th (log(268.72)) percentiles. The solid line represents the log(142.6).

Table EC.22 presents the two-way frequency distribution of whether the t values for the test hypothesis are greater than 1.96 and whether the F-statistic is greater than 10. We observe that approximately 70 percent of the specifications would yield a statistically significant coefficient under this practice.

Table EC.22 Frequency Table of First-stage F-statistics and t^2 at Conventional Critical Value

	$F < 10$	$F >= 10$
$t^2 < 1.96^2$	28 (0.036) [0.04]	193 (0.248) [0.199]
$t^2 >= 1.96^2$	39 (0.05) [0.051]	519 (0.666) [0.709]

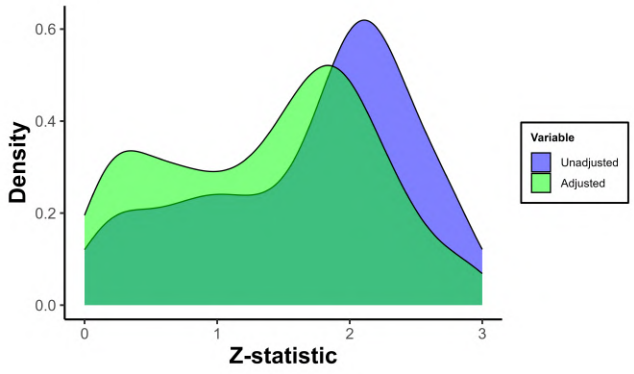
After correcting the t-ratio inference of tests using IV with the tF method by Lee et al. (2022), we find that 25 percent of significant test statistics become insignificant (specifically, 5 percent at the 1 percent level, 12 percent at the 5 percent level, and 8 percent at the 10 percent level),

[19] We also include the "Kleibergen-Paap" (KP) statistic (Kleibergen and Paap 2006) as first-stage F-statistics because in the case of a single endogenous regressor with a single instrument (i.e., just-identified model), $KP = F$ (Andrews et al. 2019).

indicating a nontrivial difference in inferences made in applied IS research. Figure EC.14 shows a concerning shift in the distribution of tests. The adjusted standard errors for these tests are, on average, 237%, 43%, and 31% larger than the conventional 2SLS standard errors at the 10%, 5%, and 1% significance levels, respectively.

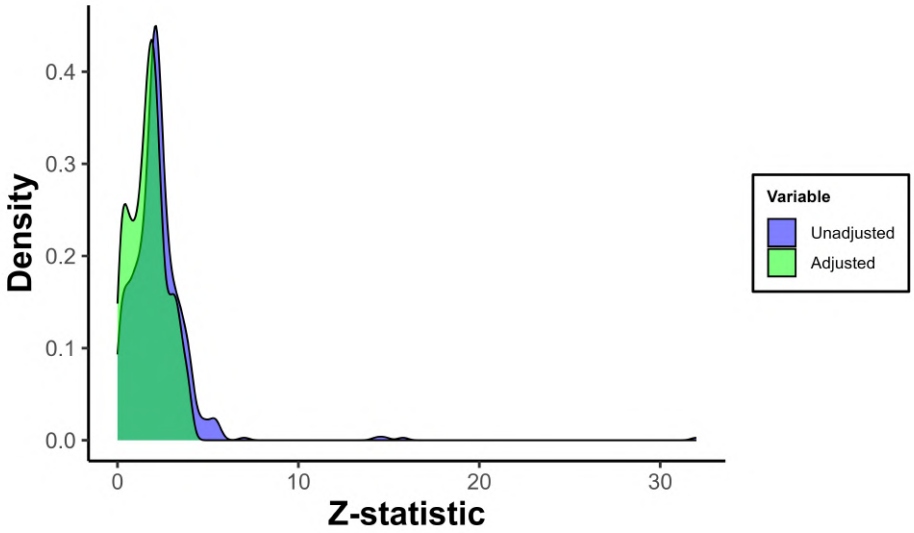
Our findings highlight opportunities to strengthen instrumental variable research through enhanced reporting practices. Given that approximately 25 percent of results are reclassified under weak-IV-robust inference, we suggest complementing conventional 2SLS with robust alternatives (Anderson-Rubin or tF-adjusted), reporting actual F-statistics alongside threshold classifications, and maintaining consistent specification choices.

Figure EC.13 Distribution of z-statistics Before and After Adjustment ($F < 142.6$, $z \in [0,3]$)



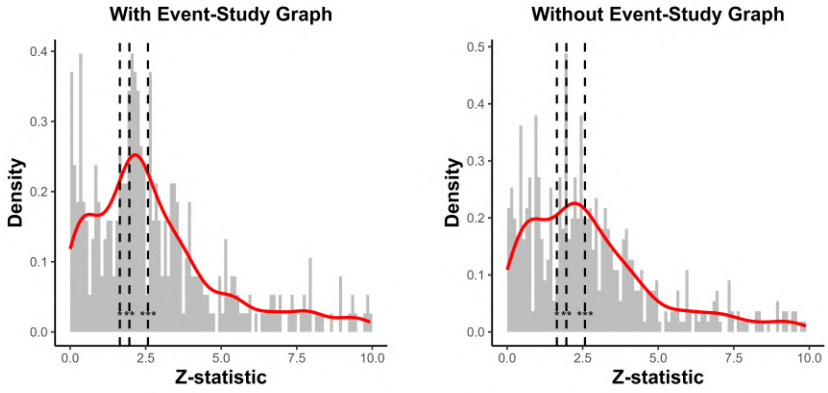
Notes: We limit the F-statistic to values under 142.6 and show $z \in [0,3]$ for better visualization. For the full distributions, see Figure EC.14.

Figure EC.14 Distribution of z-statistics Before and After Adjustment (Full Range)



Difference-in-differences: Event-study Graph

Figure EC.15 Comparison of z-statistics for DID Models with and without Event-Study Graphs



Given that convincing referees and editors of statistically significant policy impacts may be more challenging when raw data provides weak support, we examine whether the distribution of p-values differs systematically between DID articles that include versus exclude event-study graphs. Figure EC.15 presents the z-statistic distributions for these two subsamples of DID studies. While the overall density distributions exhibit some variation between DID articles with and without event-study graphs, the threshold-adjacent proportions that form the foundation of our bunching diagnostics remain similar. Our randomization tests (Table EC.14) demonstrate that across symmetric windows around the critical value of 1.96, the proportion of results falling just above versus just

below the significance threshold is statistically indistinguishable between the two subsamples. The sole exception occurs within the narrowest window (1.96 ± 0.05), where we observe a statistically significant difference in bunching patterns. Specifically, papers without event-study graphs exhibit significant evidence of bunching, while those with graphs do not. However, this finding should be interpreted cautiously, given the limited sample size for DID papers with event-study graphs ($n = 13$) in this narrow window, which substantially reduces our statistical power to detect bunching effects.

These results indicate that the inclusion of event-study graphs in DID articles does not meaningfully alter the proportion of statistically significant findings. This pattern suggests that neither the propensity for p-value manipulation nor the underlying likelihood of detecting true effects differs substantially based on the presence of event-study visualizations in the empirical analysis.

C.7. Exploratory Evidence on Revisions, Conditional on Acceptance

Peer review is the principal institutional check on selective reporting in the published record. Readers of our evidence on p-value heaping will naturally ask whether the review process mitigates its consequences. Because our data cover only accepted articles, the relevant question we can study is narrow: conditional on acceptance, do near-threshold results face more revision activity than otherwise similar results? An affirmative answer would suggest that reviewers detect statistical fragility and require additional work before publication. A null answer would suggest either that problematic manuscripts are primarily filtered at the rejection margin, which we cannot observe, or that near-threshold results are not systematically targeted for additional rounds once a manuscript proceeds to revision. Either way, documenting what happens on the acceptance path helps interpret where in the editorial pipeline any corrective pressure is likely to operate.

We focus on accepted articles because our data contain only published papers from ISJ, MS, JAIS, and ISR, the four outlets that report submission, revision, and acceptance timelines. Since we do not observe submissions that were rejected, we cannot study the rejection margin. Mechanisms are therefore not identified here. Both channels are plausible: reviewers may screen out problematic manuscripts earlier, or they may require more rounds among manuscripts that are eventually accepted.

We compare outcomes (e.g., number of revision rounds) for observations just to the right versus just to the left of the two-sided 5 percent cutoff. A naive RD that treats the running variable as unmanipulated shows a visible jump in revision rounds at $z = 1.96$. However, the z statistic is plausibly manipulated toward conventional cutoffs, so naive RD can be misleading.

To account for manipulation, we implement the RD procedure of Gerard et al. (2020). Near the cutoff, the method allows a mixture of two unobserved types: always-assigned units, which sort to one side of the cutoff, and potentially assigned units, which satisfy the standard RD conditions. Under local one-sided manipulation, which is consistent with selective movement toward statistical significance, the procedure (i) estimates the share of always-assigned units and (ii) reports intention-to-treat bounds for the potentially assigned units. Formally, with $T_i = 1$ for near-threshold significance and Y_i equal to revision rounds or review time, we bound ITT = $\mathbb{E}[Y_i(1) - Y_i(0) \mid i \text{ is potentially assigned}]$. Inference uses percentile bootstrap with 5,000 resamples.

Table EC.23 reports the estimated manipulation shares and the ITT bounds for revision rounds. For design-based causal studies, the estimated share of always-assigned units is 0.54 (95 percent CI 0.37, 0.71). For RCTs, the estimated share is 0.33 (95 percent CI 0.13, 0.53). Figure EC.16 plots the bound functions against the assumed manipulation share and marks the estimated shares with vertical lines. At the estimated shares, the confidence regions for the ITT among potentially assigned units include zero in both designs.

After accounting for local manipulation of the running variable, the manipulated-RD estimates indicate that, among potentially assigned units and conditional on acceptance, the effect of crossing the conventional threshold on revision activity is not distinguishable from zero at the estimated manipulation shares. This result does not imply that the review process is ineffective. Rather, it is consistent with the screening channel: detection may translate into earlier rejection rather than additional rounds for manuscripts that ultimately appear in print. Our data cannot adjudicate between these channels because rejected submissions are not observed. We therefore view this subsection as exploratory and descriptive, and we report it to locate the margin on which peer review may operate. The policy implication is pragmatic: interventions that target the pre-acceptance stage (e.g., statistical checklists at submission, structured robustness templates, or registered analysis plans for some designs) are likely to be more impactful than relying on incremental revisions among already accepted manuscripts.

Table EC.23 Estimated Treatment Effects of p-hacking/Publication Bias on Revisions

	Quasi-Experimental		**RCT**	
	Estimate	95% CI	Estimate	95% CI
Share of always-assigned units	0.54	[0.37; 0.71]	0.33	[0.13; 0.53]
ITT-Bounds on potentially-assigned	[0.05; 0.86]	[-0.49; 1.51]	[-0.11; 0.09]	[-0.26; 0.31]

(a) Quasi-Experimental (b) RCT

Figure EC.16 Effect of Marginally Significant p-Values from Unmanipulated Papers on Revisions in the Presence of a Share of Manipulated Papers

Notes: The horizontal dotted lines represent the confidence intervals for the lower and upper bounds of the average treatment effect. The horizontal black lines are the upper and lower bounds of the average treatment effect. The vertical solid line represents the estimated share of always-assigned units, whereas the vertical dotted lines represent the confidence interval for the estimated share.

Table EC.24 Effect of Having a Marginally Significant p-Value on Revisions

Dependent Variable:	Revisions					
Method	RCT			Quasi		
Threshold	1.64	1.96	2.58	1.64	1.96	2.58
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Constant	2.6***	2.6***	2.5***	2.3***	2.3***	2.6***
	(0.13)	(0.15)	(0.11)	(0.19)	(0.17)	(0.26)
Significance	-0.15	-0.06	0.09	0.27*	0.38**	-0.04
	(0.10)	(0.10)	(0.09)	(0.14)	(0.17)	(0.19)
Fit statistics						
R^2	0.006	0.001	0.002	0.01	0.02	0.0003
Observations	563	620	504	272	350	331

Clustered (article) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: We use an observation window for z-scores between 1.64 ± 0.5, 1.96 ± 0.5, and 2.58 ± 0.5.

C.8. Local False Discovery Rates Estimation

We adopt the two-groups model for $z \in \mathbb{R}$:

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z) \tag{4}$$

$$LFDR(z) = \frac{\pi_0 f_0(z)}{f(z)} \tag{5}$$

Null specification: estimate f_0 in two ways: (1) Theoretical nulls, $f_0 = \varphi$, (2) Empirical null with mean μ_0 and variance σ_0^2, estimated by central matching of the histogram or by maximizing the likelihood of observations in a central window $|z| \leq c$.

Marginal f: estimate f nonparametrically by either (1) kernel density estimation with bandwidth selected by biased-cross-validation, or (2) Efron's Poisson regression method that fits a natural spline to the binned z-histogram to recover a smooth f (Efron 2004).

π_0 estimation: implement Storey's direct approach using the high-p tail, with tuning parameter $\lambda \in [0.5, 0.9]$ and smoothing over λ.

Outputs: for each observed z-statistic, compute

$$\hat{LFDR}(z) = \frac{\hat{\pi}_0 \hat{f}_0(z)}{\hat{f}(z)} \tag{6}$$

$$\hat{FDR}(z) = \frac{\hat{\pi}_0 \hat{F}_0(z)}{\hat{F}(z)} \tag{7}$$

Future research can also use local false sign rate (i.e., probability of identifying the effect's sign wrong) to summarize for sign certainty, which can sometimes be more stable to π_0 misspecification (Stephens 2017).

Appendix D: Remedies

D.1. General Checklist for All Design-based Causal Methods

1. **Pre-registration and Study Design**
   (a) Pre-register the study design, hypotheses, and analysis plan.
   (b) Clearly define the research question and/or hypothesis.
   (c) Specify inclusion and exclusion criteria for the sample.
   (d) Outline the expected data sources and data collection methods.
   (e) Pre-specify the statistical models to be used.
2. **Data Transparency and Availability**
   (a) Make raw data available for replication.
   (b) Provide detailed data description and documentation.
   (c) Ensure data are anonymized to protect privacy.
3. **Robustness Checks**
   (a) Conduct robustness checks for key assumptions.
   (b) Report sensitivity analyses.

(c) Test for the presence of outliers and their influence on results.

(d) Perform placebo tests.

4. **Multiple Testing and p-Hacking Prevention**

(a) Adjust for multiple hypothesis testing using appropriate corrections (e.g., Bonferroni correction, False Discovery Rate (FDR)).

(b) Avoid post-hoc hypothesis formation (Hypothesizing After the Results are Known (HARKing)).

(c) Report all tested hypotheses, including non-significant results.

5. **Reporting and Documentation**

(a) Provide detailed descriptions of methods and procedures.

(b) Report effect sizes along with p-values.

(c) Discuss potential limitations and biases.

D.2. Specific Checklist for Each Method

D.2.1. Difference-in-Differences (DID)

1. **Study Design and Treatment Rollout**

(a) Start by plotting the treatment rollout to visualize the timing and extent of treatment across units (e.g., use the `panelView` R package).

(b) Document how many units are treated in each cohort.

(c) Plot the evolution of average outcomes across cohorts.

2. **Selection of Comparison Groups and Assumptions**

(a) Understand the comparison groups and assess the parallel trends assumption carefully to examine the exogeneity assumption of the event of interest:

i. Who/What decides treatment?

ii. What do they know?

iii. What type of selection is allowed?

(b) Test for parallel trends in the pre-treatment period.

(c) Provide graphical evidence of parallel trends.

(d) Use falsification tests to check for violations.

3. **Event Study and Specification Tests**

(a) Conduct event study analysis without any covariates to assess if the parallel trends assumption is plausible.

(b) Test for the presence of anticipatory effects.

(c) Explore dynamic treatment effects using event study analysis.

4. **Incorporating Covariates**

(a) If unconditional parallel trends are not plausible, incorporate covariates into the analysis. Covariates should be variables that affect the growth of untreated outcomes, allowing for covariate-specific trends.

(b) When using covariates, check for overlap:

i. If control groups are small, problems with overlap will probably arise.

ii. If you are okay with extrapolation, use regression adjustment DID procedures.

(c) Add additional controls or introduce more fixed effects to account for potential confounders.

5. **Staggered Adoption Settings**

(a) Use appropriate estimators for staggered treatment adoption, for example:

i. De Chaisemartin and d'Haultfoeuille (2020)

ii. Callaway and Sant'Anna (2021)

iii. Sun and Abraham (2021)

iv. Wooldridge (2023)

v. Stacked DID.

6. **Heterogeneous Treatment Effects**

(a) Explore and report heterogeneous treatment effects.

(b) Check for treatment effect heterogeneity across subgroups.

7. **Placebo and Sensitivity Tests**

(a) Implement placebo tests by assigning treatment to different periods or groups to validate the results.

(b) Conduct sensitivity analysis for violations of parallel trends (e.g., use the `honestDiD` R package).

(c) If conditional parallel trends are not plausible, consider alternative methods.

8. **Triple Difference Design**

(a) Conduct a triple difference (difference-in-difference-in-differences) design to further validate the results by adding an additional comparison group or time period.

9. **Reporting and Documentation**

(a) Ensure transparency in reporting:

i. Report all tested hypotheses, including non-significant results.

ii. Provide detailed descriptions of methods and procedures.

(b) Discuss potential limitations and biases.

D.2.2. Instrumental Variables (IV)

1. **Instrument Validity and Strength**

(a) **Relevance of the Instrument**

i. Justify the relevance of the instrument:

A. Provide a theoretical rationale for the instrument's relevance.

B. Discuss any potential violations of the exclusion restriction.

(b) **Tests for Instrument Strength, for example:**

i. Use the Stock and Yogo (2005) critical values to assess weak instrument concerns.

ii. Perform the Kleibergen and Paap (2006) rk LM statistic for weak identification.

iii. Conduct the Cragg and Donald (1993) Wald F-statistic test.

iv. Use the Anderson and Rubin (1949) test for weak instruments.

v. Apply the Sanderson and Windmeijer (2016) test or Angrist and Pischke (2009) multivariate F-test for multiple endogenous regressors.

vi. Assess the potential bias using the Bound, Jaeger, and Baker (BJB) approach (Bound et al. 1995).

(c) **Correcting for Weak Instruments** Use weak-instrument robust inference methods such as:

- The Anderson-Rubin test (Anderson and Rubin 1949).
- The conditional likelihood ratio (CLR) test by Moreira (2009).
- The tF procedure by Lee et al. (2022), which is recommended.

2. **Exclusion Restriction**

(a) Provide evidence that the instrument affects the outcome only through the endogenous regressor:

i. Discuss the theoretical basis for the exclusion restriction.

ii. Use economic or institutional arguments to support the exclusion restriction.

3. **Monotonicity Assumption**

(a) Ensure that the instrument satisfies the monotonicity assumption:

i. The instrument should affect the treatment status in one direction (either positively or negatively) for all units.

ii. Discuss theoretical reasons why the monotonicity assumption is likely to hold.

4. **Overidentification Tests**
   (a) Conduct overidentification tests if multiple instruments are used:
      i. Use Sargan-Hansen's J test to assess the validity of the instruments (Sargan 1958, Hansen 1982).
      ii. Conduct the Basmann test for overidentification (Basmann 1960).
5. **Robustness Checks**
   (a) **Alternative Instruments and Specifications**
      i. Re-estimate the model using alternative instruments that are theoretically valid.
      ii. Test the sensitivity of the results to different sets of instruments.
      iii. Use the Generalized Method of Moments to check robustness.
      iv. Apply heteroskedasticity-robust standard errors.
      v. Conduct a leave-one-out analysis by excluding one instrument at a time to check the stability of results.
   (b) **Graphical Methods**
      i. Plot the relationship between the instrument and the endogenous regressor.
      ii. Visualize the first-stage regression to assess the instrument's explanatory power.
   (c) **Heterogeneity Analysis**
      i. Explore treatment effect heterogeneity across different subgroups.
      ii. Use interaction terms to assess heterogeneous effects.
   (d) **Falsification Tests**
      i. Implement placebo tests by using instruments that should theoretically have no effect.
      ii. Test the robustness of results by introducing irrelevant instruments and checking for changes in the estimated coefficients.

D.2.3. Synthetic Control (SC)

1. **Selection of Control Units**
   (a) Justify the choice of control units.
   (b) Select control units that are similar to the treated unit in pre-treatment characteristics.
   (c) Use economic, demographic, or other relevant criteria to justify the selection.
   (d) Ensure that control units are not affected by the treatment.
   (e) Confirm that the control units do not experience spillover effects from the treatment.
   (f) Check for potential interference between treated and control units.
2. **Pre-treatment Fit**

(a) Demonstrate a good pre-treatment fit between the treated unit and the synthetic control.

(b) Plot the pre-treatment outcome trajectories of the treated unit and the synthetic control.

(c) Calculate pre-treatment root mean squared prediction error (RMSPE) to quantify the fit.

(d) Report the weights used to construct the synthetic control.

(e) Provide the exact weights assigned to each control unit.

(f) Discuss the rationale for the weight distribution.

3. **Variable Selection**

(a) Choose predictor variables carefully.

(b) Include both pre-intervention values of the outcome variable and other relevant predictors.

(c) Use a data-driven method to evaluate the predictive power of alternative sets of predictors.

(d) Consider cross-validation to select optimal weights for predictors.

4. **Robustness Checks**

(a) **Placebo Tests**

i. Conduct placebo tests using other untreated units or periods to validate the synthetic control.

ii. Generate placebo synthetic controls for units that did not receive the treatment.

iii. Compare the post-treatment outcomes of the treated unit to the distribution of placebo outcomes.

iv. Conduct placebo tests to check for the timing of the effect: Assign the treatment to a different time period and re-estimate the synthetic control to ensure that significant effects are not observed in periods where no treatment was applied.

(b) **Sensitivity Analysis**

i. Test the sensitivity of results to the exclusion of specific control units: Perform leave-one-out analysis by sequentially excluding each control unit and re-estimating the synthetic control.

ii. Examine the robustness of results to changes in the pre-treatment period length: Vary the length of the pre-treatment period to check for consistency in the synthetic control's performance.

iii. Assess the impact of different sets of predictors used in constructing the synthetic control: Re-estimate the synthetic control using alternative sets of pre-treatment characteristics.

(c) **Covariate Balance**

i. Check the balance of covariates between the treated unit and the synthetic control.

ii. Ensure that the pre-treatment characteristics of the synthetic control closely match those of the treated unit.

iii. Report standardized differences in covariates before and after constructing the synthetic control.

5. **Bias and Model Fit**

(a) Assess potential biases: Evaluate the risk of overfitting, especially with a large donor pool and a small number of pre-treatment periods.

(b) Ensure the model fit: Check if the synthetic control method reproduces the trajectory of the outcome variable for the treated unit over an extended period of time.

6. **Inference**

(a) Use permutation tests to derive p-values for the estimated effects.

(b) Compare the estimated treatment effect to the distribution of placebo effects.

(c) Measure the ratio of post-intervention fit relative to pre-intervention fit using the root mean squared prediction error (RMSPE).

(d) Consider one-sided inference for greater power in small sample settings.

(e) Report the permutation distribution of test statistics or placebo gaps for additional information.

D.2.4. Regression Discontinuity (RD)

1. **Design and Assumptions**

(a) Determine the validity of RD design:

- Report the process for assigning ratings and determining the cut-point.
- Conduct graphical and empirical analyses to confirm validity.

(b) Test for continuity of the running variable at the cutoff.

(c) Check for manipulation of the running variable (e.g., McCrary's test).

(d) Assess local randomization by testing covariate balance around the cutoff.

(e) Determine if the design is sharp or fuzzy; in fuzzy designs, estimate treatment probability and check for compliance.

2. **Bandwidth Selection**

(a) Justify bandwidth choice using data-driven methods (e.g., Imbens-Kalyanaraman).

(b) Test results with different bandwidths.

(c) Use parametric and non-parametric approaches for robustness.

3. **Functional Form**

(a) Test for appropriate functional form (e.g., linear, polynomial).

(b) Provide graphical analysis around the cutoff.

4. **Robustness Checks**
   (a) Conduct placebo tests at different cutoffs.
   (b) Check covariate balance on either side of the cutoff.
   (c) Perform sensitivity analysis by varying bandwidth and polynomial order.
5. **Inference**
   (a) Use robust standard errors and bias-corrected confidence intervals.
   (b) Apply bootstrap methods for small sample sizes.
   (c) For sharp RD, estimate the local average treatment effect (LATE) and use permutation tests for p-values.
   (d) For fuzzy RD, use IV techniques to estimate treatment effect, calculating the ratio of jumps in outcome and treatment probability at the cutoff.

D.2.5. Randomized Controlled Trials (RCT)

1. **Design, Registration, and Estimands**
   (a) **Registration and Pre-Analysis Plan (PAP)**
   i. Register the trial before data collection (e.g., AEA registry or OSF), and archive a time-stamped pre-analysis plan that specifies hypotheses, primary and secondary outcomes, analysis populations, and analysis methods (Schulz et al. 2010).
   ii. Disclose any deviations from the pre-analysis plan and flag exploratory analyses as such.
   (b) **Estimand Definition**
   i. Declare the primary estimand clearly, for example Intention-to-Treat for the average effect of assignment, and, if noncompliance exists, the Local Average Treatment Effect (LATE) for compliers under standard assumptions (Angrist and Pischke 2009, Imbens and Rubin 2015).
   ii. Define secondary estimands (e.g., quantile treatment effects, spillover effects, or subgroup effects), and state how they relate to decision objectives.
2. **Unit of Randomization, Blocking, and Assignment**
   (a) **Unit and Clustering**
   i. Justify the unit of randomization, individual or cluster, relative to the intervention and expected interference.
   ii. If clustering is used, report the intracluster correlation coefficient (ICC), design effect, and cluster size distribution.
   (b) **Blocking or Stratification**

i. Pre-specify blocking or stratification variables that are stable, predictive, and measured pre-treatment.

ii. Report the exact algorithm, for example permuted blocks, matched pairs, or re-randomization with a criterion such as Mahalanobis distance (Morgan and Rubin 2012).

(c) **Assignment Procedure and Allocation Concealment**

i. Describe the random seed generation, software, and any constraints such as capacity limits.

ii. Document allocation concealment and blinding or masking where feasible to reduce selection and measurement biases (Schulz et al. 2010).

3. **Sample Size, Power, and Minimum Detectable Effect (MDE)**

(a) **Power Calculations**

i. Provide ex ante power analysis, with assumed variance, ICC if clustered, expected take-up, and planned covariate adjustment.

ii. Report minimum detectable effects for the primary outcomes and for key subgroups.

(b) **Small Samples**

i. For few clusters, prefer randomization inference, wild cluster bootstrap, or small sample adjustments (Cameron et al. 2008, Tipton 2015). For more reference, please check "sandwich" package in R.

4. **Implementation Integrity and CONSORT Flow**

(a) **Trial Flow**

i. Provide a CONSORT-style flow diagram with counts at each stage: assessed, eligible, randomized, received treatment, analyzed, and reasons for exclusions (Schulz et al. 2010).

(b) **Treatment Delivery**

i. Describe protocol adherence, timing, intensity, and any deviations.

ii. Report contamination or crossovers, for example controls receiving elements of the treatment.

(c) **Measurement**

i. Document the timing of outcome measurement, instruments used, and enumerator training.

ii. If blinding of assessors is possible, report whether it was implemented.

5. **Baseline Balance and Randomization Checks**

(a) **Balance Diagnostics**

i. Present means by arm and standardized differences for pre-treatment covariates.

ii. Report an omnibus balance test, for example joint F-test or Mahalanobis distance.

(b) **Randomization Inference**

i. Where sample sizes are modest, compute randomization inference p-values for baseline balance and for primary outcomes using the exact assignment mechanism (Fisher 1949, Rosenbaum 2010).

(c) **Specification of Covariate Adjustment**

i. If using covariate adjustment, pre-specify the set and the functional form, and favor ANCOVA with the baseline outcome when available for higher power (McKenzie 2012, Lin 2013).

6. **Attrition, and Missing Data**

(a) **Attrition**

i. Report overall and differential attrition by arm with reasons.

ii. Test for differential attrition using pre-specified methods and assess robustness to attrition via Lee bounds or inverse probability weighting when appropriate.

(b) **Missing Data**

i. Describe missingness mechanisms and justify imputation or weighting schemes.

ii. Conduct sensitivity analyses to alternative assumptions about missing data.

7. **Spillovers, Interference, and General Equilibrium**

(a) **Stable Unit Treatment Value Assumption (SUTVA) and Interference**

i. Articulate the interference structure (e.g., within networks, or firms) and justify SUTVA or its relaxation.

ii. If interference is plausible, use designs that allow estimation of direct and spillover effects, such as two-stage randomization, partial population designs, or saturation designs (Hudgens and Halloran 2008, Athey and Imbens 2022).

(b) **Analysis under Interference**

i. Define exposure mappings (e.g., treated neighbors share) and report effects by exposure condition.

ii. Adjust inference for network or spatial correlation when relevant (Aronow and Samii 2017).

8. **Outcomes, Multiplicity, and P-hacking Safeguards**

(a) **Outcomes**

i. Clearly label primary outcomes and families of related secondary outcomes. Justify any indices or composites and provide construction details (Anderson 2008).

(b) **Multiplicity**

i. Control the familywise error rate or the false discovery rate across outcome families using pre-specified procedures (Holm 1979, Westfall and Young 1993, Benjamini and Hochberg 1995).

(c) **Specification Transparency**

i. Publish a complete specification table that lists every model estimated, covariate set, sample restrictions, and outcome transformations. Identify confirmatory versus exploratory specifications.

ii. Provide a specification curve or multiverse analysis figure in the appendix to show robustness to reasonable analytic choices (Simonsohn et al. 2020).

(d) **Distributional Diagnostics**

i. Report p-value and z-score distributions for the full set of confirmatory outcomes and pre-registered hypotheses. Discuss local false discovery rate or p-curve results where informative (Efron 2010, Simonsohn et al. 2014).

9. **Robustness and Sensitivity**

(a) **Alternative Specifications**

i. Present unadjusted and ANCOVA estimates, alternative reasonable covariate sets, and alternative functional forms for outcomes as applicable.

(b) **Leverage and Influence**

i. Report leave-one-cluster-out and leave-one-block-out estimates for cluster or blocked designs.

ii. Provide influence diagnostics for outliers and high leverage observations.

(c) **Placebos and Negative Controls**

i. Include placebo outcomes measured prior to treatment or known to be unaffected, and negative control groups if available.

10. **External Validity and Transportability**

(a) **Study Context**

i. Describe the sampling frame, participation rates, and setting features that limit or enable transportability.

(b) **Transport Methods**

i. If extrapolating, document and justify any reweighting or transportability methods and report sensitivity to alternative target populations.

11. **Ethics, Data, and Replicability**

(a) **Ethical Oversight**

i. State IRB approval, informed consent procedures, and data protection measures.

(b) **Replication Package**

i. Provide de-identified data, complete code, randomization scripts, and a README file that allows exact reproduction of tables and figures.

ii. Archive the replication package in a stable repository and include a persistent identifier.

D.3. Reviewer and Editor Checklist

1. **Replicability and Transparency**

(a) Ensure data and code are shared for replication:

i. Provide access to raw data and code through public repositories (e.g., GitHub, OSF).

ii. If data are proprietary or cannot be shared, provide detailed descriptions of the data and the steps to obtain similar data.

iii. Where possible, provide synthetic datasets that mirror the properties of the original data without compromising privacy or proprietary concerns.

(b) Check for completeness and transparency in reporting:

i. Ensure that all variables, data cleaning steps, and transformations are clearly documented.

ii. Provide a clear and detailed codebook that explains each variable in the dataset.

iii. Report any deviations from the pre-registered analysis plan, with justifications.

2. **Evaluation of Assumptions**

(a) Critically evaluate the validity of key assumptions for each method:

i. For DID: Validate the parallel trends assumption using pre-treatment data.

ii. For IV: Justify the relevance and exogeneity of the instrument.

iii. For Synthetic Control: Demonstrate a good pre-treatment fit between treated and synthetic units.

iv. For RD: Test for the continuity of the running variable at the cutoff.

(b) Look for evidence of robustness checks and sensitivity analyses:

i. Conduct robustness checks for alternative model specifications.

ii. Perform sensitivity analyses to test the impact of key assumptions and potential violations.

iii. Report results of placebo tests to validate findings.

3. **Balance of Reporting**

(a) Ensure that both significant and non-significant results are reported:

i. Report all hypothesis tests conducted, including those that do not support the hypothesis.

ii. Provide a balanced discussion that considers all findings, not just significant ones.

(b) Assess the discussion of potential biases and limitations:

i. Clearly articulate any potential biases introduced by the study design or data limitations.

ii. Discuss the generalizability of the findings and any limitations related to the sample or context.

iii. Suggest areas for future research to address unresolved issues or limitations.

4. **Ethical Considerations**

(a) Ensure ethical standards are maintained in data collection and analysis:

i. Obtain necessary ethical approvals or waivers from relevant review boards.

ii. Ensure informed consent is obtained from participants, where applicable.

iii. Maintain the confidentiality and privacy of participants' data.

(b) Verify that conflicts of interest are disclosed:

i. Disclose any financial, personal, or professional conflicts of interest that may influence the research.

ii. Include a statement on the funding sources and their role, if any, in the research process.

Appendix E: Code Book for Academic Paper Data Collection

Overview

This code book describes the variables collected from primary academic papers. The dataset includes both paper-level characteristics and author-level information, with a focus on empirical methods, statistical reporting, and author demographics.

PAPER-LEVEL VARIABLES

Table EC.25 Basic Paper Information

Variable	Description	Type	Notes
article	Article identifier	Text	Unique identifier for each paper
doi	Digital Object Identifier	Text	Standard academic identifier
year	Publication year	Numeric	Year the paper was published
journal	Journal name	Text	Full name of publishing journal
title	Paper title	Text	Complete title of the paper
url	Paper URL	Text	Direct link to paper
num_authors	Number of authors	Numeric	Total count of authors
num_funders	Number of funders	Numeric	Total count of funding sources

Table EC.26 Methodology and Research Design

Variable	Description	Type	Coding
method	Primary research method	Text	DID, IV, RCT, SC, RD
table	Table number analyzed	Text/Numeric	Specific table containing main results
event_study_graph	Event study visualization	Binary	1 = includes event study graph, 0 = no graph

Table EC.27 Statistical Results and Reporting

Variable	Description	Type	Notes
t	t-value	Numeric	t-value of interest
mu	Treatment effect/coefficient	Numeric	Main coefficient of interest
sd	Standard error	Numeric	Standard error of main coefficient
p-value	p-value	Numeric	Statistical significance level
star05	Significance at 5% level	Binary	1 = significant at $p < 0.05$, 0 = not significant
side	Sidedness of test	Numeric	2 = "two-sided" or 1 = "one-sided"
ci_lower	Confidence interval lower bound	Numeric	Lower bound of 95% CI
ci_upper	Confidence interval upper bound	Numeric	Upper bound of 95% CI
obs	Number of observations	Numeric	Sample size

Table EC.28 Instrumental Variables (IV) Analysis

Variable	Description	Type	Notes
fstat_iv (1st stage)	First-stage F-statistic	Numeric	F-stat from first stage regression
fstat_iv_table	IV F-stat table reference	Text	Table/section where F-stat appears
stock_yogo_weak_iv	Stock-Yogo weak IV test	Numeric	Critical value comparison
Cragg–Donald F	Cragg-Donald F-statistic	Numeric	Test for weak instruments
Kleibergen-Paap rk Wald F	Kleibergen-Paap F-statistic	Numeric	Robust weak instrument test
iv_10_percent	IV critical value (10%)	Numeric	10% maximal IV size
just-identified (1)	Just-identified model	Binary	1 = just-identified, 0 = over-identified

Table EC.29 Robustness Tests and Citations

Variable	Description	Type	Notes
Anderson and Rubin (1949)	Anderson-Rubin test	Binary/Numeric	1 = test performed/reported
Angrist and Kolesár (2023)	Recent IV methodology	Binary	1 = cites/uses this approach
Stock and Yogo (2005)	Stock-Yogo methodology	Binary	1 = cites/uses this approach
Lee et al. (2022)	Recent methodology reference	Binary	1 = cites/uses this approach
comment	Additional notes	Text	Free-text field for qualitative notes

AUTHOR-LEVEL VARIABLES (Paper Dataset)

Table EC.30 Author Identification (up to 7 authors per paper)

Variable	Description	Type	Notes
full_name_1 to full_name_7	Author full names	Text	Complete name of each author
affiliation_name_1 to affiliation_name_7	Author affiliations	Text	Institution at time of publication
ORCID_1 to ORCID_7	ORCID identifiers	Text	Unique researcher identifiers
sequence_1 to sequence_7	Author order sequence	Numeric	Position in author list
funder_name_1	Primary funder	Text	Main funding organization

AUTHOR-LEVEL VARIABLES (Author Dataset)

Table EC.31 Basic Author Information

Variable	Description	Type	Notes
article	Article identifier	Text	Links to paper dataset
full_name	Author full name	Text	Complete author name
year	Publication year	Numeric	Year of publication
university	Current affiliation	Text	University at time of data collection

Table EC.32 Academic Position and Career

Variable	Description	Type	Coding
title_at_pub	Title at publication	Text	Academic rank when paper published
title_at_pub (notes)	Additional title notes	Text	Clarifications or special circumstances
phd_school	PhD institution	Text	University where PhD was earned
year_grad	PhD graduation year	Numeric	Year PhD was completed
year_in_phd	Years taken to finish PhD	Numeric	Years in PhD program
first_placement	First academic position	Text	Initial post-PhD appointment

Table EC.33 Editorial and Demographic Information

Variable	Description	Type	Coding
was_editor	Editorial board member	Binary	1 = editor at top IS journals (MISQ, ISR, MS, JAIS), 0 = not editor
female	Gender	Binary	1 = Female, 0 = Male

DATA COLLECTION NOTES

1. Scope and Coverage:

- **Field**: Information Systems
- **Time Period**: 2000-2023
- **Journal Coverage**: MIS Quarterly, Information Systems Research, Journal of the Association for Information Systems, Management Science (IS Department only), Journal of Management Information Systems, Journal of Information Technology, Information Systems Journal.

2. Selection Criteria

2.1 Inclusion Criteria

Study Design:

- Empirical papers using quantitative methods
- Papers with clearly reported statistical results (coefficients, standard errors, p-values)
- Studies with identifiable treatment effects or causal claims
- Papers using experimental, quasi-experimental, or observational designs with causal inference

Statistical Reporting Requirements:

- Must report coefficient estimates and standard errors (or t-statistics)
- Must include significance testing results (p-values or significance stars)
- Sufficient detail to extract effect sizes and precision measures

2.2 Exclusion Criteria

Study Design:

- Purely theoretical or conceptual papers
- Literature reviews, meta-analyses, or survey papers
- Case studies without statistical analysis
- Qualitative studies without quantitative components

Statistical Reporting:

- Papers with insufficient statistical detail for effect size extraction
- Papers with unclear or non-standard statistical reporting
- Conference papers or working papers (journal publications only)

Language and Accessibility:

- Non-English publications
- Papers behind paywalls without institutional access
- Retracted or corrected papers (unless correction provides adequate data)

Coding Procedures:

- **Inter-rater Reliability**: 99.4%.
- **Missing Data**: We contacted authors for missing data.

Key Limitations:

- Some variables may not be available for all papers.
- Author demographic information may be incomplete.
- IV statistics only available for papers using instrumental variables.

VARIABLE RELATIONSHIPS

Cross-References

- Paper-level and author-level datasets linked via article identifier.
- Multiple authors per paper captured in separate author-level records.

Derived Variables

- Author sequence determined from publication order.

Appendix F: Open Science in IS

In our analysis, a significant number of IV papers lacked detailed reporting on the strength of the instruments used. Specifically, 57% of the papers failed to report the exact values of the first-stage F-test, and 38% did not explicitly present any form of IV strength test, such as the first-stage F-test, Stock-Yogo, Cragg-Donald, or Kleibergen-Paap Wald F Statistic. Some papers mentioned IV

strength tests in general terms but failed to document the specific statistics explicitly. To increase transparency in IS, results of valid tests should be documented explicitly and transparently.

To assess transparency, particularly for the IV method, we contacted authors of 181 papers. The response rate was low; only 57 authors (31.49%) replied to our inquiry. Out of this responsive subset, merely 38 authors (20.99%) were able to locate and were willing to share their data and code. We were unable to contact six authors (3.31%) because their email addresses were no longer valid. Additionally, 19 authors (10.50%) responded that they could not access the data and/or code at the time of our request. One observation from this exercise is that IS researchers do not have a robust procedure to document their data and code more thoroughly after publishing their papers to facilitate future reference and inquiry.

References

Anderson ML (2008) Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association* 103(484):1481–1495.

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

Andrade C (2021) HARKing, cherry-picking, p-hacking, fishing expeditions, and data dredging and mining as questionable research practices. *The Journal of Clinical Psychiatry* 82(1):20f13804.

Andrews I, Stock JH, Sun L (2019) Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics* 11:727–753.

Angrist JD, Pischke J (2009) *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Aronow PM, Samii C (2017) Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics* 11(4):1912–1947.

Athey S, Imbens GW (2017) The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives* 31(2):3–32.

Athey S, Imbens GW (2022) Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics* 226(1):62–79.

Ayorinde AA, Williams I, Mannion R, Song F, Skrybant M, Lilford RJ, Chen Y-F (2020) Publication and related biases in health services research: a systematic review of empirical evidence. *BMC Medical Research Methodology* 20:1–12.

Bartoš F, Schimmack U (2022) Z-curve 2.0: Estimating replication rates and discovery rates. *Meta-Psychology* 6.

Basmann RL (1960) On finite sample distributions of generalized classical linear identifiability test statistics. *Journal of the American Statistical Association* 55(292):650–659.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57(1):289–300.

Bound J, Jaeger DA, Baker RM (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.

Brodeur A, Cook N, Heyes A (2020) Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics. *American Economic Review* 110(11):3634–3660.

Brodeur A, Cook N, Heyes A (2022) Methods matter: p-hacking and publication bias in causal analysis in economics: Reply. *American Economic Review* 112(9):3137–3139.

Brunner J, Schimmack U (2020) Estimating population mean power under conditions of heterogeneity and selection for significance. *Meta-Psychology* 4.

Callaway B, Sant'Anna PHC (2021) Difference-in-differences with multiple time periods. *Journal of Econometrics* 225(2):200–230.

Cameron AC, Gelbach JB, Miller DL (2008) Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics* 90(3):414–427.

Cova F, Strickland B, Abatista A, Allard A, Andow J, Attie M, Beebe J, Berniūnas R, Boudesseul J, Colombo M, et al. (2021) Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology* 12:9–44.

Cragg JG, Donald SG (1993) Testing identifiability and specification in instrumental variable models. *Econometric Theory* 9(2):222–240.

De Chaisemartin C, d'Haultfoeuille X (2020) Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review* 110(9):2964–2996.

De Long JB, Lang K (1992) Are all economic hypotheses false?. *Journal of Political Economy* 100(6):1257–1272.

Easterbrook PJ, Gopalan R, Berlin JA, Matthews DR (1991) Publication bias in clinical research. *The Lancet* 337(8746):867–872.

Efron B (2010) *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press.

Elston DM (2021) Cherry picking, HARKing, and P-hacking. *Journal of the American Academy of Dermatology*. `https://doi.org/10.1016/j.jaad.2021.06.844`.

Eysenbach G (2004) Tackling publication bias and selective reporting in health informatics research: register your eHealth trials in the International eHealth Studies Registry. *Journal of Medical Internet Research* 6(3):e134.

Fisher RA (1949) *The design of experiments*. Oliver & Boyd.

Franco A, Malhotra N, Simonovits G (2014) Publication bias in the social sciences: Unlocking the file drawer. *Science* 345(6203):1502–1505.

Fraser H, Parker T, Nakagawa S, Barnett A, Fidler F (2018) Questionable research practices in ecology and evolution. *PloS One* 13(7):e0200303.

Freedman DA (2008) On regression adjustments in experiments with several treatments. *The Annals of Applied Statistics* 2(1):176–196.

Friedman CP, Wyatt JC (2001) Publication bias in medical informatics. *Journal of the American Medical Informatics Association* 8(2):189–191.

Gerard F, Rokkanen M, Roth C (2020) Bounds on treatment effects in regression discontinuity designs with a manipulated running variable. *Quantitative Economics* 11(3):839–870.

Gerber A, Malhotra N (2008) Do statistical reporting standards affect what is published? Publication bias in two leading political science journals. *Quarterly Journal of Political Science* 3(3):313–326.

Gerber AS, Malhotra N (2008) Publication bias in empirical sociological research: Do arbitrary significance levels distort published results?. *Sociological Methods & Research* 37(1):3–30.

Hall AR, Rudebusch GD, Wilcox DW (1996) Judging instrument relevance in instrumental variables estimation. *International Economic Review* 37(2):283–298.

Hansen LP (1982) Large sample properties of generalized method of moments estimators. *Econometrica* 50(4):1029–1054.

Harrison JS, Banks GC, Pollack JM, O'Boyle EH, Short J (2017) Publication bias in strategic management research. *Journal of Management* 43(2):400–425.

He J, King WR (2008) The role of user participation in information systems development: implications from a meta-analysis. *Journal of Management Information Systems* 25(1):301–331.

Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015) The extent and consequences of p-hacking in science. *PLoS Biology* 13(3):e1002106.

Hollenbeck JR, Wright PM (2017) Harking, sharking, and tharking: Making the case for post hoc analysis of scientific data. *Journal of Management* 43(1):5–18.

Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6(2):65–70.

Hudgens MG, Halloran ME (2008) Toward causal inference with interference. *Journal of the American Statistical Association* 103(482):832–842.

Imbens GW, Rubin DB (2015) *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jeyaraj A, Dwivedi YK (2020) Meta-analysis in information systems research: Review and recommendations. *International Journal of Information Management* 55:102226.

Kepes S, Thomas MA (2018) Assessing the robustness of meta-analytic results in information systems: Publication bias and outliers. *European Journal of Information Systems* 27(1):90–123.

Kleibergen F, Paap R (2006) Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics* 133(1):97–126.

Lee DS, McCrary J, Moreira MJ, Porter J (2022) Valid t-ratio Inference for IV. *American Economic Review* 112(10):3260–3290.

Lin W (2013) Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics* 7(1):295–318.

McKenzie D (2012) Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics* 99(2):210–221.

Moreira MJ (2009) Tests with correct size when instruments can be arbitrarily weak. *Journal of Econometrics* 152(2):131–140.

Morgan KL, Rubin DB (2012) Rerandomization to improve covariate balance in experiments. *The Annals of Statistics* 40(2):1263–1282.

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science* 349(6251):aac4716.

Rosenbaum PR (2010) *Design of observational studies*. Springer.

Sanderson E, Windmeijer F (2016) A weak instrument F-test in linear IV models with multiple endogenous variables. *Journal of Econometrics* 190(2):212–221.

Sargan JD (1958) The estimation of economic relationships using instrumental variables. *Econometrica* 26(3):393–415.

Schulz KF, Altman DG, Moher D, Consort Group, et al. (2010) CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC Medicine* 8.

Siegel M, Eder JSN, Wicherts JM, Pietschnig J (2022) Times are changing, bias isn't: A meta-meta-analysis on publication bias detection practices, prevalence rates, and predictors in industrial/organizational psychology. *Journal of Applied Psychology* 107(11):2013.

Simonsohn U, Simmons JP, Nelson LD (2020) Specification curve analysis. *Nature Human Behaviour* 4(11):1208–1214.

Song F, Hooper L, Loke YK (2013) Publication bias: what is it? How do we measure it? How do we avoid it?. *Open Access Journal of Clinical Trials* 71–81.

Staiger D, Stock JH (1997) Instrumental variables regression with weak instruments. *Econometrica* 65(3):557–586.

Stanley TD (2005) Beyond publication bias. *Journal of Economic Surveys* 19(3):309–345.

Stefan AM, Schönbrodt FD (2023) Big little lies: A compendium and simulation of p-hacking strategies. *Royal Society Open Science* 10(2):220346.

Stephens M (2017) False discovery rates: a new deal. *Biostatistics* 18(2):275–294.

Stock JH, Yogo M (2005) Testing for weak instruments in Linear IV regression. Andrews DWK, Stock JH, eds. *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg* (Cambridge University Press, Cambridge).

Sun L, Abraham S (2021) Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics* 225(2):175–199.

Thornton A, Lee P (2000) Publication bias in meta-analysis: its causes and consequences. *Journal of Clinical Epidemiology* 53(2):207–216.

Tipton E (2015) Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods* 20(3):375–393.

Tumber MB, Dickersin K (2004) Publication of clinical trials: accountability and accessibility. *Journal of Internal Medicine* 256(4):271–283.

Weiß B, Wagner M (2011) The identification and prevention of publication bias in the social sciences and economics. *Jahrbücher für Nationalökonomie und Statistik* 231(5-6):661–684.

Weinhardt C, van der Aalst WMP, Hinz O (2019) Introducing registered reports to the information systems community. *Business & Information Systems Engineering* 61:381–384.

Westfall PH, Young S (1993) *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley & Sons.

Wooldridge JM (2010) *Econometric analysis of cross section and panel data*. MIT Press.

Wooldridge JM (2023) Simple approaches to nonlinear difference-in-differences with panel data. *The Econometrics Journal* 26(3):C31–C66.