American Economic Review 2020, 110(11): 3634–3660
https://doi.org/10.1257/aer.20190687

Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics[†]

By Abel Brodeur, Nikolai Cook, and Anthony Heyes*

The credibility revolution in economics has promoted causal identification using randomized control trials (RCT), difference-in-differences (DID), instrumental variables (IV) and regression discontinuity design (RDD). Applying multiple approaches to over 21,000 hypothesis tests published in 25 leading economics journals, we find that the extent of p-hacking and publication bias varies greatly by method. IV (and to a lesser extent DID) are particularly problematic. We find no evidence that (i) papers published in the Top 5 journals are different to others; (ii) the journal "revise and resubmit" process mitigates the problem; (iii) things are improving through time. (*JEL* A14, C12, C52)

The credibility revolution in empirical economics has been marked by a shift towards using methods explicitly focused on causal inference (Angrist and Pischke 2010). Experimental and quasi-experimental methods, namely randomized control trials (RCT), difference-in-differences (DID), instrumental variables (IV), and regression discontinuity design (RDD), have become the norm in applied microeconomics (Biddle and Hamermesh 2017, Panhans and Singleton 2017).

In this paper we explore the relationship between inference method and statistical significance. Evidence of selective publication and specification searching in economics and other disciplines is, by now, voluminous (Ashenfelter et al. 1999, Bruns et al. 2019, Casey, Glennerster, and Miguel 2012, De Long and Lang 1992, Havránek 2015, Henry 2009, Ioannidis 2005, Ioannidis, Stanley, and Doucouliagos 2017, Leamer 1983, Leamer and Leonard 1983, Lenz and Sahn forthcoming, McCloskey 1985, Simmons et al. 2011, Stanley 2008). Publication bias, whereby the statistical significance of a result determines the probability of publication, is likely a reflection of the peer review process. The term *p*-hacking refers to a variety of practices that a researcher might (consciously or unconsciously) use to generate

* Brodeur: Department of Economics, University of Ottawa (email: abrodeur@uottawa.ca); Cook: Department of Economics, University of Ottawa (email: ncook@uottawa.ca) Heyes: Department of Economics, University of Ottawa and University of Exeter Business School (email: anthony.heyes@uottawa.ca). Stefano DellaVigna was the coeditor for this article. We are grateful to four anonymous referees for suggestions and comments. We also thank Andrew Foster, Jason Garred, Dan Hamermesh, Fernando Hoces de la Guardia, Pierre Mouganie, Jon de Quidt, Matt Webb, and seminar participants at the 4th IZA Junior/Senior Symposium, ASSA annual meeting, BITSS annual meeting, Carleton University, MAER-Net Colloquium, and the University of Ottawa for useful remarks and encouragement. We thank Mohammad Al-Azzam, Lauren Gallant, Joanne Haddad, Jessica Krueger, and Taylor Wright for research assistance. Lastly, Abel is forever thankful to Jedi Marc, Mathias, and Yanos. Errors are ours.

[†] Go to https://doi.org/10.1257/aer.20190687 to visit the article page for additional materials and author disclosure statement.

"better" *p*-values, perhaps (but not necessarily) in response to the difficulty of publishing statistically insignificant results (Abadie 2020, Blanco-Perez and Brodeur 2020, Doucouliagos and Stanley 2013, Furukawa 2019, Havránek Irsova, and Vlach 2018, Stanley 2005).[1] The link between method and statistical significance could be of interest to policymakers or others who use empirical evidence to inform decisions and policies, as publication bias and *p*-hacking will create literatures with an artificially high percentage of false positives.

The central questions in this paper are (i) what is the extent of *p*-hacking and publication bias in leading economics journals? (ii) does it depend upon the method of inference used, or other author and article characteristics? (iii) does the review process exacerbate or attenuate the problem? (iv) is there improvement over time?

To answer these and a number of secondary questions, we harvest the universe of hypothesis tests reported in papers using these four methods in 25 top economics journals for the years 2015 and 2018.

Taken as a whole, the distribution of published test statistics exhibits a two-humped or camel shape, with "missing" tests just before conventional significance thresholds, i.e., $z = 1.65$, and a "surplus" just after (Brodeur et al. 2016). The pattern is similar across Top 5 and non-Top 5 journals, and there is no discernible change in pattern over time. We also find much less *p*-hacking in our sample of tests from economics journals than has been found in other disciplines such as political science and sociology (Gerber and Malhotra 2008a, Gerber and Malhotra 2008b).

We use three approaches to document the differences in *p*-hacking, all of which compare the quasi-experimental methods against the benchmark of RCTs. Ravallion et al. (2018) observes how the RCT, randomization by the researcher, has come to be widely regarded as the gold standard against which to compare observational results. Imbens (2010, p. 407) asserts that "(r)andomized experiments occupy a special place in the hierarchy of evidence, namely at the very top."[2]

First, we test for discontinuities in the probability of a test statistic appearing just above or below a conventional statistical threshold. If the underlying distribution of test statistics (for any method) is continuous and infinitely differentiable, any surplus of outcomes just above a threshold is taken as evidence of publication bias or *p*-hacking. We find that IV and DID test statistics are not distributed equally around the one- and two-star significance thresholds. Within 10 percent of the threshold ($1.76 < z < 2.16$), there are 18 percent more significant than insignificant IV test statistics. For DID, there are 25 percent more. In contrast, RDD has only 3 percent more, while RCT has fewer statistically significant tests than insignificant tests.

Second, we apply a caliper test, as in Gerber and Malhotra (2008a). Caliper tests also focus on the distribution of *p*-values, close to arbitrary significance thresholds. We find that the proportion of tests that are marginally significant in IV articles is about 10 percentage points higher than the 47 percent for RCTs. In contrast, we find no evidence that the portion of tests that are marginally significant in RDD articles is significantly higher than for RCTs.

[1] Such practices might include continuing to collect data, strategically selecting covariates, or imposing sample restrictions until a significance threshold is met.

[2] It is worth noting that there have been thoughtful critiques of RCT as gold standard, see for example Deaton and Cartwright (2018). As far as the propensity for the published literature using a particular method to exhibit *p*-hacking and publication bias, our results indicate RCT outperforms the other methods.

A potential explanation is that different authors or fields might be more or less prone to *p*-hacking or may be more or less likely to rely on one of the four methods. For instance, Brodeur et al. (2016) provide suggestive evidence that less experienced researchers, on average, *p*-hack more. We show that controlling for author characteristics (e.g., experience and institution ranking) has no impact, suggesting that selection of authors into the use of particular methods is unlikely to drive our results. The inclusion of field and journal fixed effects decreases the gap between IV and RCT estimates, but they remain large, positive, and statistically significant. However, the inclusion of field and journal fixed effects reduces the size of the DID estimate and makes it not significantly different than RCT at conventional levels.

Third, we extend the methodology in Brodeur et al. (2016) to quantify the excess (or dearth) of *z*-values over significance regions by comparing the observed distribution of test statistics for each method to a counterfactual distribution that we would expect in the absence of *p*-hacking and publication bias. The results are consistent with our previous findings; the extent of misallocated tests differs substantially between methods. About 16 percent of statistically insignificant IV results are "missing," later to be found as statistically significant. In comparison, misallocation for RCTs is one-tenth the size of IV, at 1.5 percent.

Considering each method's body of published research as a distinct literature, our results suggest that the IV and, to a lesser extent, DID research bodies have substantially more *p*-hacking and/or selective publication than those based on RCT and RDD. This leads naturally to the question of why we find differences across methods. While we show that author and article characteristics do not appear important, another potential explanation is that some methods offer researchers different degrees of freedom than others. For instance, when using a non-experimental method like IV there are many points at which a researcher exercises discretion in ways that could affect statistical significance. With regard to the first stage of IV, we document a sizable over-representation of first stage *F*-statistics just over the conventional threshold of 10. Interestingly, the degree of *p*-hacking in the second stage is related to strength in the first stage. Second stage results from relatively weak IVs have a much higher proportion of *z*-statistics around conventional thresholds. We also provide evidence that IV results in RCTs with partial compliance display less *p*-hacking than IV results in observational studies.[3]

Another potential explanation for our main observations is that the attitudes of editors and/or referees toward null results vary systematically with method. For example, there may be more tolerance of a null result if it is the result of an RCT. We investigate the role of the review process by comparing the distributions of test statistics in the published version of each article with those from earlier working paper versions, and find no meaningful difference.

Our paper contributes to a discussion of the trustworthiness of empirical claims made by economics researchers (see Christensen and Miguel 2018 for a recent literature review). Using test statistics from three prestigious economics journals,

[3] Our findings are broadly consistent with a growing literature discussing model misspecification for IV regressions (see, for instance, Andrews, Stock, and Sun 2019 for a discussion on weak instruments). Using 1,359 instrumental variables regressions from 31 published studies, Young (2020) show that more than one half of the statistically significant IV results depend on either one or two outlier observations or clusters.

Brodeur et al. (2016) provide evidence that 10 to 20 percent of marginally rejected tests are false-positives. We extend this in several ways by, for example, comparing the Top 5 with other top journals and investigating the role of the review process. Our findings suggest that *p*-hacking is not related to researcher "pedigree." Another important study, Vivalt (2019), investigates the extent of *p*-hacking for a large set of impact evaluations. Vivalt (2019) and Brodeur et al. (2016) both point to *p*-hacking being smaller for RCT than for other methods. We complement these studies by partitioning *p*-hacking for quasi-experimental methods; the most commonly used identification strategies in many social sciences.

We also contribute to a growing literature on transparency (Miguel et al. 2014)[4] and editorial choice (e.g., Card and DellaVigna 2020 and Ellison 2011). To some extent, our findings suggest that improved research design may itself partially constrain *p*-hacking and that RCTs and RDDs appear to have another potential scientific benefit, i.e., beyond improving internal validity they also appear to reduce tendentious reporting. Our results point to the importance of identifying and correcting publication bias (Andrews and Kasy 2019) and that the appropriate correction is sensitive to method. They may also explain divergences documented in meta-analyses in the size and precision of estimates within a given literature (e.g., Havránek and Sokolova 2020).

Section I details data collection. Section II shows the distribution of tests for the whole sample, over time, and by method. We present between-method comparisons in Section III. Section IV explores the role of authors and the review process. Section V concludes.

I. Data Collection

We collect the universe of articles published by 25 top journals in economics during 2015 and 2018. Table 1 provides the complete list of journals. We selected the top journals as ranked using RePEc's Simple Impact Factor excluding any journal that did not publish at least one paper using one of the methods of interest.[5]

In selecting our samples we followed a rule-based exclusion procedure. For each method we began by searching the entire body of published articles for keywords related to that method.[6] These keywords provide four bodies of papers, one for each method.[7] We manually removed articles if they employed a sub-method that alters researcher freedoms. We thus removed papers that use matching (DID) and papers that use instruments as part of a fuzzy RDD, focusing on two stage least squares (IV). We also removed papers using a Structural Equation Model.

[4] See Blanco-Perez and Brodeur (2019) for a survey of editorial policies such as data and code availability policies.

[5] RePEc's 2018 Simple Impact Factor, calculated over the last ten years. This measure uses a citation count and scales it by the number of articles in each journal. Within-journal citations are not included (accessible at https://ideas.repec.org/top/top.journals.simple10.htm).

[6] For DID: "difference-in-difference*," "differences-in-difference*," "difference in difference*," and "differences in difference*." For IV: "instrumental variable*." For RCT: "randomized." For RDD: "regression discontinuity." Where * represents a wild card in the text search, allowing for plurals to be captured with the same search string.

[7] We manually excluded articles that contained the search term—for example in contextual discussion—but did not apply one of the four methods.

Table 1—Summary Statistics

	DID (1)	IV (2)	RCT (3)	RDD (4)	Articles (5)	Tests (6)
American Economic Journal: Applied Economics	12	13	23	4	46	2,242
American Economic Journal: Economic Policy	25	9	5	8	42	1,263
American Economic Journal: Macroeconomics	-	5	-	-	5	54
American Economic Review	21	23	14	3	55	1,740
Econometrica	2	4	1	4	10	307
Economic Policy	2	4	-	-	6	80
Experimental Economics	-	2	4	-	6	79
Journal of Applied Econometrics	-	4	-	1	5	86
Journal of Development Economics	13	25	30	3	64	2,818
Journal of Economic Growth	2	7	-	-	8	100
Journal of Financial Economics	25	16	-	3	40	635
Journal of Financial Intermediation	7	6	-	3	16	285
Journal of Human Resources	4	10	5	3	21	752
Journal of International Economics	7	13	-	1	19	510
Journal of Labor Economics	5	4	8	4	20	653
Journal of Political Economy	4	8	5	2	18	761
Journal of Public Economics	28	18	18	15	74	2,605
Journal of Urban Economics	10	16	-	3	26	660
Journal of the European Economic Association	8	7	6	2	20	491
Review of Financial Studies	25	16	-	7	39	963
The Economic Journal	13	22	1	4	38	891
The Journal of Finance	7	15	5	2	27	1,135
The Quarterly Journal of Economics	5	9	8	6	23	840
The Review of Economic Studies	2	3	2	-	7	306
The Review of Economics and Statistics	14	22	10	7	49	1,484
Total articles	241	281	145	85	684	
Total tests	5,853	5,170	7,569	3,148		21,740

Notes: This table alphabetically presents our sample of Top 25 journals identified using RePEc's Simple Impact Factor: https://ideas.repec.org/top/top.journals.simple10.html. Some top journals did not have any eligible articles in the first data collection period: *Journal of Economic Literature*, *Journal of Economic Perspectives*, *Journal of Monetary Economics*, *Review of Economic Dynamics*, *Annals of Economics and Finance*, and the *Annual Review of Economics*. We also excluded *Brookings Papers on Economic Activity* from the sample. In some research articles, multiple methods were used. This explains why the sum of articles for the four methods is greater than 684.

See online Appendix Table A1 for an example of our data collection, the *American Economic Journal: Applied Economics* for 2015. Ultimately, we collected statistics from 684 articles.

From the included articles, we collected estimates only from results tables. Our goal was to collect only coefficients of interest, or main results, excluding regression controls, constant terms, balance and robustness checks, heterogeneity of effects, and placebo tests. Coefficients drawn from multiple specifications of the same hypothesis were collected. All reported decimal places were collected. For DID, we collected only the main interaction term, unless the non-interacted terms are described by the author(s) as coefficients of interest. For IV, we only collected the coefficient(s) of the instrumented variable(s) presented in the second stage. For RDD, we only collected estimates for the preferred bandwidth. We identify the preferred bandwidth by reading the text where the estimates are described. In case of ambiguity, we chose the optimal bandwidth (Imbens and Kalyanaraman 2012). We also excluded specification checks, such as controlling for third or higher-degree polynomials of the forcing variable. Last, for papers that use more than one method,

we collect estimates from each; e.g., if a paper uses both DID and IV, we collect estimates for both.[8]

Each article was independently coded by two of the three authors. This allowed us to reproduce the work of one another and to make sure we only selected coefficients of interest. Note that we collected the same test statistics for the vast majority of the articles and revisited test statistics for which there was initial disagreement. In the end, we collected the same tests or easily reached agreement for 98.5 percent of collected test statistics.

All of the test statistics in our sample relate to two-tailed tests. Most (91 percent) are reported as coefficients and standard errors, others as *t*-statistics (4 percent), or *p*-values (5 percent). Because degrees of freedom are not always reported, we treat coefficient and standard error ratios as if they follow an asymptotically standard normal distribution. When articles report *t*-statistics or *p*-values, we transform them into equivalent *z*-statistics.

For each article, we also collected information about the authors and their affiliations. A manual search for curriculum vitae allowed us to collect the following information for 96.7 percent of authors (98 percent of test statistics): gender, year and institution of PhD, and whether the author was an editor of an economics journal.

We also revisited articles and test statistics from Brodeur et al. (2016) using the same rule-based exclusion procedure, categorizing articles by method and keeping only coefficients of interest. This results in 17,518 test statistics from 266 articles published in three of the Top 5 journals from 2005 to 2011. This additional data is used to explore *p*-hacking over time beyond our 2015 and 2018 sample.

A. *Descriptive Statistics*

Following the above procedure we collected 21,740 test statistics. On average, there are 24 estimates from each DID article, 18 per IV article, 52 per RCT article, and 37 per RDD article. Including article weights to prevent articles with more tests from having a disproportionate effect has no effect on our main conclusions. Table 1 provides summary statistics. DID, IV, RCT, and RDD respectively contribute 27 percent, 24 percent, 35 percent, and 14 percent of the sample.

Online Appendix Figure 1 illustrates the proportion of articles by method over the time period 2005–2011, 2015, and 2018 for the *American Economic Review*, *Journal of Political Economy*, and the *Quarterly Journal of Economics*. (See online Appendix Figure 2 for Top 25 for 2015 and 2018.) There is a sizable increase in the use of RDDs, with about 5 percent of articles (among those using one of the four methods) using RDD in 2005–2006 rising to 12 percent in 2015 and 2018. In contrast, the share of IV articles decreased from about 50 percent to 40 percent over this period. The share of DID and RCT articles is more stable over time.

Table 2 and online Appendix Table A2 provide descriptive statistics for article and author characteristics. The unit of observation is test statistic in Table 2 and article in online Appendix Table A2. In our sample, the mean academic year of graduation is 2005–2006. A rough categorization of institutions into top and non-top reveals

[8] For field experiments with partial compliance, we add the intention-to-treat estimates to the RCT sample and the IV estimates to the IV sample. Only five studies used both IV and RCT.

Table 2—Article and Author Characteristics

	DID (1)	IV (2)	RCT (3)	RDD (4)	2015 (5)	2018 (6)	Top 5 (7)	Non top 5 (8)	Total (9)
Top 5	0.16	0.15	0.22	0.18	0.19	0.18	1.00	0.00	0.18
	(0.37)	(0.36)	(0.41)	(0.38)	(0.39)	(0.38)	(0.00)	(0.00)	(0.39)
Editor present	0.63	0.60	0.73	0.52	0.66	0.63	0.81	0.61	0.64
	(0.48)	(0.49)	(0.45)	(0.50)	(0.47)	(0.48)	(0.39)	(0.49)	(0.48)
Solo-authored	0.19	0.22	0.12	0.37	0.22	0.18	0.13	0.22	0.20
	(0.39)	(0.41)	(0.33)	(0.48)	(0.42)	(0.38)	(0.34)	(0.41)	(0.40)
Average experience	9.92	10.75	12.28	8.49	10.97	10.47	11.43	10.57	10.73
	(5.07)	(6.26)	(5.86)	(5.20)	(6.10)	(5.48)	(6.57)	(5.62)	(5.82)
Female authors	0.22	0.27	0.38	0.26	0.28	0.31	0.27	0.30	0.29
	(0.30)	(0.37)	(0.32)	(0.35)	(0.33)	(0.35)	(0.32)	(0.34)	(0.34)
Top institutions	0.23	0.31	0.34	0.26	0.33	0.25	0.55	0.23	0.29
	(0.33)	(0.37)	(0.36)	(0.38)	(0.38)	(0.33)	(0.36)	(0.33)	(0.36)
Top PhD institutions	0.36	0.36	0.51	0.28	0.33	0.48	0.55	0.37	0.40
	(0.39)	(0.40)	(0.37)	(0.37)	(0.38)	(0.39)	(0.37)	(0.39)	(0.39)
Test statistics	5,853	5,170	7,569	3,148	11,211	10,529	3,954	17,786	21,740

Notes: Each observation is a test. The Top 5 journals in economics are the *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economic Studies*. Average experience is the mean of years since PhD for an article's authors. Share of female authors, share of authors affiliated with top institutions, and share of authors who completed a PhD at a top institution.

that more almost 30 percent of authors are from a top institution.[9] This increases to 40 percent for the proportion of authors who gained their PhD from a top institution. Last, 20 percent are solo-authored and 71 percent of authors are males.

A decomposition by method reveals that authors working in (or who graduated from) top institutions are disproportionately more likely to use RCT. Authors using RDD earned their PhD relatively more recently and are more likely to be solo-authored. We include in our model author and article characteristics to control for these compositional differences. Last, female authors are more likely to use RCT, and less likely to use DID.[10]

II. Plotting Test Statistics

Figure 1 presents the raw distribution of z-statistics in our sample.[11] Each bar has a width of 0.10 and the interval $z \in [0, 10]$ was chosen to create 100 bins. Reference lines are provided at the conventional two-tailed significance levels. The distribution exhibits a two-humped (or camel) shape: a first hump with low z-statistics and a second hump between 1.65 and 2.5. The distribution exhibits a local minimum around 1.35, suggesting misallocated z-statistics. About 56, 48,

[9] We define "top" for this purpose using the highest rated 20 in RePec's ranking of top institutions at the time of writing (https://ideas.repec.org/top/top.econdept.html). The following 20 institutions are coded as top: Barcelona GSE, Boston University, Brown, Chicago, Columbia, Dartmouth, Harvard, MIT, Northwestern, NYU, Princeton, PSE, TSE, UC Berkeley, UCL, UCSD, UPenn, Stanford, and Yale.

[10] Journal articles published in Top 5 journals were significantly more likely to have been written by authors affiliated with (and to have graduated from) a top institution. In contrast, solo-authorship and experience were not significantly related to the Top 5 status of a journal.

[11] Online Appendix Figure A5 illustrates weighted distribution of tests. The weighting schemes either put equal weight on each article or on each table. The shape of the distribution remains similar to the unweighted distribution.

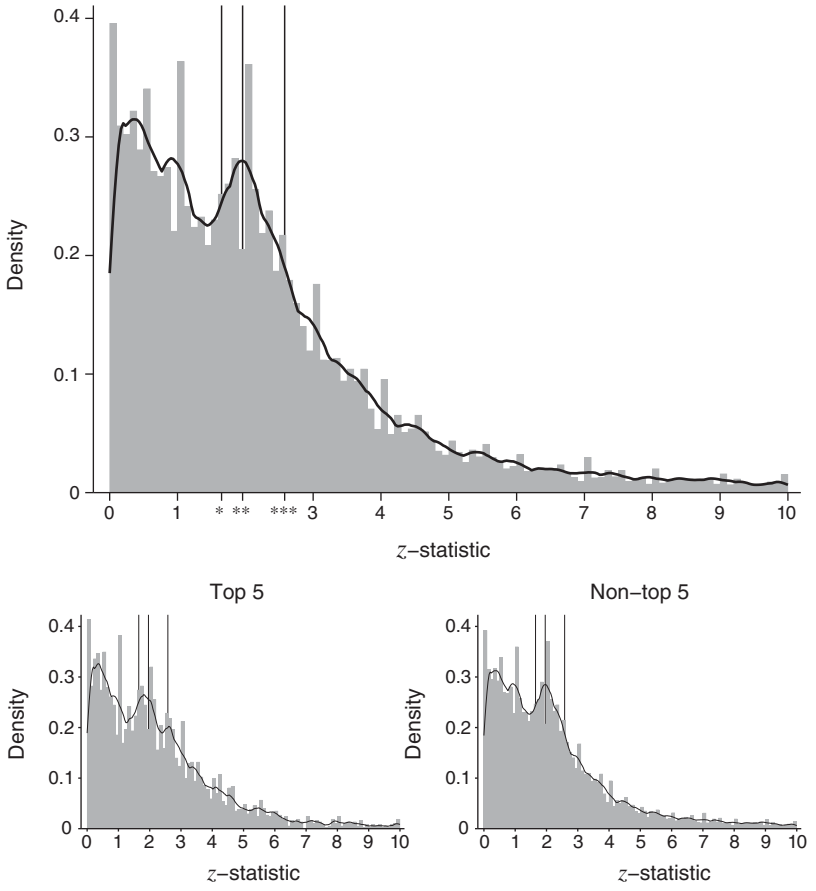

FIGURE 1. z-STATISTICS IN 25 TOP ECONOMICS JOURNALS

Notes: The top panel displays histograms of test statistics for $z \in [0, 10]$. Bins are 0.1 wide. Reference lines are displayed at the conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. The bottom left panel presents test statistics from the Top 5 journals (*American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economic Studies*). The bottom right panel presents test statistics from the remainder of the sample. We do not weight articles.

and 34 percent of test statistics are significant at the 10, 5, and 1 percent levels, respectively. This is consistent with Brodeur et al. (2016) who documented that 54 percent of tests were significant at the 5 percent level in three top economics journals.

A. *Test Statistic Plots by Journal Ranking*

Figure 1 splits the full sample of z-statistics by journal rank. In particular, the left panel restricts the sample to the Top 5,[12] while the right panel shows the distribution of tests for the remaining journals. Both distributions feature a similar

[12] Top 5 journals in economics are *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economic Studies*.

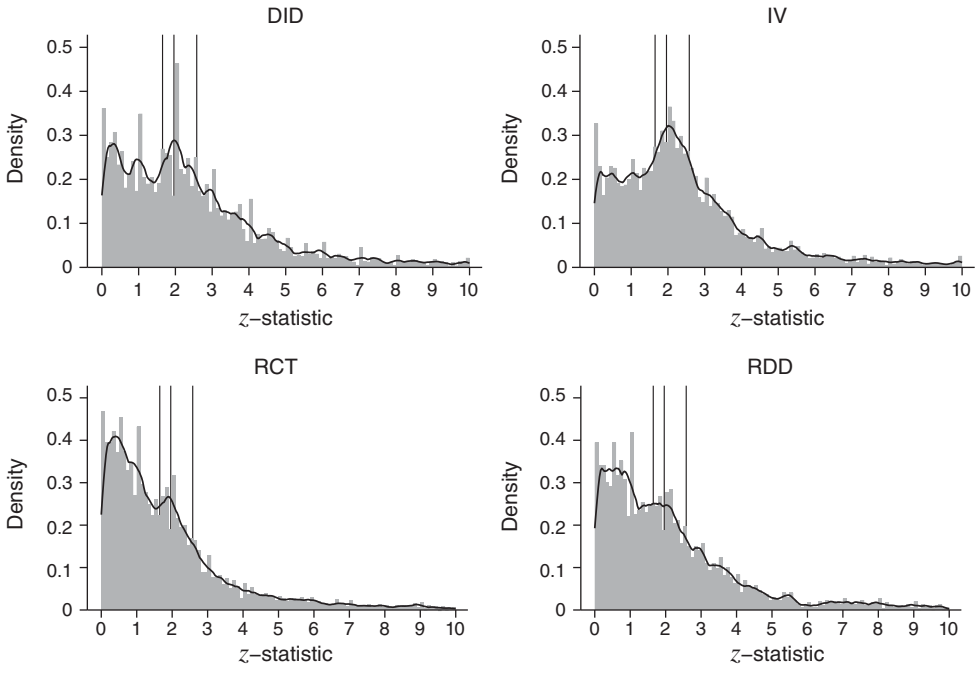

FIGURE 2. z-STATISTICS BY METHOD

Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ by method: difference-in-differences (DID), instrumental variables (IV), randomized control trial (RCT), and regression discontinuity design (RDD). Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We do not weight articles.

two-humped shape. This finding suggests that journal ranking is not related to the extent of *p*-hacking in our sample of top 25 journals. We formalize this in Section III.

B. *Test Statistic Plots by Method*

Figure 2 displays the distribution of z-statistics for each of the four methods. (See online Appendix Figure A3 for the weighted distributions.) We create z-curves by imposing an Epanechnikov kernel density (also of width 0.10). A kernel smooths the distribution, softening both valleys and peaks. In online Appendix Figure A4, we plot the same z-curves into a single panel.

The shapes (and their differences) are striking. The distributions for IV and DID present a global and local maximum around 2 (where a p-value of 0.05 is achieved). DID and IV seem to exhibit a mass shift away from the marginally statistically insignificant interval (just left of $z = 1.65$) into regions conventionally accepted as statistically significant. The extent of misallocation seems to be the highest for IV with a sizable spike and maximum density around 1.96. The distributions for IV and DID are increasing over the interval $[1.5, 2]$.

In stark contrast, RDD presents an almost monotonically falling curve with maximum density close to 0. The distribution for RCT is similar, but also features

a much smaller local maximum near 2. This suggests the extent of misallocated tests in RCT and RDD articles is much more limited than those using IV and DID.

Visual inspection of the patterns suggests two important differences between these two groups of methods. First, looking at the whole of the distributions we can see that many (around one half) of RCT and RDD studies report null results with large *p*-values as their main estimates, whereas IV and DID studies typically reject the null. Second, DID and IV are more likely to report *marginally* significant estimates than RCT and RDD. We confirm this visual analysis using the Kolmogorov-Smirnov test (KS) which confirms that the IV distribution statistically differs from the RCT distribution over the whole interval as well as in the marginally significant interval ($z = [1.65, 1.96]$).

We check whether these patterns are visible for different subsamples. Online Appendix Figures A6–A12 display decompositions by methods and the following characteristics: top 5, number of authors, institution rank, PhD institution rank, years of experience since PhD, editor of an economic journal, and gender, respectively. For these decompositions we offer some observations. The spike at about $z = 2$ is particularly striking for solo-authored RCT and IV studies. There are also many tests with high *p*-values (low *z*-statistics) and virtually no bunching around $z = 2$ for RCTs with at least one author at a top institution (or that graduated from a top institution). RDD articles from authors with greater experience have relatively more tests with high *p*-values and no apparent spike at about $z = 2$. Similarly, RDD articles in top 5 journals have relatively more tests with high *p*-values than those in other journals.

C. *Test Statistic Plots over Time*

It is unclear a priori whether we should expect the extent of *p*-hacking to have changed over time. On one hand, new tools such as pre-analysis plans and data availability policies might have decreased its extent through increased awareness of the issue among reviewers and editors. On the other hand, there is growing evidence that it is increasingly difficult to publish in top journals (Card and DellaVigna 2013) which could have increased the *incentives* to *p*-hack.

Figure 3 (top left) and Figure 3 (top right) split the sample of *z*-statistics by year of publication. Figure 3 (top left) are tests from the years 2005–2011 and 2015 and 2018 (top right) for three top journals, whereas the Figure 3 bottom panels provide a comparison for the years 2015 and 2018 using the top 25 journals. Comparison of the samples from different time periods point to no discernible change over time in either journal group.

We also explore whether the pattern by method documented above was already visible in 2005–2011 in online Appendix Figure A13. We find that the pattern is the same for RCT articles in 2005–2011 as in 2015 and 2018. In contrast, the extent of *p*-hacking appears larger for more recent IV articles. There are not enough sharp RDD studies in the early period to allow for meaningful comparison.

In this regard our findings differ from Vivalt (2019), which studies only development programs. She finds that RCTs have exhibited less *p*-hacking over time (pre- versus post-2010), but not much difference for non-RCT studies.

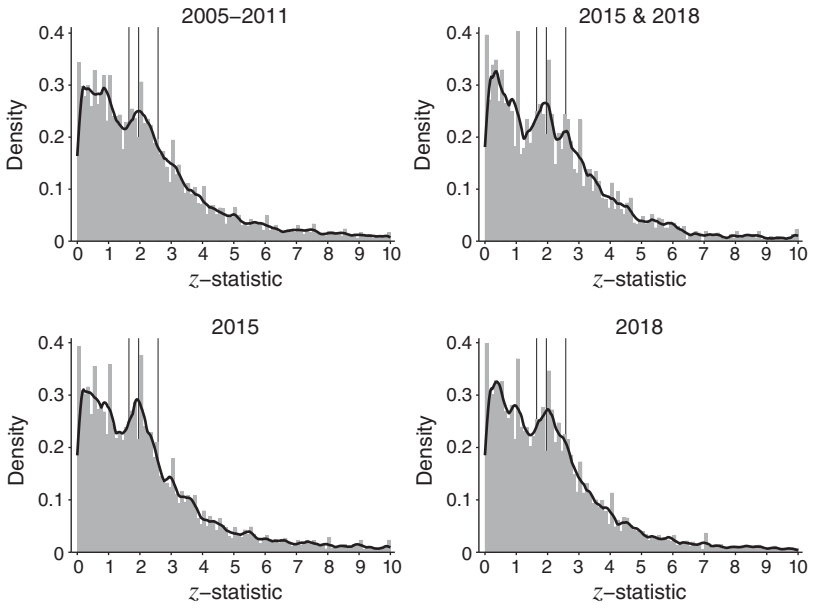

FIGURE 3. z-STATISTICS OVER TIME

Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ over time. The top panels are from the *American Economic Review*, *Journal of Political Economy*, and the *Quarterly Journal of Economics*. The top left panel uses data from Brodeur et al. (2016) and the top right uses the Top 3 journals during our sample period. The bottom left panel is Top 25 journals in 2015 and the bottom right is Top 25 journals in 2018. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We do not weight articles.

III. Further Analysis

To investigate the variations of p-hacking by method, we report three complementary analyses. First, using randomization tests to identify discontinuities in the probability of a test statistic appearing just above or below a statistical threshold. Second, we apply caliper methods to compare test statistics within narrow bands around thresholds. Third, we compare each distribution to its own calibrated counterfactual.

A. *Randomization Tests*

We first rely on what we call randomization tests. The aim of this approach is to confirm the visually obvious discontinuities around the conventional statistical thresholds. We compare whether the mass of test statistics just above, versus just below, the conventional statistical significance thresholds differ significantly by underlying identification method. The benefit of this method is its minimal assumption; in a sufficiently small window, the probability of being just above, versus just below, any threshold should be equal.

Method.—We assume that the underlying distribution of z-statistics (for any research method) is continuous and infinitely differentiable following Andrews and

Kasy (2019). From this assumption, any discontinuity in observed z-statistics must arise from p-hacking or publication bias.[13] We do this by testing if the observed test statistics are binomial-distributed around a threshold with equal probability, as in Andrews and Kasy (2019).

Let N be the number of tests observed for a method in a window of half-width h around the statistical threshold. Further, let k_{obs} be the observed number of successes (significant test statistics) and let $p = 0.5$ be the hypothesized probability of success on a trial. Then the probability of observing the same or greater proportion of significant tests k_{obs} is

$$\Pr(k \geq k_{obs}) = \sum_{m=k_{obs}}^{N} \binom{N}{m} p^m (1-p)^{N-m}. \tag{1}$$

As an added note, publication bias is likely to only work in a single direction (towards significance) as too *many* successes is more indicative of publication bias than too *few*. This makes it appropriate to consider one-sided p-value for our tests.[14] In online Appendix Tables A3, A4, and A5 we account for sampling uncertainty by estimating the proportion of successes p directly. The point estimates are unchanged, and standard errors are very small.

Results.—The results are reported in Table 3 for the 5 percent threshold. In the top panel, we examine a window of half-width $h = 0.5$ around the two-star significance threshold. Here, 1,412 IV test statistics can be found with 53.9 percent statistically significant. In comparison, 1,719 RCT test statistics can be found in the same region with 46.7 percent statistically significant. We then test whether each method is equally likely to be significant and nonsignificant. That is, is the random variable $z_{method} \sim Binomial(p = 0.5)$? The probability of observing 53.9 percent or greater statistically significant IV tests is 0.015. Both DID and IV test statistics have a statistically significant discontinuity in the distribution around the threshold. Greatly reducing the window width in successive panels does not alter this finding.

Interestingly, all methods have a statistically significant discontinuity when the analysis window becomes small enough, even with the reduced sample size. This confirms the earlier visual inspection—even RCTs seem to suffer from some publication bias.

In online Appendix Tables A6 and A7 we use the 10 percent and 1 percent significance thresholds, respectively. (See online Appendix Tables A8–A10 for weighted estimates.) For the 10 percent threshold, we find that regardless of window width IV test statistics are statistically more likely to be "successes," whereas DID test statistics are only more likely to be successful in large windows. RDD test statistics, in almost all cases, are not statistically differently distributed around $z = 1.65$. For the 1 percent significance threshold, we find that no method is ever meaningfully more likely to be successful than chance at this high significance level. This

[13] Bugni and Canay (forthcoming) apply a similar methodology to check for jumps in the density in regression discontinuity settings.

[14] Note that the binomial test is most appropriate when each of the realizations of a random variable are independent. In the online Appendix we repeat the exercise with only one randomly selected test statistic from each table in every article, finding that results are unchanged.

Table 3—Randomization Tests, 5 Percent Significance Threshold

	DID (1)	IV (2)	RCT (3)	RDD (4)
Proportion significant in 1.96 ± 0.5	0.530	0.539	0.467	0.472
One sided p-value	0.015	0.002	0.997	0.939
Number of tests in 1.96 ± 0.5	1,365	1,412	1,719	706
Proportion significant in 1.96 ± 0.4	0.532	0.533	0.479	0.488
One sided p-value	0.016	0.012	0.948	0.733
Number of tests in 1.96 ± 0.4	1,137	1,166	1,416	582
Proportion significant in 1.96 ± 0.3	0.532	0.526	0.485	0.494
One sided p-value	0.030	0.064	0.840	0.611
Number of tests in 1.96 ± 0.3	881	917	1,098	453
Proportion significant in 1.96 ± 0.2	0.556	0.541	0.493	0.508
One sided p-value	0.003	0.022	0.669	0.408
Number of tests in 1.96 ± 0.2	606	619	755	295
Proportion significant in 1.96 ± 0.1	0.631	0.575	0.547	0.542
One sided p-value	0.000	0.005	0.035	0.178
Number of tests in 1.96 ± 0.1	352	315	393	142
Proportion significant in 1.96 ± 0.075	0.684	0.597	0.560	0.565
One sided p-value	0.000	0.002	0.021	0.096
Number of tests in 1.96 ± 0.075	269	238	298	115
Proportion significant in 1.96 ± 0.05	0.707	0.601	0.641	0.614
One sided p-value	0.000	0.005	0.000	0.024
Number of tests in 1.96 ± 0.05	208	168	209	83

Notes: In this table we present the results of binomial proportion tests where a success is defined as a statistically significant observation at the threshold level. In the first panel we use observations where $(1.46 < z < 2.46)$. The other panels use observations for smaller windows. In the first panel, 53.9 percent of the 1,412 IV tests within this window are significant. We then test if this proportion is statistically greater than 0.5. The associated p-values are then reported. We do not weight articles.

is consistent with a reduction in the incentive to p-hack above the arguably more critical two-star threshold.

Comparison to Other Disciplines.—We can compare our randomization test results for the economics literature to those previously conducted on test statistics in political science and sociology. Tests for political science are from Gerber and Malhotra (2008a), while tests for sociology are from Gerber and Malhotra (2008b).[15] Online Appendix Table A11 provides a break down of the number of tests that fell within the range $1.76 < z < 2.16$ for our sample and the non-economics journals. We find that the ratio of tests just above and below 1.96 is only 1.10 in economics in comparison to over 2 for political science and sociology. This result provides strong evidence that the extent of p-hacking is much smaller in economics (at least when using these inference methods) than in other disciplines.[16]

[15] Test statistics are from the *American Political Science Review* and the *American Journal of Political Science* for the time period 1995–2007, and from journal articles published in the *American Sociological Review*, the *American Journal of Sociology*, and the *Sociological Quarterly* for 2003–2005.

[16] Restricting the sample to top 5 outlets or to the sample of top journals in Brodeur et al. (2016) for the years 2005–2011 yield similar conclusions.

B. *Caliper Test*

The caliper test compares the number of estimates in a narrow range above and below a statistical significance threshold. An advantage of this approach over the previous is that this allows us to control for author and article characteristics.

Method.—We estimate the following equation:

$$(2)\quad \Pr\big(Significant_{ij} = 1\big) = \Phi\big(\alpha + \beta_j + X'_{ij}\delta + \gamma DID_{ij} + \lambda IV_{ij} + \phi RDD_{ij}\big),$$

where $Significant_{ij}$ is an indicator variable that test i is statistically significant in journal j for a given threshold. We include journal indicators and report marginal effects of a probit model throughout.[17] Standard errors are clustered at article level.

Challenges to our claim that our approach identifies *p*-hacking, as opposed to publication bias, is that editor and referee preferences for null results may differ by method, or that the extent of *p*-hacking by method could be related to the types of authors that tend to use that method. We tackle these issues by including the term X_{it} in our model. In addition to indicator variables for how results are reported (i.e., whether an article reports *p*-values or *t*-statistics) this vector includes author characteristics. We also include field and journal fixed effects in some models.

A criticism of caliper methods is that bunching near statistical thresholds may reflect prior knowledge about the sample size necessary to obtain a marginally significant estimate.[18] We think it is unlikely a problem here for two reasons. First, it is in RCTs that researchers are most likely to be able to choose their sample size based on power calculations. Second, sample size for articles in our sample is much smaller for RCTs than for the other methods, especially DID. If bunching reflects good priors and power calculations, then the bunching should be most pronounced in the RCT sample, against which the comparisons are made.

Results.—Table 4 presents estimates of equation (2) where the dependent variable indicates whether a test statistic is statistically significant at the 5 percent level. In columns 1–4, we restrict the sample to $z \in [1.46, 2.46]$. Our sample size consists of 5,202 observations. The coefficients presented are increases in the probability of statistical significance relative to the baseline category (RCT). We report standard errors adjusted for clustering by article in parentheses. We also present bootstrapped errors, clustered by article in online Appendix Table A13. We use the inverse of the number of tests presented in the same article to weight observations.[19] This weighting scheme is used to prevent tables with many test statistics to be overweighted.

In the most parsimonious specification, we find that DID and IV estimates are about 10 percentage points more likely to be statistically significant than a RCT estimate. The estimates are statistically significant at the 1 percent level.

[17] Using logit yields similar results, see online Appendix Table A12.

[18] See Ioannidis, Stanley, and Doucouliagos (2017) for an investigation of statistical power and bias in economics. They document that many research areas in economics have nearly 90 percent of their results underpowered.

[19] See online Appendix Table A14 for the unweighted estimates.

Table 4—Caliper Test, Significant at the 5 Percent Level

	(1)	(2)	(3)	(4)	(5)	(6)
DID	0.095	0.088	0.055	0.051	0.052	0.027
	(0.034)	(0.033)	(0.032)	(0.033)	(0.037)	(0.047)
IV	0.102	0.097	0.073	0.080	0.091	0.089
	(0.034)	(0.034)	(0.033)	(0.033)	(0.037)	(0.045)
RDD	0.058	0.057	0.026	0.016	0.025	0.012
	(0.047)	(0.048)	(0.045)	(0.046)	(0.049)	(0.055)
Top 5		−0.051	−0.010			
		(0.045)	(0.084)			
Year = 2018		0.021	0.030	0.024	0.010	0.043
		(0.028)	(0.027)	(0.027)	(0.030)	(0.035)
Experience		−0.002	−0.006	−0.005	−0.006	0.009
		(0.007)	(0.007)	(0.007)	(0.008)	(0.009)
Experience^2		−0.005	0.005	0.006	0.014	−0.028
		(0.018)	(0.018)	(0.019)	(0.020)	(0.025)
Top institution		0.019	0.026	0.025	−0.001	−0.005
		(0.050)	(0.044)	(0.043)	(0.046)	(0.055)
Top PhD institution		−0.011	−0.030	−0.023	0.023	0.067
		(0.039)	(0.037)	(0.038)	(0.040)	(0.048)
Reporting method		Y	Y	Y	Y	Y
Solo-authored		Y	Y	Y	Y	Y
Share female authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Field FE			Y			
Journal FE				Y	Y	Y
Observations	5,202	5,202	5,202	5,202	3,798	2,273
Window	$[1.96 \pm 0.50]$	$[1.96 \pm 0.50]$	$[1.96 \pm 0.50]$	$[1.96 \pm 0.50]$	$[1.96 \pm 0.35]$	$[1.96 \pm 0.20]$
RCT sig rate	0.47	0.47	0.47	0.47	0.48	0.49

Notes: This table reports marginal effects from probit regressions (equation (2)). The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. In columns 1–4, we restrict the sample to $z \in [1.46, 2.46]$. Column 5 restricts the sample to $z \in [1.61, 2.31]$, while columns 6 restricts the sample to $z \in [1.76, 2.16]$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

In contrast, RDD estimates are *not* statistically more likely than RCT estimates to be statistically significant.

One potential explanation for our findings is that authors who are more/less prone to *p*-hacking may select into methods more/less amenable to it. We provide suggestive evidence that this is not the case by enriching our specifications with authors' and articles' characteristics. In column 2, we control for the average years of experience since PhD (and its square), the share of authors at top institutions, the share of female authors, the share of authors who graduated from a top institution, and an indicator for whether at least one of the authors was an editor of an economics journal at the time of publication. We also add dummy variables for Top 5 journals, the year 2018, and for reporting a *t*-statistic or *p*-value instead of the most common coefficient and standard error. The estimates for DID, IV, and RDD remain unchanged.

The estimate for the dummy variable Top 5 is statistically insignificant even at the 20 percent level, suggesting that *p*-hacking is not meaningfully related to the Top 5

status of the journal.[20] Similarly, we find no evidence that the extent of *p*-hacking differs across years.

Some covariates are significantly related to the likelihood the null hypothesis be rejected. For instance, we document a U-shaped relationship in experience. In contrast, the estimates for the share of authors at top institutions and the share who graduated from top institutions are very small and insignificant.[21]

Another potential explanation is that certain types of method are more/less likely to be used in fields where rejection rates are high/low. We explore this by including eight field fixed effects (column 3) and journal fixed effects (column 4). Our IV estimates remain statistically significant at the 5 percent level across specifications and range from 7 to 8 percentage points. In contrast, our DID estimates lose much of their statistical significance at conventional levels and fall to about 5 percentage points. The RDD estimates are very small and statistically insignificant.

In columns 5 and 6, we show that our caliper findings for the 1.96 cutoff are robust to alternative windows: 1.96 ± 0.35 and ± 0.20.[22] IV estimates are about 9 percentage points more likely to be statistically significant than an RCT estimate, and estimates remain significant. The estimates for DID are positive but statistically insignificant.

Online Appendix Tables A16, A17, A18, and A19 replicate Table 4 for the two other common significance thresholds (with and without weights). IV articles remain significantly more likely to report marginally significant tests at the 10 percent level than RCTs. The estimates are all significant and range from 7 to 9 percentage points. RDD estimates are negative, but small and not significantly different to RCT articles. There is no significant differences between DID and the other methods. Last, we do not find evidence of differential bunching by method for the 1 percent significance threshold. Once this very high level of significance is reached, differences between methods become small.

We report several robustness checks in the online Appendix such as excluding papers for which there was initial disagreement between authors in data collection, papers using multiple methods, or excluding subsets of journals based on field or type. Online Appendix Table A22 tackles another potential issue. While we exclude robustness checks and heterogeneity analyses from our sample of tests collected, it is plausible that studies using some methods may be more likely to include tables of results that are either low-power estimates of the effect or with smaller/larger samples. We explore this by restricting the sample to test statistics from the first results table in each article. This exercise decreases our sample to 1,566 test statistics. Nonetheless, our main findings by method are robust.

[20] We also do not find much evidence that the extent of *p*-hacking varies by field. The estimates reported in online Appendix Table A15 suggest that the likelihood to report marginally significant estimates is not significantly different for Top 5, other general interest journals, macroeconomics, development, labor, public, and urban economics than for international trade (the omitted category). The only fields for which there is some evidence of more (less) *p*-hacking is finance (experimental).

[21] We report estimates for the other control variables in online Appendix Table A15.

[22] A potential issue of applying caliper methods in our setting is that each method may have a different underlying distribution. We show our results are robust to increasingly smaller windows which reduces the assumption of distributional equivalence in online Appendix Tables A20 and A21. We display estimates for the following windows: 1.96 ± 0.60, ± 0.50, ± 0.40, ± 0.30, ± 0.20, and ± 0.10. The point estimates for IV are all positive, statistically significant at conventional levels, and range from 7 to 10 percentage points (with our full set of controls and journal fixed effects).

In fact, the size of the estimates for IV is now much larger, ranging from 15 to 20 percentage points in comparison to RCT. Our estimates for DID are insignificant and range from 6 to 9 percentage points.

C. *Excess Test Statistics*

Third, we compare each observed distribution of test statistics to a counterfactual distribution. This requires additional assumptions about what the observed distribution would look like absent publication bias or *p*-hacking. A counterfactual distribution allows us to examine the absolute *level* of publication bias, expanding on the previous results which have been constrained to be relative to RCTs. We expand on the framework introduced in Brodeur et al. (2016), who hypothesized that the underlying distribution of test statistics follows a *t*-distribution with 1 degree of freedom. Here, we make the same distributional assumption but relax it by flexibly calibrating a different counterfactual *t*-distribution to each method, endogenizing the potential differences between methods that would affect its shape, location, and scale.

Method.—This exercise is meant to determine the location and extent of excess test statistics. The challenge is to define an appropriate counterfactual distribution, what should be observed in the absence of publication bias or *p*-hacking. Here, we formalize Brodeur et al.'s (2016) methodology by calibrating a non-central input distribution by method. We assume that the observed test statistic distribution above $z = 5$ should be free of *p*-hacking or publication bias—the incentives to *p*-hack in a range so far above the traditional significance thresholds are plausibly zero. We then produce a non-central *t*-distribution for each method that closely fits the observed distribution in the range $z > 5$ by calibrating the degrees of freedom and non-centrality parameter. Note that while the degrees of freedom parameter is defined over real values, we focus only on positive integers. As the degrees of freedom increase, the tail of the *t*-distribution becomes thinner. We optimize in steps of 1. The non-centrality parameter of the *t*-distribution is positive and real valued. We optimize in steps of 0.01. Increasing the non-centrality parameter in our case makes the distribution's tail thicker (since we take the absolute function of the test statistics earlier in the process.)

This presents us with an optimization problem with countervailing forces. Our approach is the following. For 0 to 10 degrees of freedom, we calculate the non-centrality parameter that minimizes the difference in the $z > 5$ mass of the observed distribution and the expected distribution. We then choose the "best" of the 10 optimized *t*-distributions by degree of freedom. In this manner we explore the entire region of $0 < df < 10$ and $0 < np < 4$.

Figure 4 presents the calibrated input distributions with the observed distributions. Our formalization yields very precise fitting curves. For the distribution of DID test statistics which has 15.2 percent of its mass in the tail, our algorithm produces a *t*-distribution with a mass of 15.1 percent in its tail, choosing 2 degrees of freedom and a non-centrality parameter of 1.81. The remaining methods also optimize at 2 degrees of freedom, the optimal non-centrality parameter varies across methods.

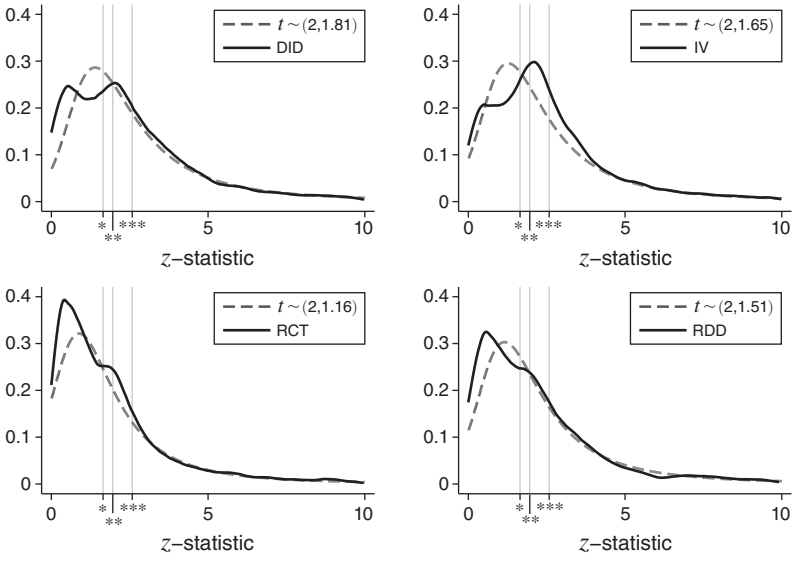

FIGURE 4. EXCESS TEST STATISTICS BY METHOD

Notes: This figure presents the calibrated input distributions with the observed distributions. We optimize for each method a student *t*-distribution with 2 degrees of freedom. The optimal non-centrality parameter varies across methods. See Section IIIC for more details.

To calculate the excess test statistics in a particular region, we use the CDF of observed *t*-statistics $\hat{F}(upper) - \hat{F}(lower)$ and from this subtract $F_{t(2,1.81)}(upper) - F_{t(2,1.81)}(lower)$. In this way we calculate the excess mass of test statistics as the difference between the mass observed and the mass expected, given that our expectations are calibrated only by information contained in the tail.

Another approach to measuring excess test statistics is to compare each method's excess masses to the excesses of a common baseline—in our case RCT. Making this comparison additionally assumes homogeneous effect distributions and that *p*-hacking and publication bias similarly distort test statistics across methods. The conclusions using this approach are similar, and are presented in online Appendix Table A23.[23]

Results.—In Figure 4 we present the observed and calibrated *t*-distributions (for table form see online Appendix Table A24). We first remark that our tail fitting has succeeded visually. For each method, the calibrated *t* closely matches the observed distribution in $[5 < z < \infty)$. This is confirmed in online Appendix Table A24, where the difference in mass between calibrated and observed is at most 0.001 in $[5 < z < \infty)$.

[23] Another approach would be to use maximum likelihood to fit a (non-central or even generalized) *t*-distribution to the tail of each method's observed distribution. The entire observed distribution is then compared to the fitted *t*-distribution. The results of this maximum likelihood exercise are presented in online Appendix Figure A14. There we present results using both information from the tail of $z > 5$, a less stringent tail of $z > 3$, and the inclusion of the *t*-distribution's scale parameter. Our conclusions remain unchanged as this approach generates curves similar to those in Figure 4 and generally larger estimates of publication bias.

For the $[0 < z < 1.65)$ region, the mass difference between expected and observed is small with the exception of IV, which has a dearth equal to 6.3 percent of its total mass. Compared to the expected IV mass, approximately 16 percent of insignificant IV test statistics are missing.

For the $[1.65 < z < 1.96)$ region, the mass difference between expected and observed is small for every method (although IV is the only method with excess mass).

The most striking result comes from the $[1.96 < z < 2.58)$ region, where IV has an excess of 4.1 percent of its total mass, or *30 percent more statistically significant test statistics than expected*. The size of the IV excess is more than 5 times as large as the excess for DID and RCT. DID and RCT both exhibit a degree of distortion, each having 0.8 percent too much mass (6.0 percent and 6.5 percent more significant test statistics than expected in this region) respectively. RDD performs well, consistently having *less* statistically significant test statistics than expected.

For the $[2.58 < z < 5)$ region, IV has an excess total mass of 1.9 percent, or 7.8 percent too many statistically significant test statistics. The remaining methods have too few; DID has 3.7 percent, RDD has 10.4 percent, and RCT has 15.3 percent less statistically significant test statistics than expected.

Comparison with RCT.—In order to benchmark the size of our results, we apply a similar approach in online Appendix Table A23 which uses the observed RCT distribution in place of the calibrated *t*-distributions. We take the mass of test statistics observed (e.g., $\hat{F}_{IV}(2.58) - \hat{F}_{IV}(1.96)$) and subtract the RCT mass ($\hat{F}_{RCT}(2.58) - \hat{F}_{RCT}(1.96)$). While relaxing the assumption that the underlying tests are *t*-distributed, this approach no longer endogenizes method differences in how test statistics are treated by researchers or reviewers.

The results are similar. The first panel examines the statistically insignificant region. DID and IV have too few insignificant test statistics, each dearth more than double that of RDD. In the just-significant region, there is very little difference between the quasi-experimental and RCT distributions (although IV is the only method with excess mass). In the two-star significance region, we estimate that 5.4 percent of all IV estimates are misallocated, or that 43.3 percent of two star IV results should instead be found in the insignificant region (the only region with too little mass). The estimate for IV is twice that of DID and eleven times that of RDD. The weakness of this simpler approach becomes apparent in the $[2.58 < z < 5)$ range, where all methods are considered to have far "too many" significant results. This is due to the implicit assumption that effect sizes are homogeneous between literatures. For this reason we favor our calibrated input distribution approach.

D. *Estimating the Amount of Distortion*

Our setting is well suited to applying the Andrews and Kasy (2019) measurement of publication bias. Recall that publication bias is present when the probability a result is published is a function of its statistical significance. This measurement makes distributional assumptions for the sample's *effect sizes* and assumes effect size estimates with smaller standard errors do not relate to different estimates. The measure of publication bias is the relative publication probability of a statistically

TABLE 5—RELATIVE PUBLICATION PROBABILITIES

	DID (1)	IV (2)	RCT (3)	RDD (4)
Panel A				
$\beta_{[0<Z<1.96]}$	0.237	0.214	0.522	0.355
	(0.010)	(0.010)	(0.024)	(0.020)
Location	0.006	0.021	0.020	0.004
	(0.001)	(0.003)	(0.002)	(0.001)
Scale	0.004	0.013	0.013	0.002
	(0.000)	(0.002)	(0.001)	(0.000)
Degrees of freedom	2.249	2.464	2.335	2.100
	(0.010)	(0.060)	(0.051)	(0.080)
Panel B				
$\beta_{[0<Z<1.65]}$	0.181	0.159	0.493	0.301
	(0.009)	(0.008)	(0.029)	(0.021)
$\beta_{[1.65<Z<1.96]}$	0.465	0.559	0.835	0.660
	(0.028)	(0.034)	(0.051)	(0.057)
$\beta_{[1.96<2.58]}$	0.732	0.834	1.079	0.863
	(0.039)	(0.046)	(0.062)	(0.070)
Location	0.006	0.018	0.019	0.003
	(0.001)	(0.003)	(0.002)	(0.001)
Scale	0.003	0.011	0.012	0.002
	(0.001)	(0.002)	(0.002)	(0.000)
Degrees of freedom	2.408	2.589	2.329	2.193
	(0.050)	(0.063)	(0.053)	(0.095)

Notes: In panel A, $\beta_{[0<Z<1.96]}$ is the relative publication probability of a statistically insignificant test. For example, if a statistically significant IV test has a 50 percent chance of being published, then a statistically insignificant one has a $50\% \times 21.4\% = 10.7\%$ chance of being published. Panel B presents the relative publication probability of statistical significance regions as compared to the most significant test statistics ($Z > 2.58$). The table presents the results of applying the publication bias model presented in Andrews and Kasy (2019). The model assumes that the underlying effect sizes follow a generalized *t*-distribution, as elsewhere in this manuscript. We report the fitted location and scale parameters, as well as the degrees of freedom.

significant result compared to a statistically insignificant result. If a significant result is just as likely as an insignificant result to be published, publication bias must be low.

The measurement involves applying a step function at significance thresholds to the conditional probability of publication. The results are presented in Table 5. For ease of exposition we begin by comparing results that are insignificant at the 5 percent level to results that are significant at the 5 percent level. In the IV literature, a result that is statistically insignificant is only 21.4 percent as likely to be published as a significant one. Said differently, a significant IV result is almost 5 *times* more likely to be published than an insignificant IV result. In the DID literature, a statistically significant result is 4.2 times more likely to be published. For RCTs, a significant result is only 1.9 times more likely to be published. For RDDs, a significant result is 2.8 times more likely to be published. All of these estimates are statistically significant at the 1 percent level. We have also presented the generalized *t* distribution parameters the model fits for the underlying effect distribution. Reassuringly, the estimates are similar to those in previous sections.

When we apply the model which differentiates between test statistics at the one, two and three star significance levels, a similar pattern emerges. Most notably, a stark difference between statistically insignificant and significant (at any level) results. For DID, a result statistically significant at the 10 percent, 5 percent, and 1 percent level is 2.6, 4.0, and 5.5 times more likely to be published than an insignificant result, respectively. For IV those multiples are 3.5, 5.2, and 6.3 times more likely to be published than an insignificant result, respectively. RDD is less stark, at 2.2, 2.9, and 3.3. RCT behaves uniquely and arguably the best as the publication probability step between insignificant and significant results is reduced substantially. We find that an RCT result statistically significant at the 10 percent, 5 percent, and 1 percent level is 1.7, 2.2, and 2.0 times more likely to be published than an insignificant result, respectively.

IV. Exploring Channels

We now turn to possible channels through which the different methods might produce differing patterns of test statistics in the published literature.

A. *Instrumental Variables: F-statistics*

For non-experimental methods (like IV) there are many stages in the research process when researchers exercise discretion. This is in contrast to RCTs where there are fewer researcher degrees of freedom (and where pre-registration is more likely to be expected).[24] We can use the first stage estimates reported in IV studies to probe, in a different part of the analysis, researcher responses to conventional cutoffs. More concretely, we document the distribution of *F*-statistics for IV articles in our sample. The first stage *F*-statistic is typically used in IV papers to test if an instrumental variable is weak; if its correlation with the endogenous regressor is low.[25]

Interestingly, *F*-statistics were reported in only two-thirds of IV papers in our sample. On average, there were 10 *F*-statistics (standard deviation of 11) per paper.

Figure 5 (top panel) shows the distribution of *F*-statistics reported in specifications over the interval $[0, 50]$. (See online Appendix Figure A15 for $F \in [0, 100]$.) Our sample includes 2,175 *F*-statistics, of which about 12 percent are smaller than 10. This result is in line with Andrews, Stock, and Sun (2019), who find that weak instruments are frequently encountered and that virtually all published papers in their sample (17 papers published in the *American Economic Review*) reported at least one such first-stage *F*-statistic. We are interested in whether there is bunching

[24] We also investigate whether it is easier to manipulate *p*-values when there is not an event-study graph in DID articles. It is arguably harder to convince referees and editors that a policy has a statistically significant impact when the raw data suggest otherwise. Online Appendix Figure A16 directly compares the distribution of test statistics for DID articles without and with an event-study graph, whereas online Appendix Table A26 shows caliper tests for the 5 percent significance level. In our sample, about three-quarter of DID articles have such a graph. Our estimates suggest that DID articles with an event-study graph are not significantly more likely to reject the null hypothesis than the other DID studies.

[25] Many studies in our sample mentioned Stock and Watson's recommendation (or, more generally, the problem of weak instruments) that first-stage *F*-statistic(s) should be larger than 10. This suggests that authors are aware of and use this threshold.

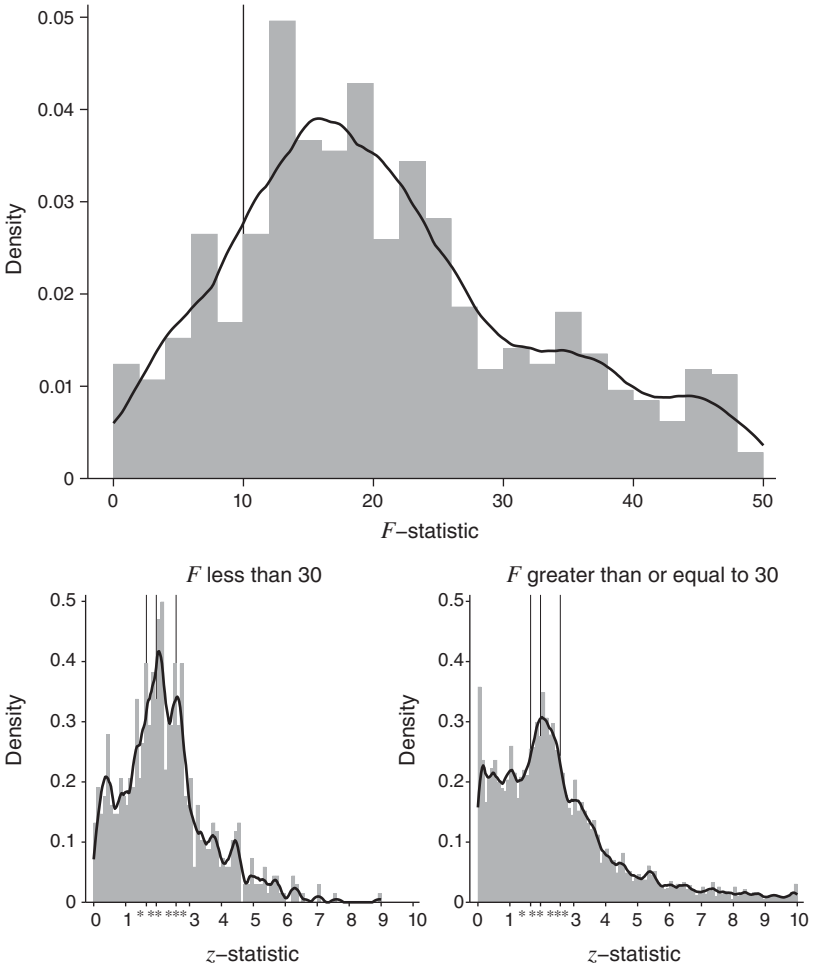

FIGURE 5. INSTRUMENTAL VARIABLES: FIRST STAGE *F*-STATISTICS AND ASSOCIATED *z*-STATISTICS

Notes: This figure displays a histogram of first stage *F*-statistics of instrumental variables for $F \in [0,50]$. This is the raw distribution. Bins have a width of 2. A reference line is provided at the standard "weak" instrument threshold of 10. The bottom left panel displays the distribution of test statistics for IVs with a relatively low *F*-statistic (below median), while the bottom right panel displays the distribution of tests for IVs with a relatively high *F*-statistic (above median). The median *F*-statistic in our sample is just over 30. Because not all IV statistics have an associated *F*-statistic, a total of 1,414 statistics are used in this analysis. The bottom left panel contains 681 tests, while the bottom right contains 733 tests.

at 10 and find that the distribution has a maximum density near to but above 10 and that approximately 52 percent are in the interval $[10,50]$. There is a sizable under-representation of weak instruments relatively to *F*-statistics just over the threshold of 10 but also to (very) large *F*-statistics.

Online Appendix Table A25 formally tests for discontinuities using randomization tests. In this table, we present the results of binomial proportion tests where a success is defined as a first stage *F*-statistic above 10. Reported *p*-values are the probability of the observed (or greater) proportion given a hypothesized equal probability of being just above and below the threshold. There is a statistically significant difference in the proportion around 10 using windows as small as $5 < F < 15$.

Acknowledging that we can expect the proportion of tests between $0 < F < 10$ and $10 < F < 35$ to be very different due to the sheer width differences in the interval, we prefer the results from randomization tests using widths of 10 and smaller.

Overall, the results indicate that both the first and second stages of IV studies display an excess of marginally significant test statistics. We then check whether the degree of *p*-hacking in the second stage is related to the strength of the first stage. Figure 5 (bottom panels) shows that second stage results from comparatively "weak" instruments have a much higher proportion of *z*-statistics centered around conventional thresholds, suggesting that the weaker the IV, the greater the extent of *p*-hacking.

We also find evidence that IV results in RCT studies with partial compliance display a markedly smaller degree of *p*-hacking than IV in purely observational studies (online Appendix Figure A17). This points us to suspect that the reception of IV—rather than the methodology itself—is generating this distinctive curve.[26] See the online Appendix for additional results and discussion on IV.

B. *Role of the Journal Review Process*

We now explore the role of journal editors and referees, and test whether the editorial process exacerbates or attenuates *p*-hacking. To do this, we compare the distribution of test statistics in published journal articles to that in the antecedent working papers. Arguably, observed differences between the working paper version and the published version can be thought to capture the direct impact of the review process.

We proceed as follows. First, we collected all working papers of the articles in our sample. Second, we kept only working papers released before the date of submission to the journal. Unfortunately, the date of submission was not available for 11 journals and for those we kept only working papers released at least 2 years prior to publication. We managed to obtain at least one (valid) working paper for 279 articles (41 percent of sample).[27] Third, for journal articles with multiple (valid) working papers, we chose that closest to the date of submission (or the two-year threshold), with a preference for CEPR, IZA, or NBER discussion/working papers when multiple working papers have similar dates. For a paper published in 2015, for example, it likely means it was first a working paper in 2012 or 2013, given editorial delays.

We then collect test statistics in the working papers using the same methodology as for the published version. For some papers, tables were added or removed, or test statistics have different *p*-values, e.g., by having a different clustering technique.

Figure 6 compares the distribution of test statistics in the working paper versus published version. Panel A is restricted to the published papers with a working paper, whose test statistics are presented in panel B. (See online Appendix Figure A19 for the unbalanced comparison between all working papers and all journal articles.)

[26] We also present a related exercise in which we compare IV test statistics in RCT papers to RCT test statistics. Admittedly this is an unbalanced sample—many RCT studies do not report IV estimates to cope with partial compliance. Online Appendix Figure A18 shows that the distribution of test statistics is quite similar in these two subsamples.

[27] We were significantly more successful at finding working papers for RCT articles than for the other methods. See online Appendix Table A27 for details.

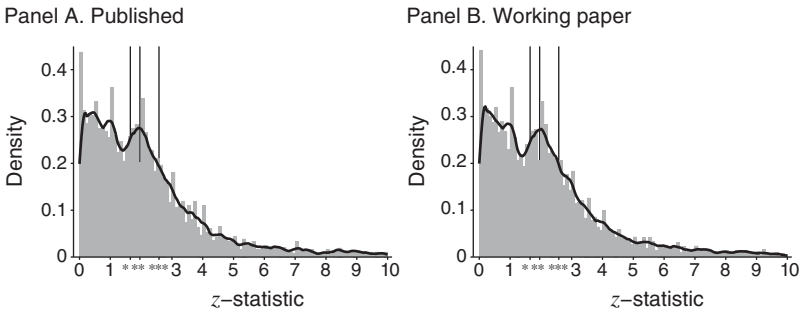

FIGURE 6. HISTOGRAM BY PUBLICATION STATUS—BALANCED SAMPLE

Notes: This figure displays histograms of test statistics for $z \in [0, 10]$. Panel A restricts the sample to journal articles. Panel B restricts the sample to working papers. For the published version, the sample is restricted to journal articles for which we could find a working paper. Bins are 0.1 wide. Reference lines are displayed at the conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We do not weight articles.

The distributions are strikingly similar with a two-humped shape, suggesting that conditional on an article being published, the editorial process has little impact on the extent of *p*-hacking.

Online Appendix Figure A20 repeats this exercise but for each method separately. We do not find much evidence that the distribution of tests differs from the working paper and published version for any of the four methods.

We formalize this analysis in online Appendix Table A28 which reports caliper tests where the dependent variable indicates whether a test statistic is statistically significant at the 5 percent level. The sample is 4,305 tests from working papers and their subsequent published version. The variable of interest is a dummy for whether a test comes from the working paper or the published version. We include fixed effects for each paper in our model, estimating *within* article changes to significance. In column 1 the estimated effect of the publication process is very small, negative, and statistically insignificant. This leads us to believe the editorial process does not change the extent of *p*-hacking. Columns 2–5 restrict the sample to DID, IV, RCT, and RDD articles. Results are similar in each case.[28]

V. Conclusion

The credibility revolution in empirical economics has promoted causal identification using experimental (RCT) and natural-experimental methods (IV, DID, and RDD) (Angrist and Pischke 2010). The associated change in the focus of empirical economics towards explicit causal inference is arguably the most important re-orientation in the discipline of the past two decades. Such design-based research methods deliver many well documented benefits. They may also bring

[28] Online Appendix Figure A21 presents histograms of test statistics in working papers by method and journal ranking (i.e., Top 5 and non-Top 5). Online Appendix Figure A22 presents the same for subsequent published versions. The figures are strikingly similar, confirming that the editorial process appears to not change the extent of *p*-hacking by method or at Top 5 and other top outlets.

opportunities for questionable research practices (of the sort that have collectively come to be known as *p*-hacking) and be differently subject to publication bias.

The primary aim of this study is to investigate the extent of the *p*-hacking and publication bias problems both in aggregate and by method. Our analysis points to significant between-method differences, with papers using IV and DID identified as particularly problematic. We believe this to be roughly consistent with an unspoken hierarchy in the profession, which typically regards the RCT as gold standard and IV with skepticism.

Our secondary results find no discernible difference between papers published in the Top 5 compared to those in other leading economics journals. The *p*-hacking or publication bias pattern also appears common across author characteristics (with the exception of experience). Comparing the published version with an antecedent working paper provides little evidence of mitigation by the peer-review process. Despite recent awareness to the issues of *p*-hacking and publication bias in economics, we find little evidence of a change over time. Last, we find that the extent of *p*-hacking in economics is much smaller than in other social sciences.

Several limitations and caveats of this study are worth discussing. First, our analysis does not indicate that individual researchers or reviewers are acting "dishonestly" or without integrity, and we do not use the terms *p*-hacking or publication bias in an individually pejorative way. Research, often involving a team of contributors, evolves via a sequence of decisions over a period of months or even years. The set of conscious and unconscious biases that could lead to the patterns that we see in the overall published record is not something to which we speak directly. Instead, our results suggest that, *taken as a body*, those papers that report results based on the IV method for example, appear less "trustworthy" than results based on other methods.

Second, the test statistics in our sample come from papers published in excellent general and top field journals. As such, our results document what is happening at the "top end" of publishing in the profession, and casts no light on the greater literature. It may be that marginally insignificant results from IV and DID-based research find homes at journals of lower rank, such that a sufficiently broad reading of a literature reduces the possibility that a reader might be misled by the issues identified here.

Third, the results do not necessarily point to flaws inherent in the methods themselves, but rather the way in which they are collectively executed by researchers and received by reviewers in leading journals. In terms of future solutions this is a potentially important distinction, implying that improved publication practices may eventually mitigate the problem. From the point of view of the research consumer who is interested in knowing to what extent they should be skeptical about the published literature of a topic, the distinction is less important.

Finally, while the published literature may have embedded *p*-hacking and publication bias, it still delivers valuable insights. We suggest only that a nuanced reading of research should account for the underlying method's proclivity to statistical significance. The recent progress in research transparency in the forms of data availability, pre-registrations, pre-analysis plans, and the declared openness to publishing null results may serve to meaningfully mitigate these problems.

REFERENCES

Abadie, Alberto. 2020. "Statistical Nonsignificance in Empirical Economics." *American Economic Review*: *Insights* 2 (2): 193–208.

Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review* 109 (8): 2766–94.

Andrews, Isaiah, James H. Stock, and Liyang Sun. 2019. "Weak Instruments in Instrumental Variables Regression: Theory and Practice." *Annual Review of Economics* 11: 727–53.

Angrist, Joshua D., and Jorn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30.

Ashenfelter, Orley, Colm Harmon, and Hessel Oosterbeek. 1999. "A Review of Estimates of the Schooling/Earnings Relationship, with Tests for Publication Bias." *Labour Economics* 6 (4): 453–70.

Biddle, Jeff E., and Daniel S. Hamermesh. 2017. "Theory and Measurement: Emergence, Consolidation, and Erosion of a Consensus." History of Political Economy 49: 34–57

Blanco-Perez, Cristina, and Abel Brodeur. 2019. "Transparency in Empirical Economic Research." *IZA World of Labor 2019*: 467.

Blanco-Perez, Cristina, and Abel Brodeur. 2020. "Publication Bias and Editorial Statement on Negative Findings." *Economic Journal* 130 (629): 1226–47.

Brodeur, Abel, Nikolai Cook, Anthony Heyes. 2020. "Replication Data for: Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E120246V1.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics* 8 (1): 1–32.

Bruns, Stephan B., Igor Asanov, Rasmus Bode, Melanie Dunger, Christoph Funk, and Sherif M. Hassan. 2019. "Reporting Errors and Biases in Published Empirical Findings: Evidence from Innovation Research." *Research Policy* 48 (9): 103796.

Bugni, Federico A., and Ivan A. Canay. Forthcoming. "Testing Continuity of a Density Via g-Order Statistics in the Regression Discontinuity Design." *Journal of Econometrics.*

Card, David, and Stefano DellaVigna. 2013. "Nine Facts about Top Journals in Economics." *Journal of Economic Literature* 51 (1): 144–61.

Card, David, and Stefano DellaVigna. 2020. "What Do Editors Maximize? Evidence from Four Economics Journals." *Review of Economics and Statistics* 102 (1): 195–217

Casey, Katherine, Rachel Glennerster, and Edward Miguel. 2012. "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan." *Quarterly Journal of Economics* 127 (4): 1755–812.

Christensen, Garret, and Edward Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature* 56 (3): 920–80.

Deaton, Angus, and Nancy Cartwright. 2018. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science & Medicine* 210: 2–21.

De Long, J. Bradford, and Kevin Lang. 1992. "Are All Economic Hypotheses False?" *Journal of Political Economy* 100 (6): 1257–72.

Doucouliagos, Chris, and T. D. Stanley. 2013. "Are All Economic Facts Greatly Exaggerated? Theory Competition and Selectivity." *Journal of Economic Surveys* 27 (2): 316–39.

Ellison, Glenn. 2011. "Is Peer Review in Decline?" *Economic Inquiry* 49 (3): 635–57.

Furukawa, Chishio. 2019. "Publication Bias under Aggregation Frictions: From Communication Model to New Correction Method." Unpublished.

Gerber, Alan, and Neil Malhotra. 2008a. "Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals." *Quarterly Journal of Political Science* 3 (3): 313–26.

Gerber, Alan S., and Neil Malhotra. 2008b. "Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?" *Sociological Methods & Research* 37 (1): 3–30.

Havránek, Tomáš. 2015. "Measuring Intertemporal Substitution: The Importance of Method Choices and Selective Reporting." *Journal of the European Economic Association* 13 (6): 1180–204.

Havránek, Tomáš, Zuzana Irsova, and Tomas Vlach. 2018. "Measuring the Income Elasticity of Water Demand: The Importance of Publication and Endogeneity Biases." *Land Economics* 94 (2): 259–83.

▶ **Havránek, Tomáš, and Anna Sokolova.** 2020. "Do Consumers Really Follow a Rule of Thumb? Three Thousand Estimates from 144 Studies Say 'Probably Not.'" *Review of Economic Dynamics* 35: 97–122.

▶ **Henry, Emeric.** 2009. "Strategic Disclosure of Research Results: The Cost of Proving Your Honesty." *Economic Journal* 119 (539): 1036–64.

▶ **Imbens, Guido.** 2010. "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature* 48 (2): 399–423.

▶ **Imbens, Guido, and Karthik Kalyanaraman.** 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *Review of Economic Studies* 79 (3): 933–59.

▶ **Ioannidis, John P. A.** 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8): e124.

▶ **Ioannidis, John P. A., T. D. Stanley, and Hristos Doucouliagos.** 2017. "The Power of Bias in Economics Research." *Economic Journal* 127 (605): F236-65.

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73 (1): 31–43.

▶ **Leamer, Edward E., and Herman B. Leonard.** 1983. "Reporting the Fragility of Regression Estimates." *Review of Economics and Statistics* 65 (2): 306–17.

Lenz, Gabriel, and Alexander Sahn. Forthcoming. "Achieving Statistical Significance with Control Variables and without Transparency." *Political Analysis*.

McCloskey, Donald N. 1985. "The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests." *American Economic Review* 75 (2): 201–205.

▶ **Miguel, E., C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, et al.** 2014. "Promoting Transparency in Social Science Research." *Science* 343 (6166): 30–31.

▶ **Panhans, Matthew T., and John D. Singleton.** 2017. "The Empirical Economist's Toolkit: From Models to Methods." *History of Political Economy* 49: 127–57.

Ravallion, Martin. 2018. "Should the Randomistas (Continue to) Rule?" Center for Global Development Working Paper 492.

▶ **Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn.** 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66.

▶ **Stanley, T. D.** 2005. "Beyond Publication Bias." *Journal of Economic Surveys* 19 (3): 309–45.

Stanley, T. D. 2008. "Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection." *Oxford Bulletin of Economics and Statistics* 70 (1): 103–27.

▶ **Vivalt, Eva.** 2019. "Specification Searching and Significance Inflation across Time, Methods and Disciplines." *Oxford Bulletin of Economics and Statistics* 81 (4): 797–816.

Young, Alwyn. 2020. "Consistency Without Inference: Instrumental Variables in Practical Application." Unpublished.