American Economic Review 2022, 112(9): 3124–3136
https://doi.org/10.1257/aer.20210121

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

By Sebastian Kranz and Peter Pütz*

Brodeur, Cook, and Heyes (2020) study hypothesis tests from economic articles and find evidence for p-hacking and publication bias, in particular for instrumental variable and difference-in-difference studies. When adjusting for rounding errors (introducing a novel method), statistical evidence for p-hacking from randomization tests and caliper tests at the 5 percent significance threshold vanishes for difference-in-differnce studies but remains for instrumental variable studies. Results at the 1 percent and 10 percent significance thresholds remain largely similar. In addition, Brodeur, Cook, and Heyes derive latent distributions of z-statistics absent publication bias using two different approaches. We establish for each approach a result that challenges its applicability. (*JEL* A14, C12, C52)

Abel Brodeur, Nikolai Cook, and Anthony Heyes (2020)—henceforth, BCH—collected a large, very insightful dataset of hypothesis tests from 25 economic journals. They compare articles that employ different empirical strategies to estimate causal effects and find evidence for *p*-hacking and publication bias overall and in particular for results relying on instrumental variables (IV) and to a smaller extent for difference-in-differences (DID) estimates.

We show that rounding errors in the reported coefficients and standard errors cause bunching of computed *z*-statistics, in particular at exactly $z = 2$. When adjusting for this rounding problem, evidence for *p*-hacking in the dataset substantially weakens. Replicating BCH's randomization and caliper tests at the 5 percent significance threshold, evidence only remains for IV. That being said, our adjustment has only a small impact at the 10 percent significance threshold and on the kernel density estimates of *z*-statistics; BCH's corresponding insights are not substantially changed.

Finally, we note two issues unrelated to rounding. BCH use two different approaches to recover the latent distribution of *z*-statistics absent publication bias and *p*-hacking. We prove that the first approach, based on matching probability

* Kranz: Ulm University, Department of Mathematics and Economics (email: sebastian.kranz@uni-ulm.de); Pütz: Bielefeld University, Faculty of Business Administration and Economics (email: peter.puetz@uni-bielefeld.de). Isaiah Andrews was the coeditor for this article. We thank two anonymous referees for very helpful comments and suggestions. We also thank the original authors, Abel Brodeur, Nikolai Cook, and Anthony Heyes, for helping to clarify important points and for providing an updated data set. We also thank Maike Hohberg for helpful comments.

† Go to https://doi.org/10.1257/aer.20210121 to visit the article page for additional materials and author disclosure statements.

masses in the tails, generally fails to recover the true latent distribution. For the second approach, based on Andrews and Kasy (2019), a crucial independence assumption is violated in BCH's dataset. More details and additional analyses can be found in an extended working paper version (Kranz and Pütz 2021—henceforth, KP).

I. The Rounding Problem

BCH collected data for more than 21,000 hypothesis tests. For most tests (90.2 percent) the reported coefficient μ and its standard error σ are collected and the (absolute) z-statistic $z = abs(\mu)/\sigma$ is computed from these values. For the remaining tests, the z-statistic was derived from a reported t-statistic (5.0 percent), p-value (4.7 percent), or confidence interval (0.1 percent).[1]

BCH's main statistical analyses focus on the 5 percent significance threshold. For their randomization tests, they assume that absent p-hacking or publication bias, the distribution of z-statistics would be continuous and differentiable so that in a sufficiently small window $[1.96 - h, 1.96 + h]$ around the 5 percent significance threshold, there should be roughly as many significant results with $z \geq 1.96$ as insignificant results with $z < 1.96$. They then compare the shares of significant and insignificant tests for a grid of window half-widths $h \in \{0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5\}$. The left panel of Figure 1 shows the corresponding shares of z-statistics above 1.96 for all window half-widths on a fine grid between 0.01 and 0.5 using the pooled data.

For window half-widths below 0.04, less than 50 percent of z-statistics are above the threshold of 1.96. But there is a massive, discontinuous increase of significant z-statistics once the window half-width exceeds 0.04. This jump has not been discussed by BCH whose smallest considered window half-width of $h = 0.05$ is already on the right-hand side of this discontinuity.

The discontinuity occurs because the dataset contains 260 z-statistics with a value of exactly 2. All these observations are counted as significant and thus cause the jump in the share of significant tests once $1.96 + h$ reaches 2. These 260 observations constitute 37.9 percent of the total observations in the smallest window analyzed by BCH. The left panel of Figure 2 shows the number of observations included for each window half-width and verifies the substantial jump at $h = 0.04$. The right panels of Figure 1 and Figure 2 present adjusted results and are discussed in the next section.

Rounding errors are likely the most important reason for the bunching of z-statistics at exactly $z = 2$; 68.6 percent of observations with a z-statistic of exactly 2 have just a single significant digit for the standard error and 97.7 percent have at most two significant digits. For the remaining observations these shares are just 17.2 percent and 59.8 percent, respectively.

If the coefficients and standard errors were reported with more significant digits, the computed z-statistic could well have been smaller than 1.96. For example, assume the reported standard error is $\sigma = 0.02$. Then this observation has a

[1] An initial version of our comment detected that BCH's conversion from p-values into z-statistics wrongly assumed that all p-values correspond to one-sided tests. The authors corrected this problem and also detected some smaller typos when converting the raw data. They kindly provided us with a corrected version of the dataset and also made it publicly available (Brodeur, Cook, and Heyes 2022). We use that updated dataset for all our analyses in this comment.

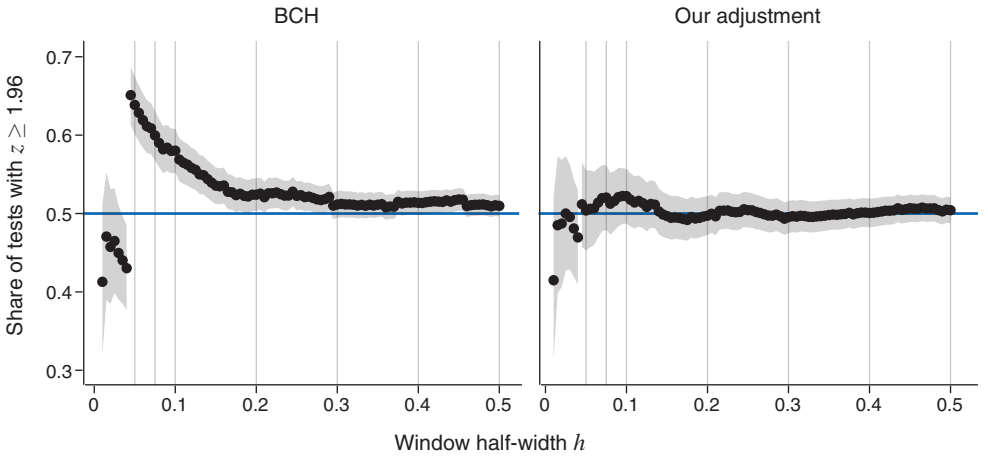

FIGURE 1. SHARE OF SIGNIFICANT RESULTS FOR DIFFERENT WINDOW HALF-WIDTHS

Notes: The left panel corresponds to the case of no adjustment for rounding errors as in BCH. The right panel shows results with our adjustment that omits all observations whose reported standard error has a significand below 37. The shaded areas indicate 95 percent confidence intervals based on binomial tests. The gray vertical lines indicate the window half-widths that BCH studied.

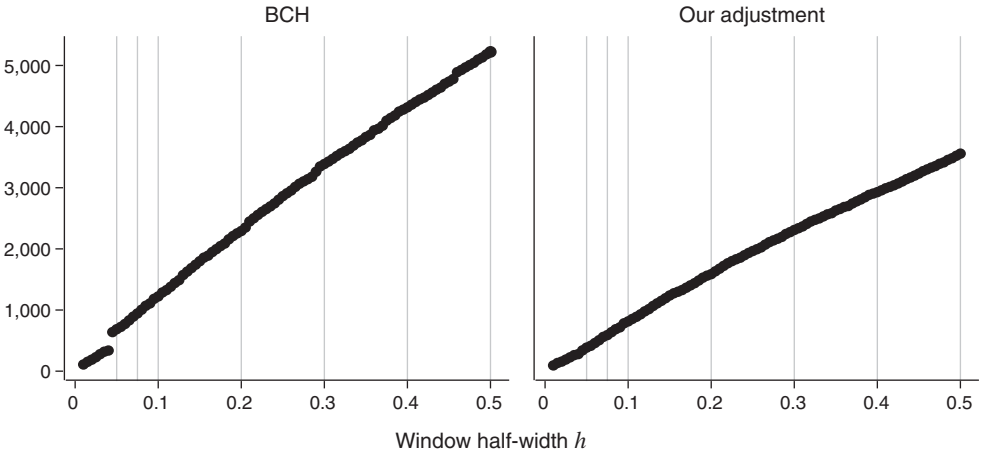

FIGURE 2. NUMBER OF INCLUDED OBSERVATIONS FOR DIFFERENT WINDOW HALF-WIDTHS

computed z-statistic of exactly $z = 2$ if the reported coefficient is also rounded to one significant digit and given by $\mu = 0.04$. If one computed the z-statistic using the original nonrounded values, it may range from an insignificant lower bound of $z = 1.4$ (i.e. $0.035/0.025$) to a highly significant upper bound of $z = 3$ (i.e. $0.045/0.015$).[2]

[2] Replying to an initial version of our comment, Brodeur et al. looked at the DID articles with the largest number of tests with $z = 2$ and collected the reported significance stars. From the collected 43 tests 5 had no stars, 14 had one star, 21 had two stars and 3 had three stars.

II. Adjusting for the Rounding Problem

There are different approaches to adjust for the rounding problem. Brodeur et al. (2016) and Bruns et al. (2019) deround reported coefficients and standard errors by assuming that missing digits are drawn from a uniform distribution. While intuitively appealing, we show in a Monte-Carlo study in the online Appendix that this uniform derounding approach may induce an attenuation bias in the randomization tests.[3] Furthermore, it is not clear how to compute valid confidence intervals for the uniform derounding approach. To overcome these issues, we implement an alternative approach that omits observations that are too coarsely rounded. Concretely, we omit all observations whose standard error has a significand s below a threshold $\bar{s} = 37$. The significand consists of the significant digit(s) written as an integer; e.g., for $\sigma = 0.012$, the significand is $s = 12$. A larger threshold $\bar{s}$ has the drawback of reducing statistical power by omitting more observations, but has the advantage of reducing the impact of rounding errors.[4] The latter effect is formalized by:

LEMMA 1: *Assume reported standard error σ and coefficient μ are rounded to the same number of decimal places. The true and reported z-statistics $\tilde{z}$ and z are guaranteed not to lie on opposite sides of an arbitrary threshold τ if the significand s of the standard error satisfies*

$$s \geq \frac{1+\tau}{2|z-\tau|}. \tag{1}$$

The proof of Lemma 1 is in the online Appendix. We say an observation is *misclassified* if its reported z-statistic z and true z-statistic $\tilde{z}$ lie on opposite sides of the significance threshold $\tau = 1.96$. Lemma 1 implies that our threshold $\bar{s} = 37$ is the smallest omission threshold guaranteeing that no observation with $z = 2$ is misclassified.[5]

Besides misclassification, rounding errors can also cause an observation to be wrongly included in a window $[1.96 - h, 1.96 + h]$ if z is inside the window but $\tilde{z}$ outside, or wrongly excluded the other way round.[6] Lemma 1 implies that our omission threshold $\bar{s} = 37$ guarantees that no remaining observation with $z = 2$ is wrongly included or excluded from windows with half-width $h \geq 0.08$. Thus, only for the two smallest window half-widths considered by BCH, we cannot rule out that after our adjustment some observations with $z = 2$ are still wrongly included in those windows.

Assume the unobserved true z-statistic $\tilde{z}$ follows a distribution function $F(\tilde{z})$ and let $F(\tilde{z} \,|\, s \geq \bar{s})$ be the corresponding distribution function of the z-statistics that we select. Crucial for our approach is:

[3] The Monte-Carlo study also confirms that without any rounding adjustment there is a substantial upward bias due to the bunching at $z = 2$.

[4] See the online Appendix for a quantification of the power loss induced by our adjustment method.

[5] The omission threshold $\bar{s} = 37$ also guarantees for our dataset that no other observation that originally has the same reported z-statistic as at least six other observations is misclassified. All the corresponding 229 distinct values of z-statistics have a larger distance to the significance threshold $\tau = 1.96$ than $z = 2$.

[6] Note that collecting data on the reported significance levels for each observation (if available) could be another way to solve the misclassification problem arising from rounding errors. But it would not solve the problems of potentially wrong inclusion or exclusion.

ASSUMPTION 1: *Our selection does not affect the distribution of true z-statistics:* $F(\tilde{z}) = F(\tilde{z} \mid s \geq \bar{s})$.

A model that satisfies Assumption 1 is the following. The true z-statistic $\tilde{z}$ is independently distributed from the true standard error $\tilde{\sigma}$ and the number of reported decimal places d. The true coefficient is given by $\tilde{\mu} = \tilde{z} \cdot \tilde{\sigma}$. The reported coefficients μ and σ are derived by rounding $\tilde{\mu}$ and $\tilde{\sigma}$ to d decimal places and the reported z-statistic is $z = \mu/\sigma$.

Assumption 1 cannot be directly tested as the true z-statistics $\tilde{z}$ are unobserved. However, it suggests that for the reported z-statistics the distribution of the selected sample should look similar to that of the full sample, except for the bunching points. Figure 6 (further below) shows that indeed the kernel density estimates of the two samples look very similar. In a similar spirit, Assumption 1 suggests that reported z-statistics and reported standard errors should not exhibit a strong correlation. Indeed, we find a correlation of only -0.0002 with a 95 percent confidence interval $[-0.0138, 0.0134]$. One plausible reason for this low correlation is that the main source of variation in σ (and μ) is different scaling of explanatory and dependent variables across the regressions, which seems uncorrelated with the true z-statistic. Also the empirical correlation between a dummy indicating whether an observation is omitted and the reported z-statistic is only -0.005 with a 95 percent confidence interval $[-0.019, 0.008]$.

One concern could be that coarse rounding is used as a form of p-hacking in order to wrongly suggest that z-statistics pass the significant threshold. This would imply that we observe more coarse rounding for z-statistics that are close to the 5 percent (or 10 percent) significance threshold where p-hacking concerns seem most relevant. However, Figure 6 shows that in that range the selected subsample without coarse rounding even has a slightly larger density than the complete sample. Also recall footnote 2, which shows that a substantial fraction of a selection of DID tests with $z = 2$ indeed reported significance levels above 5 percent, i.e., for these tests coarse rounding was not used to wrongfully suggest a 5 percent significance level.

For generating the right-hand panel of Figure 1, we have applied our adjustment procedure to the pooled data. The discontinuity at $h = 0.04$ vanishes and no clear evidence for p-hacking or publication bias remains. The shares of significant z-statistics decrease for all window half-widths; they are mostly close to 50 percent and the confidence intervals always include the 50 percent level. Overall, our adjustment omits 37.9 percent of observations, but it omits 87.3 percent of observations with $z = 2$. The right-hand panel of Figure 2 shows that with our sample selection, the number of included observations increases smoothly with growing window half-width h without any visible jumps.

III. Replication Results

In this section we replicate the main analyses from BCH using our adjusted dataset. BCH compare the evidence for p-hacking and publication bias between four different identification strategies: DID, IV, randomized control trial (RCT), and regression discontinuity design (RDD). The share of observations with a z-statistic of exactly 2 differs substantially between the subsamples corresponding

TABLE 1—RANDOMIZATION TESTS

	ALL (1)	DID (2)	IV (3)	RDD (4)	RCT (5)
Window half-width 0.05					
Proportion significant	0.504	0.446	0.581	0.395	0.5
(*p*-value)	(0.459)	(0.884)	(0.036)	(0.928)	(0.538)
Observations	385	101	136	38	110
Window half-width 0.075					
Proportion significant	0.52	0.486	0.591	0.453	0.494
(*p*-value)	(0.172)	(0.659)	(0.006)	(0.809)	(0.588)
Observations	590	148	198	64	180
Window half-width 0.1					
Proportion significant	0.522	0.507	0.56	0.476	0.51
(*p*-value)	(0.110)	(0.446)	(0.029)	(0.707)	(0.400)
Observations	814	217	266	84	247
Window half-width 0.2					
Proportion significant	0.497	0.469	0.545	0.459	0.485
(*p*-value)	(0.599)	(0.904)	(0.023)	(0.882)	(0.763)
Observations	1,587	397	506	183	501
Window half-width 0.3					
Proportion significant	0.495	0.478	0.53	0.491	0.477
(*p*-value)	(0.676)	(0.868)	(0.056)	(0.644)	(0.903)
Observations	2,309	584	736	265	724
Window half-width 0.4					
Proportion significant	0.501	0.497	0.539	0.469	0.477
(*p*-value)	(0.478)	(0.574)	(0.009)	(0.890)	(0.926)
Observations	2,928	720	929	352	927
Window half-width 0.5					
Proportion significant	0.504	0.506	0.548	0.48	0.469
(*p*-value)	(0.313)	(0.367)	(0.001)	(0.807)	(0.984)
Observations	3,558	869	1,123	431	1,135

Notes: Replicates Table 3 in BCH. We present for several windows centered around $z = 1.96$ the proportion of significant observations and test if it is statistically greater than 0.5.

to the different identification strategies. In the smallest window studied by BCH, it ranges from only 16.6 percent for IV to 50.0 percent for DID. Correspondingly, adjusting for rounding errors affects in particular the DID results.

A. *Randomization Tests*

Table 3 in BCH shows the share of observations with $z \geq 1.96$ in the window $[1.96 - h; 1.96 + h]$ for different window sizes h and different identification strategies. It also presents the p-value of a one-sided binomial test with the null hypothesis that this share does not exceed 50 percent.

Table 1 shows the results using our adjustment. We find the largest difference to BCH for DID. While in BCH the share of significant tests varied between 53.0 percent and 70.7 percent (with p-values between 0.000 and 0.030), with our adjustment for rounding the share ranges between 44.6 percent and 50.7 percent (with p-values between 0.367 and 0.904). For RDD and RCT, the share of significant tests is never significantly above 50 percent with adjustment while BCH found significant results

for smaller window sizes. For IV, the adjustment changes little compared to BCH in terms of magnitude and *p*-values.

At the 10 percent threshold our results are closer to BCH (see online Appendix Table A3). In particular, significant results for DID for larger window sizes remain. As Figure 4 shows, densities of *z*-statistics generally slope upward around the 10 percent threshold. In contrast, at the 1 percent threshold, the densities slope downward and the share of tests with $z \geq 2.576$ is significantly below 50 percent for many window half-widths (see online Appendix Table A4).

B. *Caliper Tests*

BCH proceed their analyses with so-called caliper tests. Again, all observations with *z*-statistics in a specified window around the $z = 1.96$ threshold are considered and probit regressions of the following form are performed:

$$(2) \qquad \Pr\big(Significant_i = 1\big) = \Phi\big(\alpha + X_i'\boldsymbol{\delta} + \gamma DID_i + \lambda IV_i + \phi RDD_i\big).$$

$Significant_i$ is a dummy variable indicating whether $z_i \geq 1.96$ and X_i is a vector of control variables that, depending on the specification, includes author and article characteristics as well as journal and field fixed effects. Table 2 shows our replication results of the caliper tests using the adjusted dataset.

For example, the first column shows that in the window $[1.96 - 0.5; 1.96 + 0.5]$ tests from IV studies are 10.1 percentage points more likely to be significant than tests from RCT studies of which 47 percent are significant in that window. Overall, the estimates for IV are very similar to those of BCH and remain significant when performing our adjustment for rounding errors. In contrast, effect sizes for DID substantially reduce with our adjustment and, in contrast to the findings of BCH, are not statistically significant in any specification. So after adjustment for rounding errors, the caliper tests provide no more evidence for *p*-hacking of DID studies at the 5 percent threshold than the randomization tests. Results at the 10 percent and 1 percent threshold are shown in Tables A5 and A6 in the online Appendix and are closer to BCH's original findings.

C. *Distribution of z-Statistics*

In this subsection we show the empirical distribution of *z*-statistics with our rounding adjustment for different subsamples analyzed by BCH. In general, the kernel density estimates are quite similar to BCH's and show even a slightly more pronounced second hump.[7] The main effect of the adjustment procedure is that the histograms exhibit less bunching at $z = 2$ and at other bunching points, such as $z = 1$.

Figure 3 shows the empirical distribution of *z*-statistics overall as well as the respective distributions for the top five and non-top five journals with adjustment for

[7] Note that the default kernel density is biased toward zero at the left margin, but we use that estimator for better comparison with BCH. See online Appendix Figure A3 for a version of Figure 6 using a bias-corrected density estimator.

TABLE 2—CALIPER TESTS

	(1)	(2)	(3)	(4)	(5)	(6)
DID	0.056	0.045	0.029	0.028	0.024	−0.036
	(0.038)	(0.039)	(0.040)	(0.041)	(0.045)	(0.059)
IV	0.101	0.100	0.079	0.084	0.097	0.098
	(0.034)	(0.037)	(0.038)	(0.038)	(0.041)	(0.050)
RDD	0.094	0.082	0.069	0.069	0.074	0.038
	(0.063)	(0.061)	(0.058)	(0.058)	(0.060)	(0.073)
Top 5		−0.028	−0.021			
		(0.054)	(0.110)			
Year = 2018		0.006	0.013	0.022	−0.004	0.022
		(0.033)	(0.033)	(0.034)	(0.037)	(0.041)
Experience		−0.008	−0.013	−0.013	−0.011	0.000
		(0.008)	(0.008)	(0.008)	(0.010)	(0.011)
Experience squared		0.002	0.017	0.020	0.020	−0.005
		(0.021)	(0.022)	(0.022)	(0.026)	(0.032)
Top institution		−0.001	0.003	0.002	−0.040	−0.078
		(0.054)	(0.053)	(0.052)	(0.059)	(0.069)
Top PhD institution		0.014	−0.013	−0.013	0.058	0.143
		(0.045)	(0.045)	(0.046)	(0.050)	(0.061)
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 fixed effects			Y			
Journal fixed effects				Y	Y	Y
Observations	3,558	3,558	3,558	3,558	2,626	1,585
Window	[1.96 ± 0.50]	[1.96 ± 0.50]	[1.96 ± 0.50]	[1.96 ± 0.50]	[1.96 ± 0.35]	[1.96 ± 0.20]
RCT significance rate	0.47	0.47	0.47	0.47	0.48	0.49

Notes: Replicates Table 4 in BCH. The shown coefficients are marginal effects at the means. For dummy variables we measure the effect of a change from zero to one. Standard errors in parentheses are clustered at article level. Observations are weighted by the inverse of the number of tests conducted in the same article.

the rounding problem. For top five articles no bunching right of the 5 percent significance level is observable anymore, but the density still increases strongly right of the 10 percent threshold.

Figure 4 compares the densities of z-statistics for the four strategies of causal identification. While the sizes of the second humps of the kernel density estimates don't decrease compared to BCH, bunching at exactly $z = 2$ is substantially reduced.

Figure 5 compares the distribution of z-statistics over time for three top journals (upper panels) and the top 25 journals (lower panels).[8] Omitting too coarsely rounded values does not change the main message of BCH's Figure 3, namely that the distributions for both journal groups do not change remarkably over time. See the online Appendix for versions of BCH's Figures 5 and 6 using our adjusted dataset.

[8] The data sets including z-statistics published between 2005 and 2011 are taken from Brodeur, Cook, and Heyes (2022) and Brodeur et al. (2019).

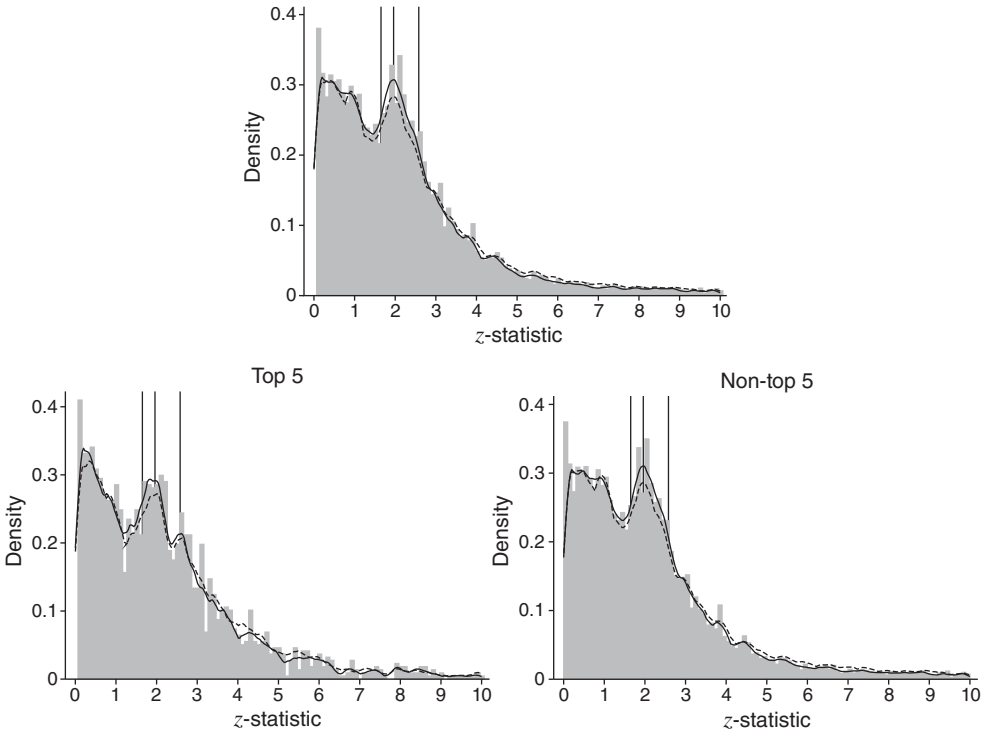

FIGURE 3. z-STATISTICS IN 25 TOP ECONOMICS JOURNALS

Notes: Replicates Figure 1 in BCH. The top panel presents the distribution of all test statistics for $z \in [0, 10]$. The bottom panels show test statistics from the top five journals (*American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economic Studies*, left panel) and the remaining sample (right panel). The vertical lines indicate the critical z-statistics at the 10 percent, 5 percent, and 1 percent significance levels. The histograms have bin size 0.1. The black lines are density estimates based on a Epanechnikov kernel with bandwidth 0.1. The dotted lines show the corresponding densities using BCH's dataset.

D. *Excess Test Statistics*

BCH hypothesize that absent publication bias and *p*-hacking, the distribution of absolute z-statistics would follow for each identification strategy a noncentral t-distribution that is truncated at zero. Figure 6 compares these t-distributions with the empirical densities.[9] Whether we adjust for rounding (black line) or take the original sample (gray line) leads visually only to small differences in the empirical kernel density estimates.

Yet, there is an issue regarding the calibration of the latent distributions absent *p*-hacking and publication bias. BCH calibrate the degrees of freedom and noncentrality parameters of these t-distributions by matching the probability mass in the tails ($z > 5$) with its analog from the empirical distributions. They base this calibration on the assumption that the observed test statistic distribution above $z = 5$ should be free of *p*-hacking or publication bias. However, even if publication

[9] BCH showed in their Figure 4 by mistake the densities of the t-distributions without accounting for the truncation at zero. Figure 6 shows the corrected densities.

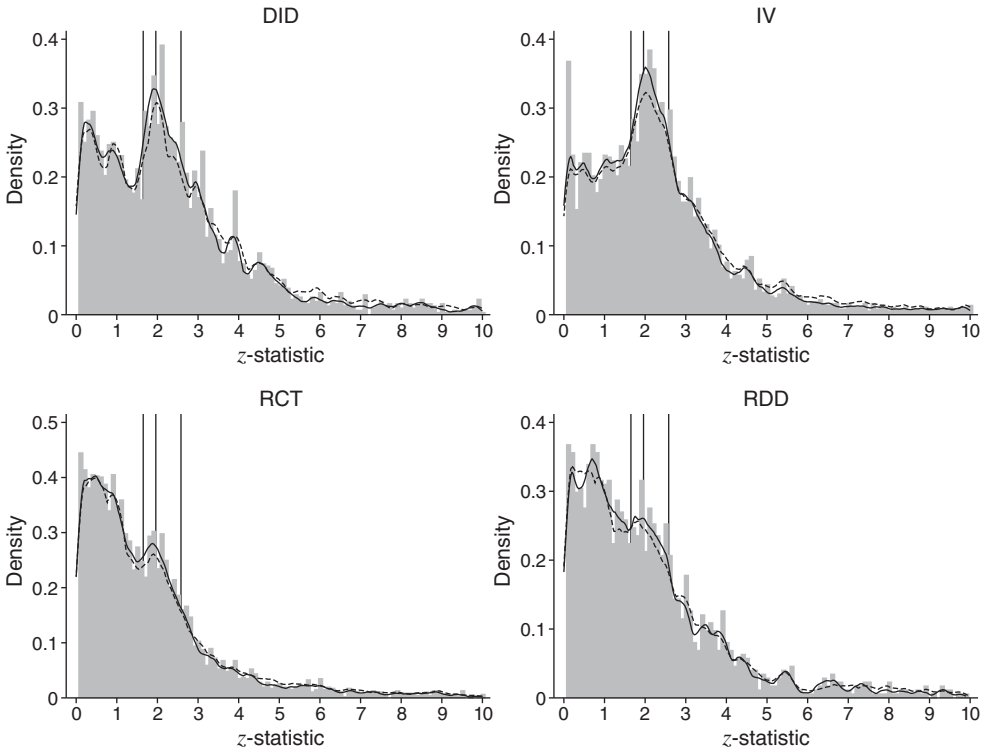

FIGURE 4. z-STATISTICS BY METHOD

Notes: Replicates Figure 2 in BCH. The figure presents the distributions of z-statistics for $z \in [0, 10]$ for each of the four identification methods: DID, IV, RCT, and RDD. The vertical lines indicate the critical z-statistics at the 10 percent, 5 percent, and 1 percent significance levels. The histograms have bin size 0.1. The black lines are density estimates based on a Epanechnikov kernel with bandwidth 0.1. The dotted lines show the corresponding densities using BCH's dataset.

bias only affects the publication probability for $z < 5$, it will affect the observed probability mass in the tails, since the total probability mass must add up to one. Correspondingly, we show in Proposition 1 in the online Appendix that BCH's calibration approach is not able to recover the true distribution of z-statistics absent publication bias, even if the correct functional form is assumed.[10]

E. *Estimating the Amount of Distortion*

In addition, BCH derive the latent distributions of z-statistics without publication bias and p-hacking using an approach developed by Andrews and Kasy (2019). Table 3 shows the replicated results using our adjusted dataset. Comparing to BCH, we see that rounding has only a small impact on those results.

[10] See KP for a discussion of additional issues of BCH's excess test statistics approach.

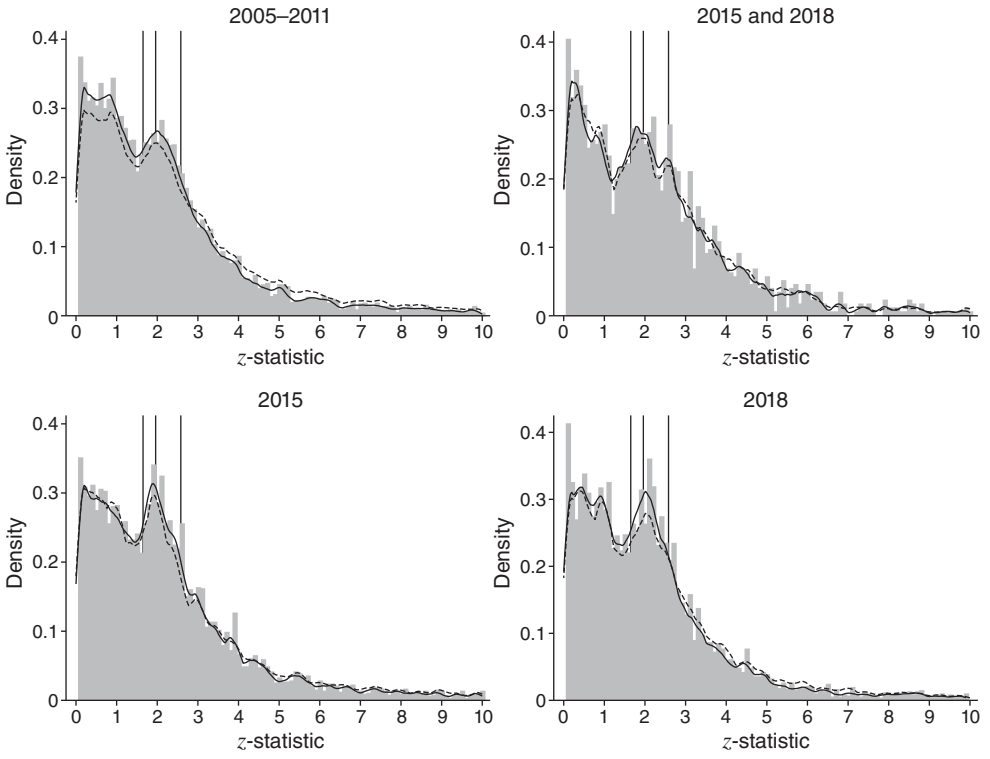

FIGURE 5. z-STATISTICS IN 25 TOP ECONOMICS JOURNALS

Notes: Replicates Figure 3 in BCH. The top panels presents the distribution of test statistics from the *American Economic Review*, *Journal of Political Economy*, and the *Quarterly Journal of Economics* for $z \in [0, 10]$ over time. The top left panel is based on data from Brodeur et al. (2016). The bottom panels show test statistics from the top 25 over time. The vertical lines indicate the critical z-statistics at the 10 percent, 5 percent, and 1 percent significance levels. The histograms have bin size 0.1. The black lines are density estimates based on a Epanechnikov kernel with bandwidth 0.1. The dotted lines show the corresponding densities using BCH's dataset.

However, the approach relies on the identifying assumption that in the latent distribution absent publication bias, the standard error σ_i and estimated coefficient μ_i are independently distributed from each other. For a statistical test, we compute for all observations the weighted correlation between $\log \sigma_i$ and $\log abs(\mu_i)$ using as weights the inverse of the estimated publication probabilities. Under the null hypothesis that all assumptions of the chosen implementation of the Andrews and Kasy (2019) approach are satisfied, this inverse probability weighting allows to recover the correlation in the unobserved latent distribution of tests if no publication bias was present.[11] Table A8 in the online Appendix shows the computed correlations and bootstrapped confidence intervals for different subsamples. The correlations range from 0.89 to 0.92 and the 95 percent confidence intervals from 0.88 to 0.94. These results suggest that the crucial independence assumption of Andrews and Kasy (2019) is strongly violated in BCH's dataset.

[11] We thank Isaiah Andrews who proposed the idea for this inverse probability weighting approach and pointed out a problem of an earlier idea of ours to test the independence assumption.

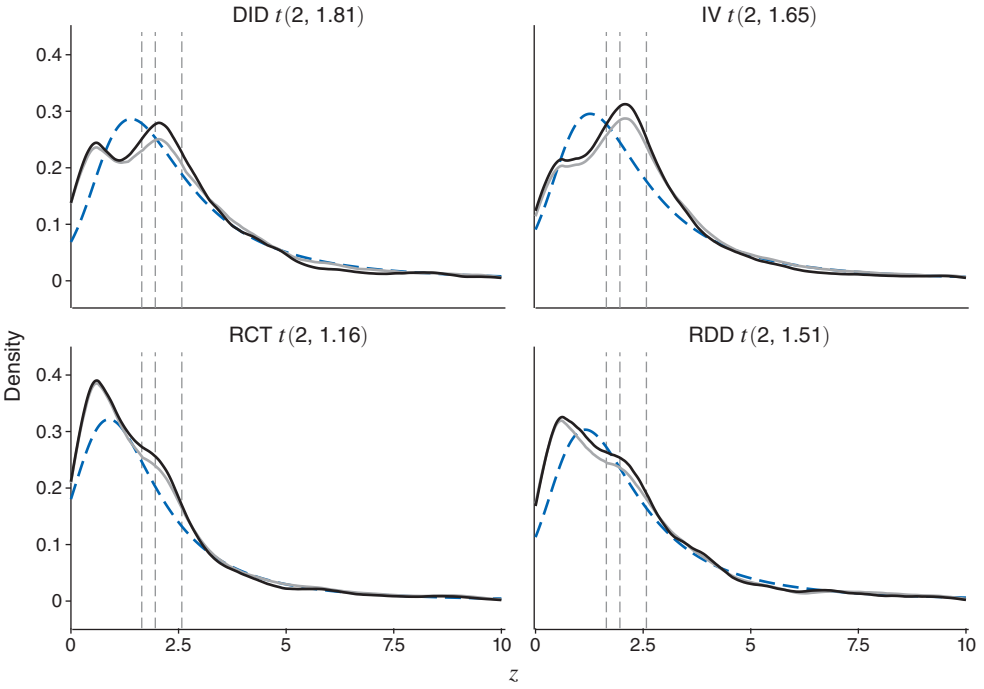

FIGURE 6. EXCESS TEST STATISTIC PLOTS

Notes: Replicates Figure 4 in BCH. The dashed blue lines show the density of BCH's originally estimated truncated, noncentral *t*-distribution. The black lines show the empirical kernel density estimate using our adjusted sample and the gray line shows its analog using the original sample.

Table 3—Relative Publication Probabilities

	DID (1)	IV (2)	RCT (3)	RDD (4)
Panel A				
$\beta_{[0<Z<1.96]}$	0.222	0.253	0.574	0.521
	(0.011)	(0.014)	(0.027)	(0.047)
Scale	0.011	0.060	0.054	0.059
	(0.002)	(0.005)	(0.002)	(0.007)
Location	0.005	0.035	0.026	0.039
	(0.001)	(0.003)	(0.001)	(0.004)
Degrees of freedom	2.317	2.660	2.475	2.139
	(0.063)	(0.076)	(0.075)	(0.123)
Panel B				
$\beta_{[0<Z<1.65]}$	0.162	0.195	0.544	0.503
	(0.009)	(0.013)	(0.031)	(0.057)
$\beta_{[1.65<Z<1.96]}$	0.505	0.647	0.947	0.916
	(0.039)	(0.046)	(0.066)	(0.113)
$\beta_{[1.96<Z<2.33]}$	0.683	0.986	1.122	1.199
	(0.049)	(0.062)	(0.076)	(0.137)
Scale	0.010	0.052	0.051	0.055
	(0.001)	(0.005)	(0.002)	(0.007)
Location	0.004	0.031	0.025	0.038
	(0.001)	(0.003)	(0.001)	(0.005)
Degrees of freedom	2.544	2.736	2.470	2.096
	(0.076)	(0.084)	(0.078)	(0.135)

Notes: Replicates Table 5 in BCH. It shows the results of applying the publication bias model presented in Andrews and Kasy (2019). Note that the third coefficient in panel B, also in BCH refers to z-statistics in the interval $[1.96, 2.33]$; the label $[1.96, 2.58]$ was a mistake.

REFERENCES

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

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review* 110 (11): 3634–60.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2022. "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/E120246V2.

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.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2019. Replication Data for: Star Wars: The Empirics Strike Back. American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E113633V1.

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

Kranz, Sebastian, and Peter Pütz. 2021. "Rounding and Other Pitfalls in Meta-Studies on p-Hacking and Publication Bias: A Comment on Brodeur et al. (2020)." SSRN 3848786.

Kranz, Sebastian, and Peter Pütz. 2022. "Replication Data for: Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics: Comment." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E159221V1.

This article has been cited by:

1. Tianyi Ma, Kai-Hong Tee, Baibing Li. 2022. On hedge fund inceptions in a competitive market. *The European Journal of Finance* **94**, 1-26. [Crossref]
2. Abel Brodeur, Nikolai Cook, Anthony Heyes. 2022. Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics: Reply. *American Economic Review* **112**:9, 3137-3139. [Abstract] [View PDF article] [PDF with links]