American Economic Review 2022, 112(9): 3137–3139
https://doi.org/10.1257/aer.20220277

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

By Abel Brodeur, Nikolai Cook, and Anthony Heyes*

In Brodeur, Cook, and Heyes (*2020*) *we present evidence that instrumental variable* (*and to a lesser extent difference-in-difference*) *articles are more p-hacked than randomized controlled trial and regression discontinuity design articles. We also find no evidence that* (*i*) *articles published in the top five journals are different;* (*ii*) *the "revise and resubmit" process mitigates the problem;* (*iii*) *things are improving through time. Kranz and Pütz* (*2022*) *apply a novel adjustment to address rounding errors. They successfully replicate our results with the exception of our shakiest finding: after adjusting for rounding errors, bunching of test statistics for difference-in-difference articles is now smaller around the 5 percent level* (*and coincidentally larger at the 10 percent level*). (*JEL* A14, C12, C52)

In Brodeur, Cook, and Heyes (2020) —henceforth, BCH—we collect test statistics from articles that use experimental and quasi-experimental inference methods. The sample of articles was published during 2015 and 2018 in 25 top economics journals. We apply three different approaches to document differences in p-hacking across inference method, journal ranking, and over time. We also investigate whether the "revise and resubmit" process mitigates p-hacking. We find that instrumental variable (IV) and to a lesser extent difference-in-differences (DID) articles are particularly problematic in comparison to articles featuring a randomized controlled trial (RCT) or a regression discontinuity design (RDD). We find no evidence that (i) articles published in the "top five" journals are different in this regard; (ii) the "revise and resubmit" process mitigates the problem; (iii) things are improving through time.

Kranz and Pütz (2022)—henceforth, KP—apply a novel adjustment to our data in order to address possible rounding errors, which arise from the coarseness of how test statistics are often reported. As explained in Brodeur et al. (2016), rational numbers that can be expressed as ratios of small integers get over-represented because of the low precision used by authors. For instance, if the estimate is reported to be 0.020 and the standard error is 0.010, then our reconstructed z-statistic would be exactly two. To address this potential issue, KP exclude all observations from

*Brodeur: Department of Economics, University of Ottawa (email: abrodeur@uottawa.ca); Cook: Department of Economics, Wilfrid Laurier University (email: ncook@wlu.ca); Heyes: Department of Economics, University of Ottawa and University of Exeter Business School (email: aheyes@uottawa.ca). Isaiah Andrews was the coeditor for this article.

[†]Go to https://doi.org/10.1257/aer.20220277 to visit the article page for additional materials and author disclosure statements.

our data that are "too coarsely rounded." Specifically, they exclude reconstructed z-statistics that have a significand below 37 (where if $\sigma = 0.012$ the significand is 12.) This adjustment omits 87.3 percent of coefficients where $z = 2$.

We would like to begin by thanking KP. Our reply is brief as using their new adjustment and sample, KP confirm all the earlier results, except for our most tentative finding: after adjusting for rounding errors, the extent of p-hacking for DID articles is now smaller at the 5 percent level (and coincidentally becomes larger at the 10 percent level). We are grateful to KP for this confirmation and for further illuminating this pattern with respect to DID. For the reply we proceed in two steps. First, we discuss derounding in our sample. Second, we compare the test statistic distributions before and after applying the KP adjustment.

The comment notes that the distribution of test statistics in BCH has a "spike" of extra mass where $z = 2$ for DID. This is indeed apparent from the main figures of the study. They continue by noting that BCH assumes that all tests with a reconstructed z-statistic above 1.96 are significant at the 5 percent level, but that some of those tests may be miscategorized because of rounding conventions. This is a great point and one we were aware of as it is discussed in Brodeur et al. (2016).[1] In preparing this reply, we re-examined DID articles reporting coefficients and standard errors for which we calculate $z = 2$, and documented the reported statistical significance by the presence of "stars." In our sample there is a total of 114 instances of $z = 2$ for DID. We recoded the articles with the largest number of $z = 2$ and added to our dataset whether the author reported one star, two stars, three stars, or none. In the end, we recoded 43 of the 114 test statistics where $z = 2$. Of these 43 tests, 14 had one star, 21 had two stars, 3 had three stars, and 5 had none. This examination suggests that about half of tests coded as $z = 2$ are plausibly incorrectly coded as $z > 1.96$. The fact that over 80 percent of these test statistics are statistically significant at the 10 or 5 percent levels suggest the rounding issue for $z = 2$ seems unlikely to change our main conclusions. Moreover, many test statistics coded as statistically significant at the 5 percent level are in fact barely statistically significant at the 10 percent level, suggesting that bunching around this significance level threshold for DID is slightly larger than originally documented.

To probe this further we then directly compare the test statistic distributions before and after applying the KP adjustment and sample restrictions. We then apply the visual inspection approach in BCH. There we find that the test statistic distribution of the "top five" journals was similar to the remainder of the sample, that the 2015 and 2018 distributions were similar, and that IV and to a lesser extent DID articles exhibited markedly different distributions compared to those using RCT and RDD. Here we present Figure 1 which reproduces the comment's Figure 4 and corresponds to Figure 2 in BCH which juxtaposes our data and theirs. We invite the reader to engage in some visual inspection as to whether the adjustment transforms what they think they learn from the study. To our eyes the conclusions are virtually the same.

We are grateful to KP for taking the time to reproduce our results and conducting a replication using their novel adjustment for possible rounding of test statistics

[1] In the online Appendix of this reply, we also show that the distribution of tests is similar if we focus just on that subsample of test statistics for which we have no measurement error—i.e., tests in which the author(s) reported a p-value, t-statistic or confidence interval.

reported in the published papers that underpin our analysis. Overall, we agree that applying their approach the extent of p-hacking for DID appears to be smaller at the 5 percent level and larger at the 10 percent level. The results with respect to the other methods, and all secondary results, are not meaningfully changed.

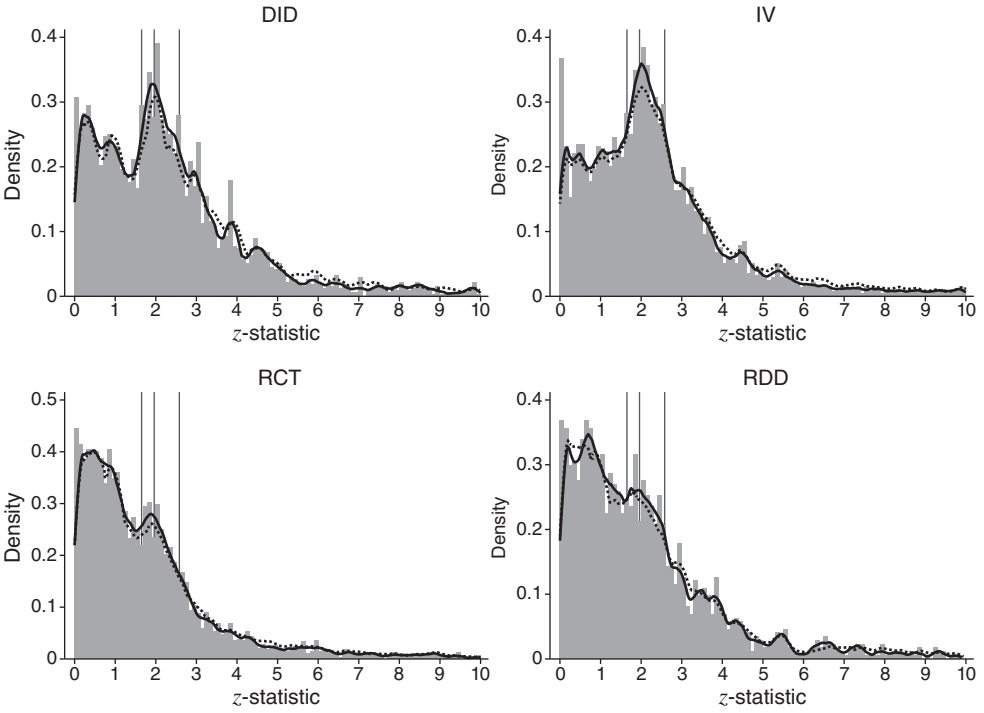

FIGURE 1. *z*-STATISTICS BY METHOD

Notes: This figure is taken from KP (Figure 4), who replicate and extend Figure 2 of BCH using a new derounding adjustment. The figure presents the distributions of z-statistics for $z \in [0, 10]$ by method: difference-in-differences (DID), instrumental variables (IV), randomized controlled trial (RCT), and regression discontinuity design (RDD). Bins are 0.1 wide. Vertical lines indicate the conventional 10, 5, and 1 percent significance levels. There are Epanechnikov kernel density estimates based on the two versions of the data. The dotted kernel corresponds to BCH. The solid kernel reflects the KP derounding. We present similar figures for the other results in an online Appendix.

REFERENCES

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: Reply." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E165621V1.

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.

Kranz, Sebastian, and Peter Pütz. 2022. "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics: Comment." *American Economic Review* 112 (9): 3124–36.