ONLINE APPENDIX
Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics: Reply

By Abel Brodeur, Nikolai Cook and Anthony Heyes[*]

In Brodeur et al. (2020) we present evidence that IV (and to a lesser extent DID) articles are more p-hacked than RCT and RDD articles. We also find no evidence that: (i) articles published in the Top 5 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 DID articles is now smaller around the 5% level (and coincidentally larger at the 10% level).

Online Appendix Figures A1, A2, and A3 compare the distribution of test statistics in our sample and in Kranz and Pütz (2022)'s restricted sample by journal ranking, over time, and by publication status, respectively. We again invite the reader to engage in some visual inspection as to whether the adjustment transforms what they think they learn from the paper. To our eyes the pictures look virtually the same.

We also show that the distribution of tests is similar if we focus only on the subsample of test statistics for which rounding errors are absent - i.e., tests in which the author(s) reported a p-value, t-statistic, or confidence interval. In Online Appendix Figure 4, we plot the distribution of tests for this subsample of tests (right graph) below and compare it to the distribution of tests for the whole sample (left graph). The distribution of tests is rather similar, with a relatively larger spike just above the 10% significance threshold for the subsample, and a similar mass of tests from 1.65 to 1.96 than the full sample. This suggests that the distribution of tests and the extent of inflation is quite similar in the subsample and full sample.

In Online Appendix Figure 5, we plot the distribution of tests by method for the reduced sample for IV, DID and RCT. (The subsample size becomes very small for RDD for the reduced sample, so we omitted that graph.) Of course, the distributions are now much noisier than those originally reported because of the considerable erosion of sample size. However, the distribution for IV still presents a pronounced global and local maximum around 2, and a mass shift away from

[*] Brodeur: abrodeur@uottawa.ca. Department of Economics, University of Ottawa. Cook: ncook@wlu.ca. Department of Economics, Wilfrid Laurier University. Heyes: aheyes@uottawa.ca. Department of Economics, University of Ottawa and University of Exeter Business School.

the marginally statistically insignificant interval (just left of 1.65). For DID, the share of tests between the 10% and 5% thresholds is even larger for the reduced sample than for the full sample with a maximum just after 1.65. Overall, this figure also suggests that our main results are robust to the use of this subsample of tests with no measurement error and as with the application of Kranz and Pütz (2022)'s adjustment, the extent of p-hacking now seems larger for DID at the 10% level than at the 5% level.

REFERENCES

Brodeur, A., Cook, N., and Heyes, A. (2020). Methods matter: P-hacking and publication bias in causal analysis in economics. *American Economic Review*, 110(11):3634–60.

Brodeur, A., Lé, M., Sangnier, M., and Zylberberg, Y. (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, 8(1):1–32.

Kranz, S. and Pütz, P. (2022). Methods matter: P-hacking and publication bias in causal analysis in economics: Comment. *American Economic Review.*

I. Appendix Figures

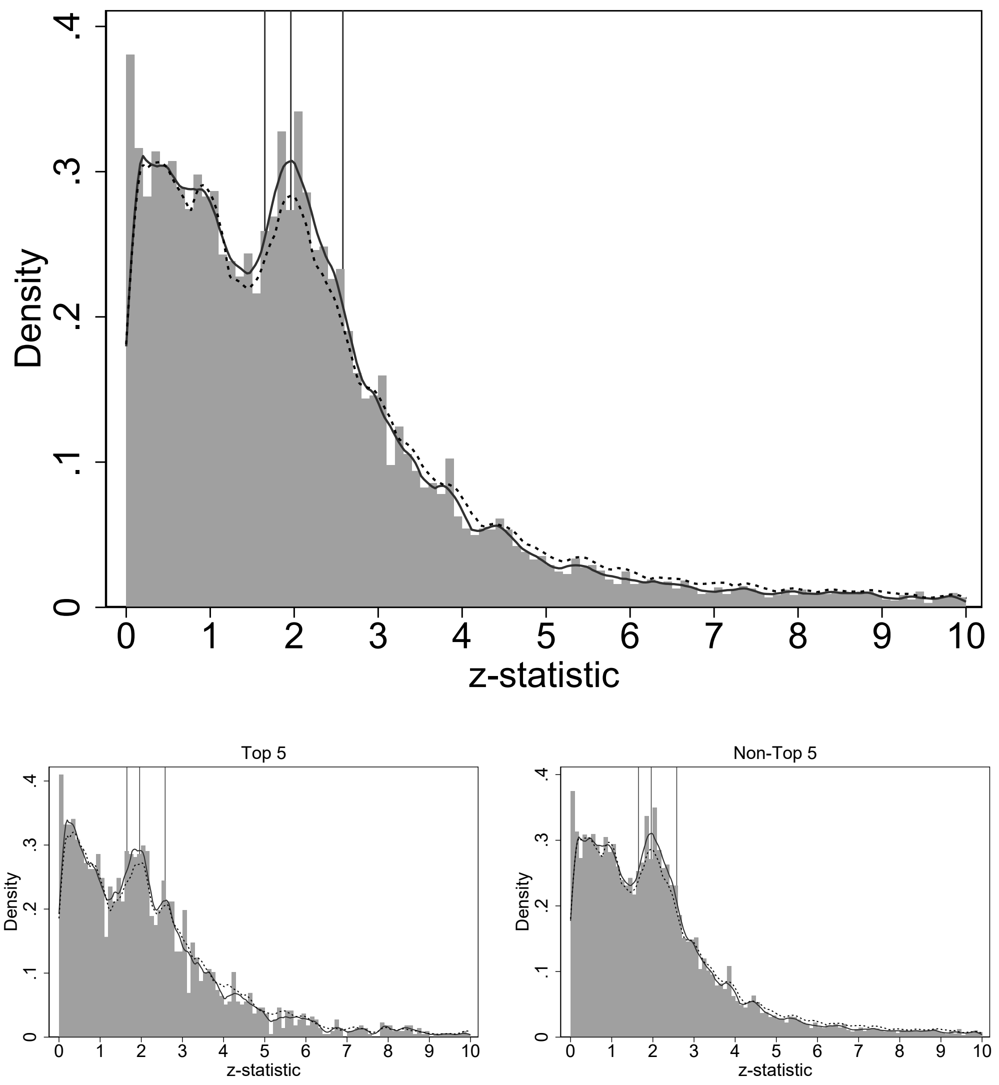

Figure A1. z-Statistics in 25 Top Economics Journals

Note: This figure is taken from Kranz and Pütz (2022, Figure 3), who replicate and extend Figure 1 of Brodeur et al (2020) using a new derounding adjustment. The top panel displays histograms of all test statistics for $z \in [0, 10]$. The bottom left panel presents test statistics from the "Top 5" (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. Vertical lines indicate the conventional 10%, 5% and 1% significance levels. There are Epanechnikov kernel density estimates based on the two versions of the data. The dotted kernel corresponds to Brodeur et al (2020). The solid kernel reflects the Kranz and Pütz (2022) derounding.

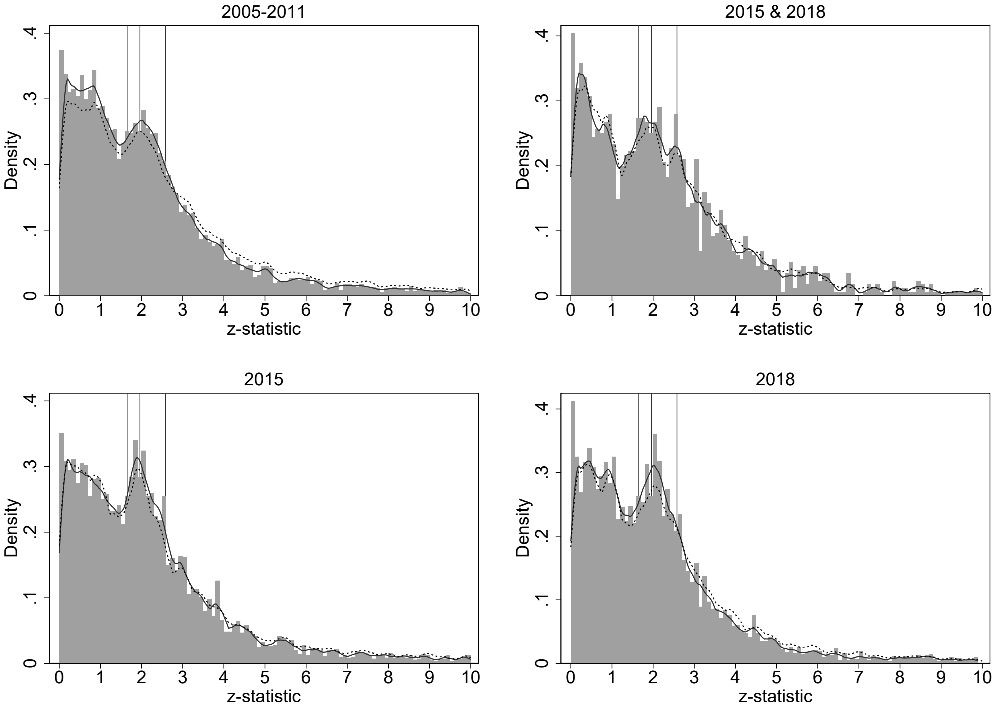

Figure A2. z-Statistics over Time

Note: This figure is taken from Kranz and Pütz (2022, Figure 5), who replicate and extend Figure 3 of Brodeur et al (2020) using a new derounding adjustment. 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 same journals during the Brodeur et al. (2020) sample period.The bottom left panel is top 25 journals in 2015 and the bottom right is top 25 journals in 2018. Vertical lines indicate the conventional 10%, 5% and 1% significance levels. There are Epanechnikov kernel density estimates based on the two versions of the data. The dotted kernel corresponds to Brodeur et al. (2020). The solid kernel reflects the Kranz and Pütz (2022) derounding.

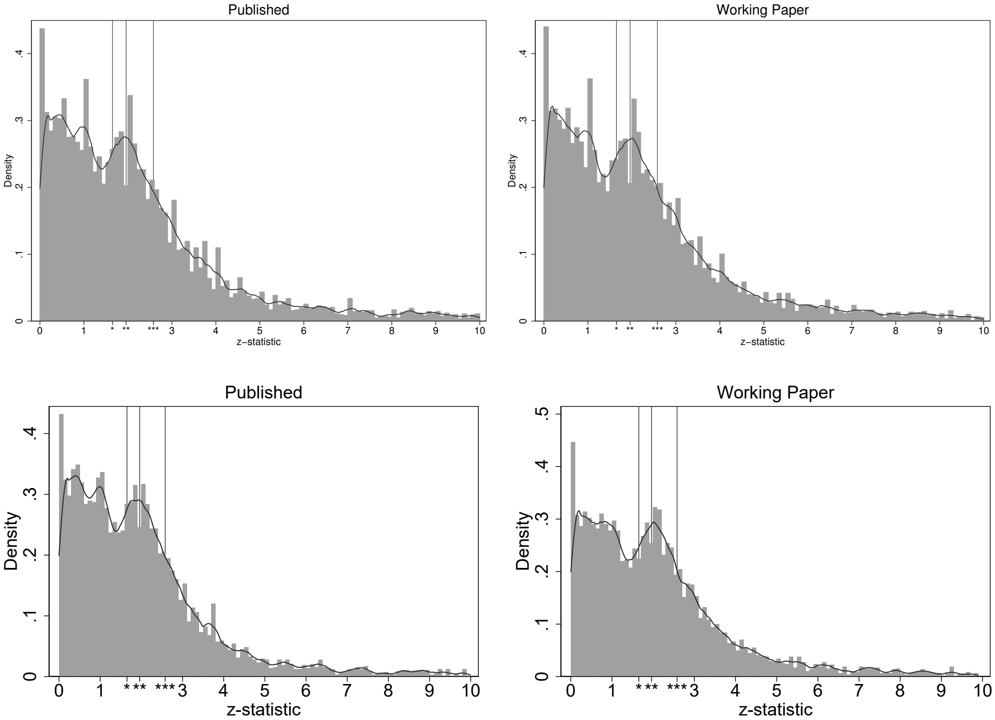

Figure A3. z-statistic by publication status – balanced sample

Note: The top 2 panels are taken from Figure 6 of Brodeur et al (2020). The bottom 2 panels are taken from Figure A2 of Kranz and Pütz (2022). Vertical lines indicate the conventional 10%, 5% and 1% significance levels. There are Epanechnikov kernel density estimates based on the two versions of the data. The dotted kernel corresponds to Brodeur et al (2020). The solid kernel reflects the Kranz and Pütz (2022) derounding.

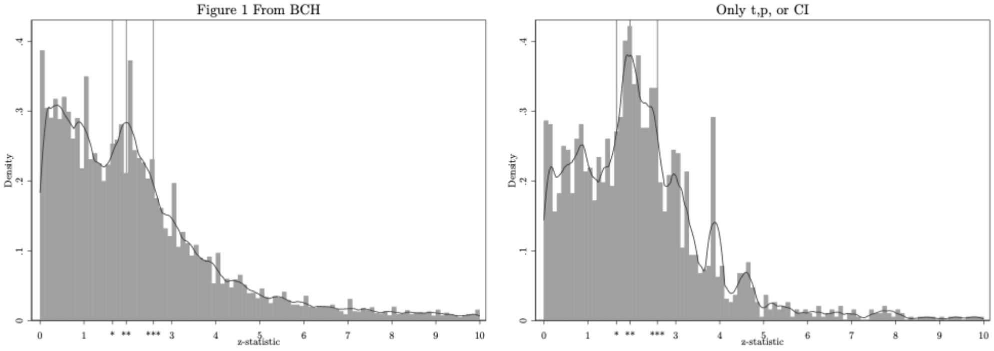

Figure A4. Distributions of tests for the subsample of test statistics for which the author(s) reported a p-value, t-statistic or confidence interval.

Note: This figure plots the distribution of tests for the subsample of test statistics for which the author(s) reported a p-value, t-statistic or confidence interval (right graph) and compare it to the distribution of tests for the whole sample (left graph). Vertical lines indicate the conventional 10%, 5% and 1% significance levels.

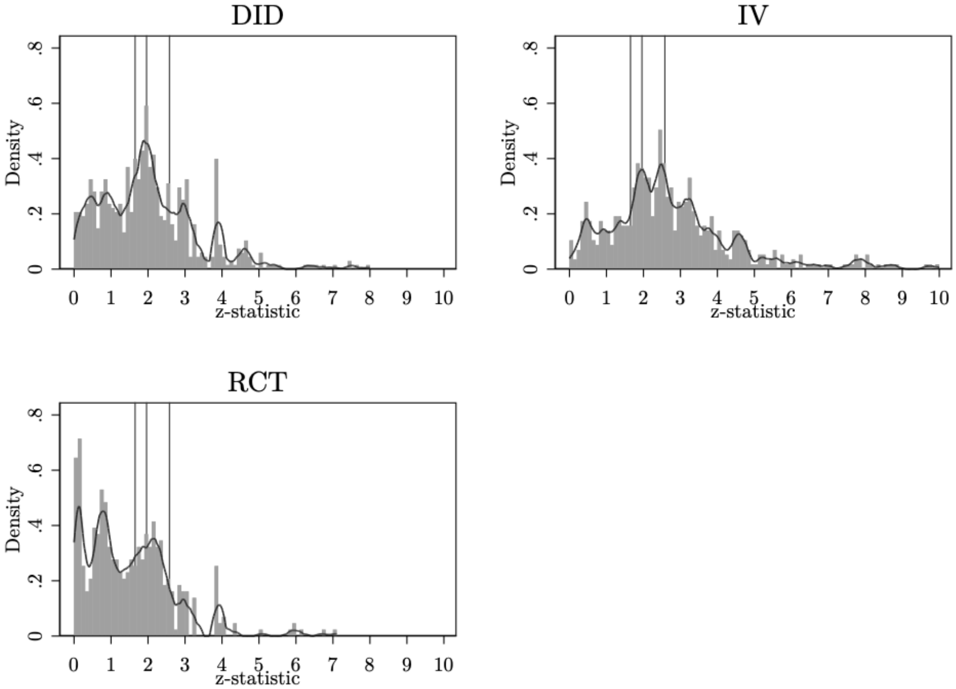

Figure A5. Distributions of tests for the subsample of test statistics for which the author(s) reported a p-value, t-statistic or confidence interval, by method.

Note: These figures plot the distribution of tests for the subsample of test statistics for which the author(s) reported a p-value, t-statistic or confidence interval, by method. Vertical lines indicate the conventional 10%, 5% and 1% significance levels.