FEDERAL RESERVE BANK OF MINNEAPOLIS | Research Division

STAFF REPORT
No. 629

Comment on "Star Wars: The Empirics Strike Back"

November 2021

Adam Gorajek
Reserve Bank of Australia

Benjamin Malin
Federal Reserve Bank of Minneapolis

DOI: https://doi.org/10.21034/sr.629
Keywords: Researcher bias; Research credibility; Research replicability; Z-curve
JEL classification: A11, C13

Comment on "Star Wars: The Empirics Strike Back"

Adam Gorajek
Reserve Bank of Australia
gorajeka@rba.gov.au

Benjamin Malin
Federal Reserve Bank of Minneapolis
benjamin.malin@mpls.frb.org

November 2021°

Using a novel meta-analytical method, Brodeur et al. (2016) argue that hypothesis tests in top economic journals have exaggerated levels of statistical significance. Brodeur et al. (2020) apply the same method to another sample of hypothesis tests, obtaining similar results. We investigate the reliability of the method by highlighting questionable assumptions and compiling a dataset to examine their merits. Our findings support the original conclusions.

JEL Classification Numbers: A11, C13
Keywords: researcher bias, research credibility, research replicability, z-curve

1. Introduction

Brodeur et al. (2016) show that hypothesis tests in top economic journals produce an unusual bimodal distribution of z-scores. Using a novel decomposition, the authors argue that the bimodality stems from missing tests just below the 5 percent significance threshold ($|z| = 1.96$) and a corresponding surplus just above it. The surplus is attributed to "inflation", a term for researcher practices that increase measured levels of statistical significance in artificial ways. Follow-on work by Brodeur et al. (2020) applies the same decomposition to a newer sample and obtains similar results. If correct, both papers have concerning implications for the credibility of published economic research. We ask, How reliable is the decomposition?

Gorajek et al. (2021) offer preliminary answers to this question as part of an investigation into the credibility of central bank research. First, they filter out hypothesis tests for which the decomposition assumptions are least credible. The results suggest that problems can arise when applying the decomposition to unfiltered samples, like the one in Brodeur et al (2016). Second, they conduct a placebo test that searches for inflation in hypothesis tests about control variables. The result is inconclusive because their placebo sample is small, but it nonetheless raises concerns about the decomposition.

In this paper, we repeat the investigations of Gorajek et al. (2021), this time applying them to hypothesis tests from the same papers assessed by Brodeur et al. (2016) rather than to central bank discussion papers. The change increases the placebo sample size and helps answer the question of

° We thank Joey Pickens, Ji Sue Song and Caitlin Treanor for research assistance. We thank Joel Bank and Andrew Staib for helpful comments and James Holt for editorial assistance. The views in this paper are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Minneapolis, the Federal Reserve System, or the Reserve Bank of Australia.

whether the decomposition has merit in applications to top journals. To conduct our investigations, we add considerable detail to the Brodeur et al. (2016) dataset.[1]

Our results support the Brodeur et al. (2016) findings. First, filtering the Brodeur et al. (2016) dataset does not change their results, because the filter removes too few hypothesis tests to matter. Still, we suggest that future work use the filter, as it might matter for some datasets. Second, the placebo test shows no traces of inflation, as should be the case when the decomposition has merit. In fact, as Figure 1 shows, if we simply take the controls distribution as a gauge of what a trustworthy distribution of z-scores would look like, the same story told by the formal decomposition emerges: researchers produce a troubling surplus of marginally significant z-scores.

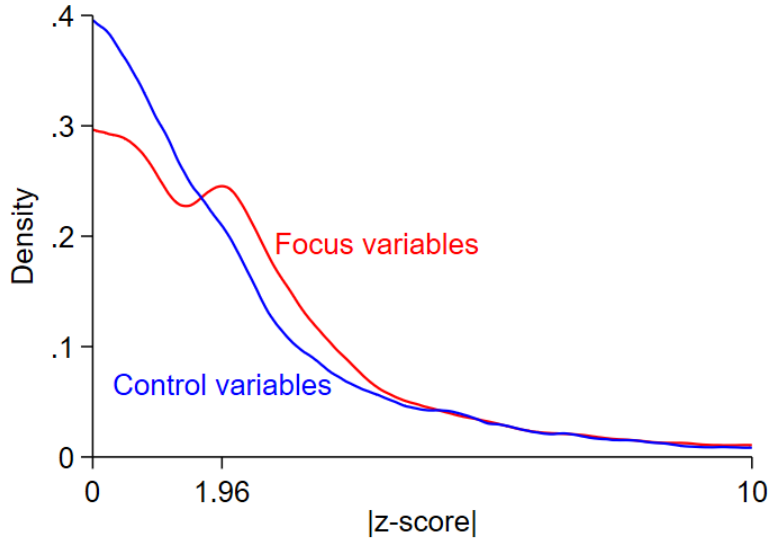

Figure 1: Distributions of z-scores for hypothesis tests in top economic journals

Notes: This figure plots kernel densities for the absolute unweighted values of t-statistics (very close to z-scores) published in the *American Economic Review*, *Journal of Political Economy*, and *Quarterly Journal of Economics*, from 2005 to 2011. If we take the controls distribution as a gauge of what trustworthy z-scores would look like, the comparison with focus variables is consistent with researchers pushing insignificant tests harder to pass the 5 percent significance threshold of |z| = 1.96. The distributions exclude tests that authors disclose as coming from data-driven model selection, or research portrayed as "reverse causal" (Gelman and Imbens, 2013), but these exclusions matter little.
Sources: Brodeur et al. (2016), *American Economic Review*, *Journal of Political Economy*, *Quarterly Journal of Economics*

To shield this analysis from inflation of our own, we pre-registered the analysis with the Open Science Framework.[2] Our online appendix includes the pre-registered analysis that has been omitted here.

2. Decomposition assumptions

On the surface, the decomposition introduced by Brodeur et al. (2016) looks complex. But Gorajek et al. (2021) explain that it can be understood as a simple application of Bayes' rule. We repeat elements of their explanation here (in places, we repeat them verbatim), as doing so is helpful for

1 We conduct our investigations on the papers examined by Brodeur et al. (2016) and not on those in Brodeur et al. (2020) because Brodeur et al. (2020) focus on papers using randomised control trials, difference-in-differences, instrumental variables, and regression discontinuity designs. For such papers, our first investigation becomes redundant, because no tests are caught by the filter. Only the placebo exercise would stand any chance of being useful.

2 See https://doi.org/10.17605/OSF.IO/58MNJ.

highlighting important assumptions. Their explanation also contains a stylised graphical example, which we do not repeat here.

Fundamentally, the decomposition aims to detect inflation in the probability distribution of test statistics that are the primary interest of research projects. Call these probabilities P[z], where z is the z-score equivalent of each test statistic of primary interest and is measured in absolute values. (Although z-scores are continuous variables, we use discrete variable notation to simplify the exposition.) The central challenge is that we observe only the z-scores that researchers publish. That is, we draw z-scores with probability P[z|published]. Following Bayes' rule, the P[z|published] distribution is a distorted version of the P[z] distribution, whereby

$$P[z|published] \propto P[published|z]P[z].$$

Brodeur et al. (2016) name the distorting term, P[published|z], "selection". It captures the fact that researchers are more likely to publish papers containing statistically significant test statistics (see Franco et al., 2014).

At a high level, the decomposition of Brodeur et al. (2016) examines whether unbiased candidates for P[z] and realistic forms of P[published|z] can explain estimates of P[z|published]. Unexplained variation in estimates of P[z|published], especially if systematic and concentrated near important significance thresholds, is attributed to inflation. The process works in four steps:

1. Identify a wide range of potential candidates for inflation-free forms of P[z]. For example, some of the candidates chosen by Brodeur et al. (2016) are empirical distributions that come from collating z-scores on millions of random regressions within major economic datasets. By construction, these distributions will be free of inflation and selection.

2. Select several preferred candidates for P[z], based on how well each one matches the estimated distribution of P[z|published] for values of z larger than 5. The "focus variables" series in Figure 1 is, in our case, the estimate for P[z|published] that the P[z] candidates need to match over this range. An underlying assumption in this step is that both inflation and selection should be "much less intense, if not absent" over these extreme values of z (Brodeur et al. 2016, p. 17).

3. For each of these preferred candidates of P[z], choose a corresponding P[published|z] that is increasing in z and best explains the estimated P[z|published], where "best" is determined by a least-squares criterion. The goal here is to explain as much of the estimated P[z|published] distribution as possible with plausible forms of selection.

4. Attribute unexplained variation in P[z|published] to inflation if it suggests a missing density of results just below the 5 percent significance threshold that can be retrieved just above it.

Brodeur et al. (2016) consider two concerns with this method. First, researchers in some contexts will favour null results, in which case P[published|z] need not be strictly increasing in z. Second, raw samples of test statistics, used to estimate P[z|published], will be clustered at the paper level. The authors produce supplementary analysis that, in our view, is effective in allaying both concerns.

Another potential concern, raised by Gorajek et al. (2021), is whether the method generates any suitable candidates for unbiased forms of P[z]; the true form would be the result of many interacting and unobservable factors, and using incorrect candidates would affect findings about inflation. One particular concern is that unfiltered samples of test statistics, like the one in Brodeur et al. (2016), will include research that is transparent about using data-driven model selection techniques, such as general-to-specific variable selection. Those techniques could plausibly generate a bunching of marginally significant results and contribute automatically to findings of inflation. Indeed, Leeb and Pötscher (2005) explain that common data-driven model selection techniques can distort test statistics in unpredictable ways, and Gorajek et al. (2021) find that test statistics associated with these techniques have concentrations of mass in the marginally significant zone. In our view, assuming that at least one of the P[z] candidates is accurate is the most questionable feature of the Brodeur et al (2016) decomposition.

3. Tests of the decomposition assumptions

We investigate these additional concerns with two tests.

The first test applies the decomposition to the same z-scores used by Brodeur et al. (2016), after filtering out the scores for which the P[z] candidates are least credible. In particular, we drop scores that authors disclose as coming from data-driven model selection techniques. We also drop scores coming from papers that are transparent about being "reverse causal", meaning they investigate the possible causes of observed outcomes (Gelman and Imbens, 2013). In those cases, data-driven model selection is often implied.

The second test is a placebo exercise that jointly examines the merits of all the decomposition assumptions. It applies the decomposition to z-scores on control variables taken from the same papers underlying the Brodeur et al. (2016) dataset. Since the statistical significance of control variables is not a selling point of research, the mechanisms that could motivate inflation should not be present. For this reason, Brodeur et al. (2016) did not include z-scores on control variables in their dataset. When we use the control z-scores for this test, we first drop those that do not necessarily apply to a specific economic hunch or theory, such as tests of fixed effects or time trends. We also drop those that authors disclose as coming from data-driven model selection or reverse causal research. If the decomposition is valid, we should not find traces of inflation in this placebo sample.

4. The new dataset

We have considerably enlarged the Brodeur et al. (2016) dataset. In particular, to enable our filter for data-driven model selection and reverse causal research, we have added classifications to each of the focus-variable observations already in the dataset. And to conduct the placebo test, we have added new observations about control variables, as well as the associated metadata.

Table 1 summarises the extended dataset, which contains almost 65,700 hypothesis tests from top journals.[3] About one-sixth of the hypothesis tests on focus variables come from either data-driven model selection or reverse causal research. In Gorajek et al. (2021), the corresponding figure in the

3 To facilitate future research, we also append to the final dataset in our replication files the data from Gorajek et al. (2021), which describe 14,800 hypothesis tests in central bank discussion papers.

dataset for central bank discussion papers is about one-third. This is consistent with central bankers needing to understand the causes of all major developments in their purview and thus insisting less than academics on pursuing tightly defined, "forward causal" research questions.

Our sample of hypothesis tests on control variables is eight times larger than the one in Gorajek et al (2021). It is similar in size to several of the subsamples investigated by Brodeur et al (2016).

Table 1: Summary statistics

Number of hypothesis tests, with shares of totals in parentheses

	Focus variables	Control variables
All Hypothesis tests	49,727	15,937
	[100]	[100]
Of which:		
- Portrayed as reverse-causal research	7,587	1,690
	[15]	[11]
- Disclosed as using data-driven model selection	395	42
	[1]	[0]
- Neither of the above	41,812	14,205
	[84]	[89]

Notes: These are counts of collated hypothesis tests published in the *American Economic Review*, *Journal of Political Economy*, and *Quarterly Journal of Economics*, from 2005 to 2011. We have slightly fewer tests on focus variables than Brodeur et al. (2016), because our reproduction of their work found some erroneous (as well as missing) entries. The effect of these changes on the results is immaterial.

Sources: Brodeur et al. (2016), *American Economic Review*, *Journal of Political Economy*, *Quarterly Journal of Economics*.

5. Results

The filters for data-driven model selection and reverse-causal research turn out to matter little. If anything, the filtered dataset shows a more exaggerated bimodal shape (Figure 2).

This result surprised us. In the central bank dataset compiled by Gorajek et al. (2021), the same filter materially changed the observed distribution of z-scores and reduced formal measures of inflation by about a third. The higher number of tests caught by the filters in that paper go some way to explain the difference. Sampling error is perhaps also playing a role, especially in the central bank dataset, which has 14,800 observations. At this point, we can only speculate.

The placebo test also supports the original conclusions of Brodeur et al (2016). Figure 1 shows that the distribution of z-scores on control variables is unimodal, and Table 2 shows that the results of the formal decomposition indicate negligible inflation. The number -0.5 in the first column of data technically reads as "assuming that the bias-free form of P[z] is the Student distribution with 1 degree of freedom, and P[published|z] is well estimated non-parametrically, there is an unexplained *shortage* of marginally significant results that amounts to 0.5 percent of all results in the marginally

significant zone of 2 < |z| < 4".[4] The figures in the focus variable columns, on the other hand, suggest material inflation; they are consistently positive and much larger in magnitude.

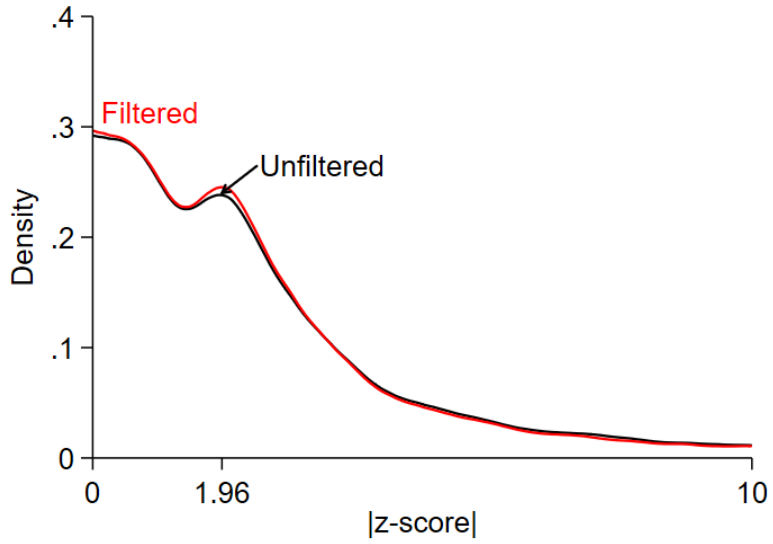

Figure 2: Distributions of z-scores for hypothesis tests on focus variables in top economic journals

Notes: This figure plots kernel densities for the absolute unweighted values of t-statistics (very close to z-scores) published from 2005 to 2011 in the *American Economic Review*, *Journal of Political Economy*, and *Quarterly Journal of Economics*. The filtered version strips out tests that authors disclose as coming from data-driven model selection and research portrayed as "reverse causal" (Gelman and Imbens, 2013). Clearly, the filter matters little in this application.
Sources: Brodeur et al. (2016), *American Economic Review*, *Journal of Political Economy*, *Quarterly Journal of Economics*

4 This presentation format differs from the one chosen by Brodeur et al. (2016), which is based around a measure of "maximum cumulated residuals". We depart from that because several of the maxima in the placebo decomposition occur well beyond the marginally significant zone (see Figure A7 in the Online Appendix), which is inconsistent with the Brodeur et al. (2016) description of inflation. The Online Appendix contains the results that use maximum cumulated residuals (Table A3). They still show that the decomposition performs well in the placebo test.

Table 2: Formal inflation estimates from decompositions

Excess z-scores in marginally significant zone (2 < |z| < 4) as a percentage share of all z-scores in marginally significant zone

	Control variables (placebo) sample		Focus variables sample	
Candidate for P[z]	Non-parametric estimation of P[published|z]	Parametric estimation of P[published|z]	Non-parametric estimation of P[published|z]	Parametric estimation of P[published|z]
Student (1)	-0.5	4.1	8.3	16.5
Cauchy (0.5)	-0.5	5.1	6.0	17.6
WDI	1.4	3.1	11.0	15.9
VHLSS	-1.8	-0.6	8.3	12.6
QOG	-0.7	0.8	7.4	13.6
PSID	-0.5	1.3	7.2	14.5

Notes: For all input functions, the estimates of inflation for the controls (placebo) sample are much lower than those for the focus variables sample, and they are typically close to zero. The focus variables' results are not identical to those presented in Table 2 of Brodeur et al. (2016), since here we have i) filtered the data for data-driven model selection and reverse causal research and ii) corrected their sample for some erroneous (as well as missing) entries.

Sources: Brodeur et al. (2016), *American Economic Review*, *Journal of Political Economy*, *Quarterly Journal of Economics*

6. Conclusion

While the work of Gorajek et al. (2021) raised concerns about the merits of the Brodeur et al. (2016) decomposition, our tests here suggest the decomposition is valid for the top-journals application. Our findings support ongoing moves to adopt research practices that, while often resource intensive, better preserve research credibility.

References

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

Brodeur A, M Lé, M Sangnier and Y Zylberberg (2016), `Star Wars: The Empirics Strike Back', *American Economic Journal: Applied Economics*, 8(1), pp 1–32.

Franco A, N Malhotra, and G Simonovits (2014), 'Publication Bias in the Social Sciences: Unlocking the File Drawer', *Science*, 345(6203), pp 1502–1505.

Gelman A and G Imbens (2013), 'Why Ask Why? Forward Causal Inference and Reverse Causal Questions', NBER Working Paper no. 19614. Accessed on 15 September 2019.
URL: https://www.nber.org/papers/w19614

Gorajek A, J Bank, A Staib, B Malin, and H Fitchett (2021), 'Star Wars at Central Banks', Reserve Bank of Australia Research Discussion Paper 2021-02.

Leeb H and B M Pötscher (2005), 'Model Selection and Inference: Facts and Fiction', *Econometric Theory*, 21(1), pp 21–59.