FEDERAL RESERVE BANK OF MINNEAPOLIS | Research Division

STAFF REPORT
No. 630

Online Appendix for: 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.630
Keywords: Researcher bias; Research credibility; Research replicability; Z-curve
JEL classification: A11, C13

ONLINE APPENDIX

Comment on "Star Wars: The Empirics Strike Back"

Adam Gorajek
Reserve Bank of Australia

Benjamin Malin
Federal Reserve Bank of Minneapolis

This appendix contains the pre-registered analysis for our comment on "Star Wars: The Empirics Strike Back" by Brodeur et al (2016). To structure the analysis, we reproduce the pre-registration; our results appear in red under each of the relevant parts. The time-stamped version of the pre-registration is available from the Open Science Framework website at the address https://doi.org/10.17605/OSF.IO/58MNJ.

To understand this appendix deeply, we recommend carefully reading Brodeur et al (2016). The body of our comment paper outlines only the intuition of their method. In some of the figures presented in this appendix, we use labels that differ from those in Brodeur et al. (2016), and we do so to more clearly connect to the intuition we offer.

A Comment on "Star Wars: The Empirics Strike Back" by Abel Brodeur, Mathias Le, Marc Sangnier, Yanos Zylberberg

PRE-ANALYSIS PLAN[1]

Hamish Fitchett
Reserve Bank of New Zealand

Adam Gorajek
University of New South Wales, Reserve Bank of Australia

Benjamin Malin
Federal Reserve Bank of Minneapolis

27 September 2020

[Updated with results in October 2021]

1.0 Research Question

Brodeur et al. (2016) develop a method to detect researcher bias, called the z-curve, and use it on a novel dataset covering papers in top econ journals. The results suggest that those papers do contain researcher bias. But how valid is the z-curve method?

Bank et al. (forthcoming) conduct two investigations into the merits of the method, using a dataset focussing on central bank discussion papers. First, they conduct a placebo test, which searches for researcher bias in hypothesis tests about control variables. The results are not definitive because the sample size is low, but they do cast doubt over the z-curve method. Second, the authors investigate potential problems with applying the method to papers that use data-driven model selection techniques. Those results also reveal potential problems with the z-curve method.

In this paper we will repeat the investigations of Bank et al. (forthcoming), this time using data about the same papers assessed by Brodeur et al. (2016). This is a useful point of difference because we expect to increase the placebo sample size dramatically. It is also unclear whether problems applying the z-curve method to central bank research would extend to assessments of top journals. To conduct our investigations we have to enlarge the dataset used by Brodeur et al. (2016), because they did not collect data on hypothesis tests of control variable parameters, nor did they collect information on the use of data-driven model selection.

In footnote 19, Brodeur et al. (2016) suggest another application for data on control variables if they are available: the data could be used to better understand what a distribution of test statistics might look like without researcher or publication bias. That is, they could be used to construct an "input function". We also follow that advice, recalculating the results of Brodeur et al. (2016) accordingly.

[1] The views expressed in this plan are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Minneapolis, the Federal Reserve System, the Reserve Bank of Australia, or the Reserve Bank of New Zealand. Use of any content from this plan should clearly attribute the content to the authors and not to the aforementioned institutions. The authors are solely responsible for any errors. Adam's contribution benefits from an Australian Government Research Training Scholarship.

2.0 Data

We will collect the data for the project from the same journal articles that make up the dataset of Brodeur et al. (2016). The attached collection instructions provide data definitions and detail the inclusion/exclusion criteria. In many places we have chosen to clarify the original instructions, where we have found them to be unclear. The clarifications were informed by the discussion in Brodeur et al. (2016).

We do not expect to drop any observations that meet these criteria. For any that look unusual, we will revisit the original paper to ensure there is no transcription error. Only if the original paper includes a definite error (unlikely we think) will we drop the observation, and be clear about each specifics of the decision in our paper.

While extending the original dataset of Brodeur et al. (2016), we discovered some erroneous entries, as well as some missing ones. The discoveries are detailed in the spreadsheet named "Brodeur_corrections.xlsx" in the raw data folder of our replication material. We decided to correct the dataset before using it. However, the corrections make an imperceptible difference to observed distributions of z-scores (Figure A1).

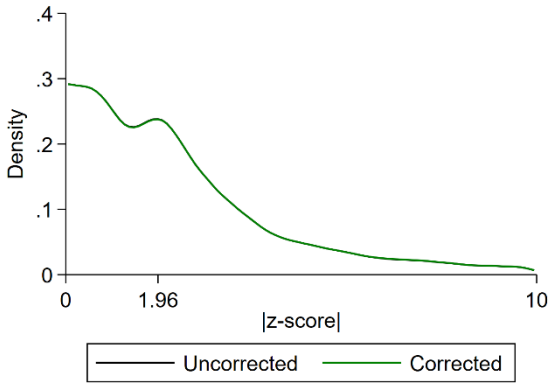


Figure A1: Corrected and uncorrected distributions of z-scores from Brodeur et al. (2016)
Focus variables

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

We do not know how large the resulting sample of control variable parameters will be. Focussing on central bank research, Bank et al. (forthcoming) collect about 15 control variable results for every 100 focus variable results. The same ratio would give us about 7000 control variable results for the top journals dataset. The z-curve method has no corresponding power calculations (or similar) to justify sample size, but Brodeur et al. (2016) do present results for subgroups of a similar size. Only for much smaller sample sizes do they note that their method is unreliable.

3.0 Presentation of Analysis

Three deviations from the Brodeur et al. (2016) presentational choices will persist throughout our analysis:

1. Whenever we present the data from Brodeur et al. (2016), we will omit their data on one-sided tests, to support a cleaner comparison.

2. Whenever we present kernel densities for the z-curve method, including when using the Brodeur et al. (2016) dataset, they will include a boundary adjustment.
3. We will always use z-scores that are unweighted and de-rounded, meaning we will smooth out anomalies relating to rounding problems, as per Panel B of Figure 1 in Brodeur et al. (2016). We will never present the weighted or rounded data, as they do. (Brodeur et al. 2016 favour their estimates that are based on the unweighted, de-rounded data.)

3.1 Summary Statistics

Our first presentation of the data will be a table of summary statistics, where we split the sample into controls vs other. Both subsamples will be split further into:

1. Those that have been disclosed as coming from a data-driven model selection process
2. Those that are disclosed as coming from reverse-causal research, which often implies data-driven model selection. (We will also include in this category straight forecasting research and general-equilibrium macroeconometric models. We expect eligible versions of these types of hypothesis tests to be rare.)
3. The ideal sample, which retains only forward causal research that does not use any data-driven model selection.

The data in the table will summarise the counts of test statistics in our sample, rather than, say, counts of papers.

Table A1: Summary statistics

Number of hypothesis tests, with shares of totals reported 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]

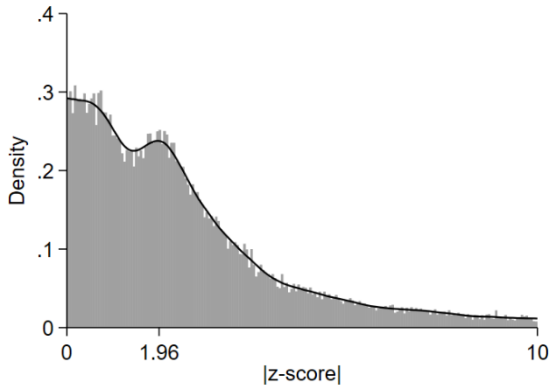
Notes: By "reverse causal" we mean research that searches for the possible causes of an observed outcome, as per Gelman and Imbens (2013). The opposite of this type of research is "forward causal" research, which studies the effects of pre-specified causes. The tests on "focus variables" are those that are discussed in the main text of a paper. We have slightly fewer tests on focus variables than appear Brodeur et al. (2016), since 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*.

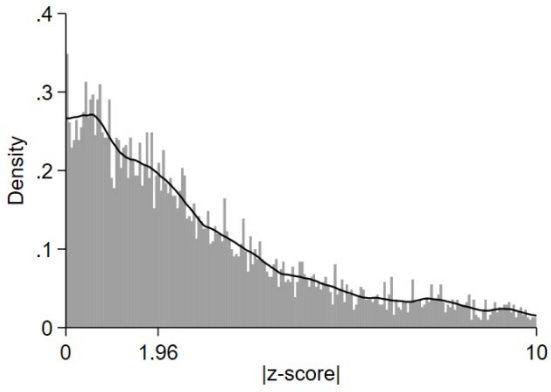
3.2 Concerns about data-driven model selection

We'll then show z-curves for each of the three groups in the summary statistics section. If the distributions are noticeably different, we will argue for focussing on the cleansed sample. Data-driven model selection could plausibly generate a mass of just-significant results, which could automatically generate findings of researcher bias.

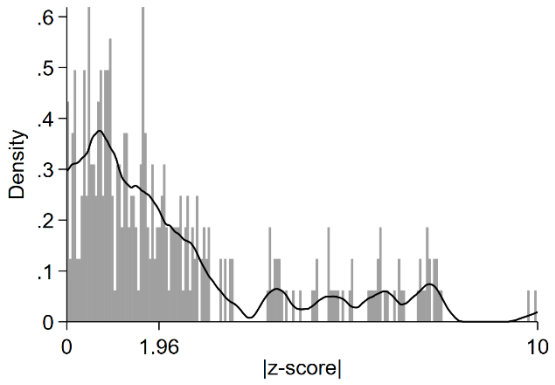
A: Tests on all focus variables



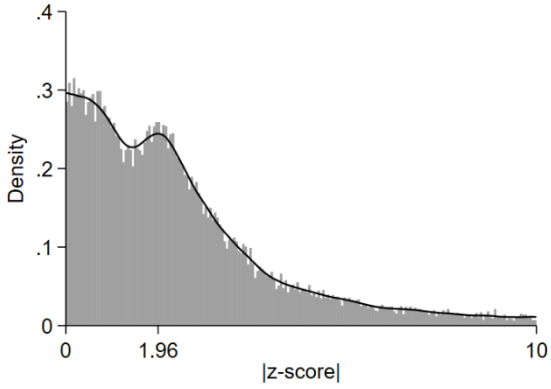
B: Tests on focus variables in research portrayed as "reverse causal"



C: Tests on focus variables disclosed as coming from data-driven model selection



D: Tests on all focus variables, excluding reverse causal research and data-driven model selection



E: Tests on focus variables, before and after filtering out those from reverse causal research and data-driven model selection

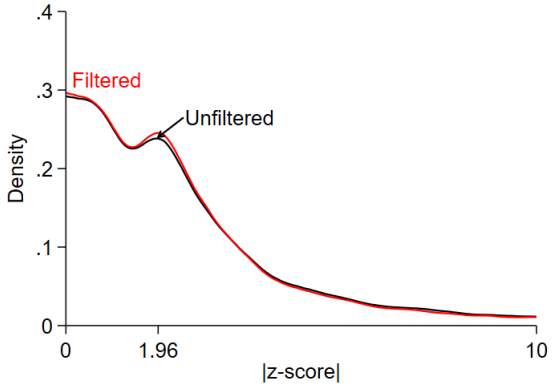


Figure A2: Distributions of z-scores for focus variables

Notes: The tests on "focus variables" are those that are discussed in the main text of a paper.
Sources: Brodeur et al. (2016), *American Economic Review*, *Journal of Political Economy*, *Quarterly Journal of Economics*

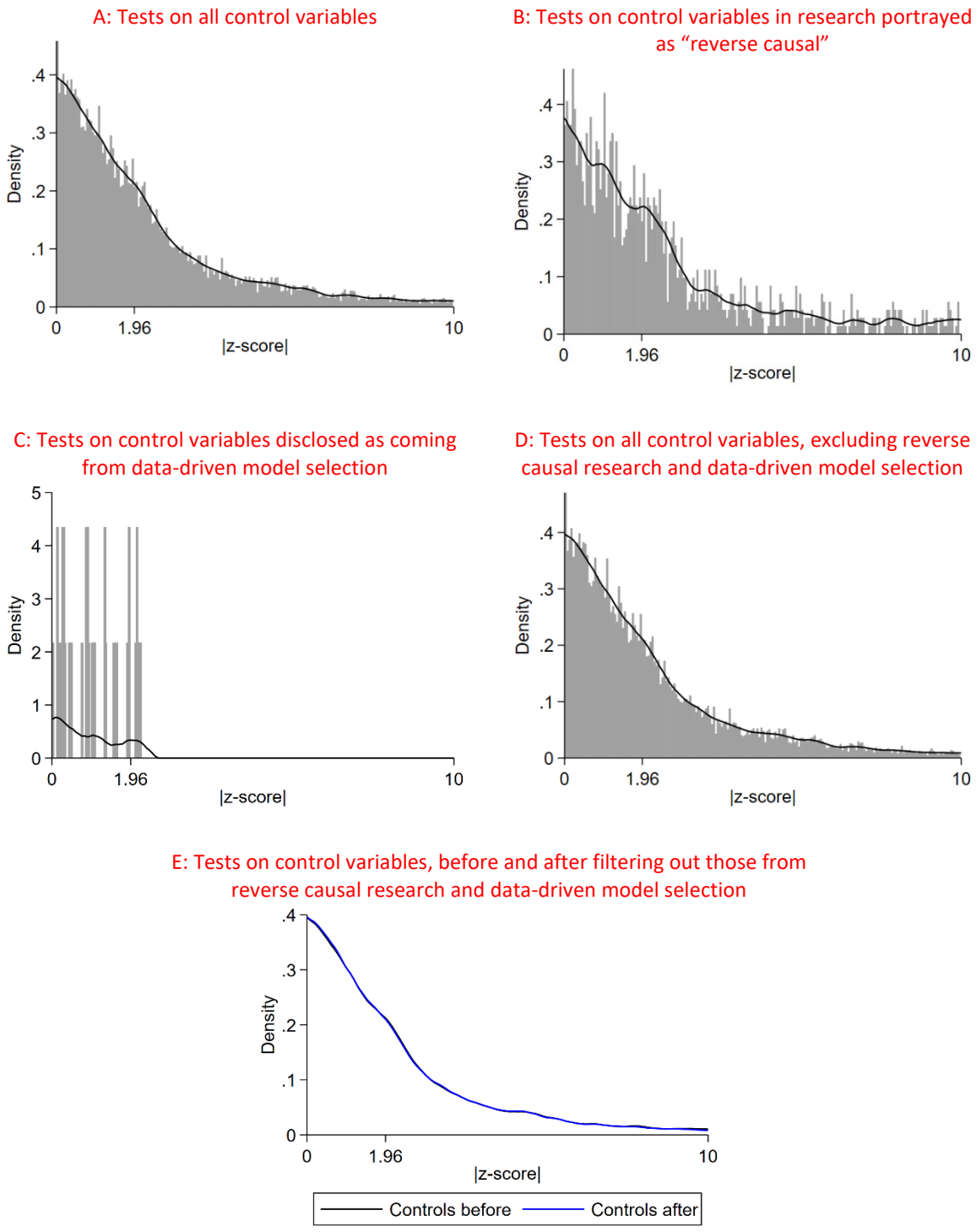


Figure A3: Distributions of z-scores for control variables

Sources: American Economic Review, Journal of Political Economy, Quarterly Journal of Economics

If these results do suggest that focussing on the ideal sample is appropriate, we would reproduce all of the results (tables and charts) in Brodeur et al. (2016), after dropping any observations that aren't in the ideal sample. The exception is the results that use weighted or rounded test statistics; we will not show those, to economise on space. Many of our results will probably go into a separate appendix.

The results (in Figures A2 and A3) show clearly that whether we use the unfiltered or filtered samples will make very little difference. The charts comparing the unfiltered and filtered distributions are remarkable for how similar the distributions are. Therefore, we have chosen not to reproduce all the results in Brodeur et al. (2016) with the filtered samples.

3.3 Using control variables as an input

Wherever Brodeur et al. (2016) produce results (tables and charts) that vary by input function, we will show the corresponding results using the control variables input function. Whenever we use the controls data, we will drop observations that come from data-driven model selection or reverse-causal research, using the same data classifications as for the main results.

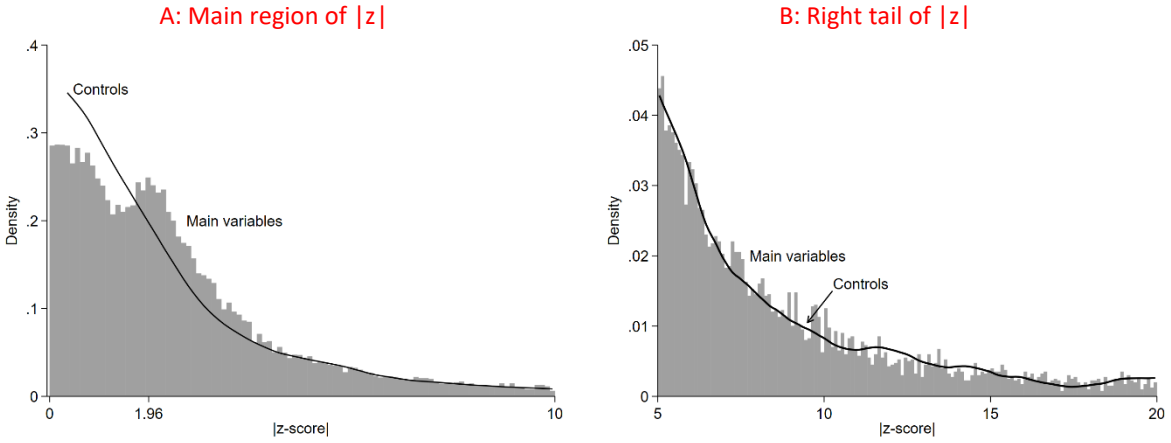


Figure A4: Distributions of z-scores on focus variables, against controls

Notes: The distributions exclude tests that authors disclose as coming from data-driven model selection techniques, as well as tests coming from research that authors portray as "reverse causal" (as per Gelman and Imbens, 2013). The effect of these omissions is very minor, as per section 3.2 of this appendix.
Sources: Brodeur et al. (2016), *American Economic Review*, *Journal of Political Economy*, *Quarterly Journal of Economics*

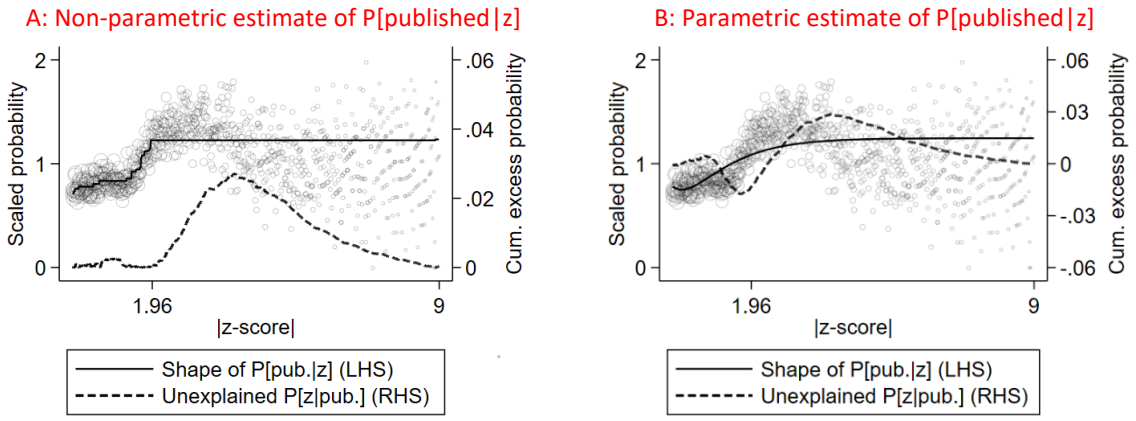


Figure A5: Unexplained variation in P[z|published] and our estimates for P[published|z], using the controls distribution as a proxy for bias-free P[z]

Notes: The graphs exclude tests that authors disclose as coming from data-driven model selection techniques, as well as tests coming from research that authors portray as "reverse causal" (as per Gelman and Imbens, 2013). The effect of these omissions is very minor, as per section 3.2 of this appendix.
Sources: Brodeur et al. (2016), *American Economic Review*, *Journal of Political Economy*, *Quarterly Journal of Economics*

The maxima of the dotted lines in both panels of Figure A5 are the counterparts to the formal inflation estimates shown in Table 2 of Brodeur et al (2016). Our maxima are very close to these inflation estimates, meaning that this extension—to use the controls distribution as a candidate for P[z]—has not materially changed conclusions.

Brodeur at al. (2016) also present formal inflation estimates for various subsamples in their Table 3. Unsurprisingly, our extension has not affected those results much either. Table A2 below shows the comparison.

Table A2: Formal subsample inflation estimates

	Maximum cumulated residuals			
	Controls distribution as P[z] candidate		WDI distribution as P[z] candidate	
Subsample	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]
Macroeconomics	0.030	0.034	0.038	0.044
Microeconomics	0.028	0.028	0.039	0.037
Positive results	0.028	0.030	0.039	0.040
Null results	0.094	0.005	0.084	0.011
Eye-catchers	0.032	0.033	0.043	0.043
No eye-catchers	0.019	0.021	0.029	0.031
Central results	0.023	0.026	0.033	0.035
Non-central	0.039	0.037	0.049	0.047
With model	0.015	0.015	0.020	0.025
Without model	0.035	0.034	0.045	0.044
Low average PhD-age	0.043	0.043	0.055	0.053
High average PhD-age	0.016	0.015	0.022	0.025
No editor	0.030	0.030	0.042	0.040
At least one editor	0.026	0.028	0.036	0.038
No tenured author	0.037	0.038	0.048	0.047
At least one tenured author	0.019	0.020	0.030	0.031
Single-authored	0.045	0.042	0.056	0.052
Co-authored	0.023	0.026	0.033	0.035
With research assistants	0.032	0.033	0.042	0.043
Without research assistants	0.021	0.022	0.033	0.033
Low number of thanks	0.022	0.024	0.030	0.033
High number of thanks	0.033	0.033	0.044	0.043
Data and codes available	0.031	0.032	0.042	0.042
Data or codes not available	0.026	0.027	0.035	0.037
Lab experiments or RCT data	0.055	0.038	0.064	0.047
Other data	0.024	0.027	0.034	0.037

Notes: Using the controls distribution as a candidate for P[z] produces similar—albeit usually slightly smaller—estimates of subsample inflation as using the empirical WDI distribution as a candidate for P[z]. The WDI results are not identical to those presented in Brodeur et al.'s (2016) Table 3, 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*

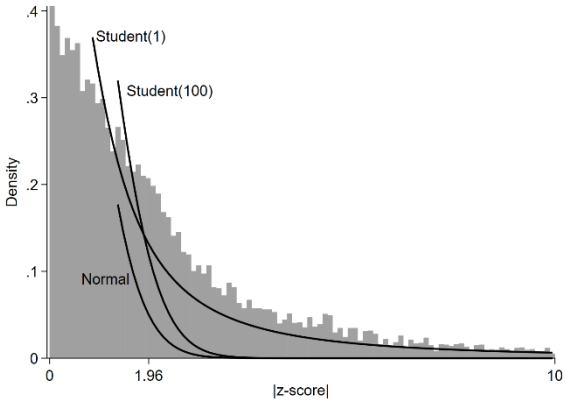
3.4 Conducting our placebo test

We will apply the z-curve method, exactly as applied in Brodeur et al. (2016), but on the controls data (again omitting observations that come from data-driven model selection or reverse-causal research). We will produce all of the same tables and charts, except that we won't show any results about subsamples of the control variables. We won't explore any of their weighting or rounded test statistic variations either.

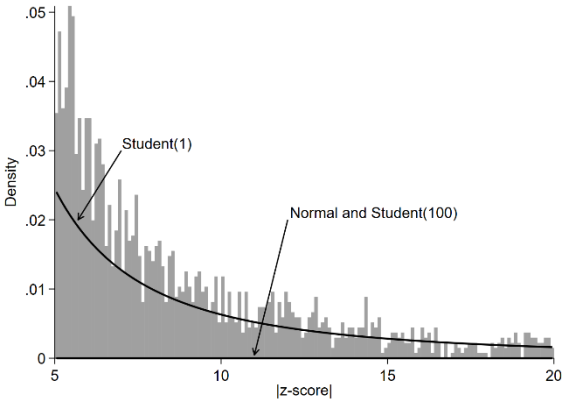
Brodeur et al. (2016) include as candidates for P[z] empirical distributions that come from collating test statistics on millions of random regressions within four economic datasets: the World Development Indicators (WDI), the Quality of Government (QOG) dataset, the Panel Study of Income Dynamics (PSID) and the Vietnam Household Living Standards Survey (VHLSS). These distributions of random test statistics will be free of researcher bias by construction. Other candidate distributions are parametric and include various Student-t and Cauchy forms.

The results of our placebo exercises show that the empirical candidates for P[z], especially the distributions from the WDI, QOG, and VHLSS datasets, closely match the controls distribution. Moreover, irrespective of which candidate P[z] is used, the formal estimates of inflation for the placebo sample are much lower than those for the focus variables sample and typically close to zero (Table A3).

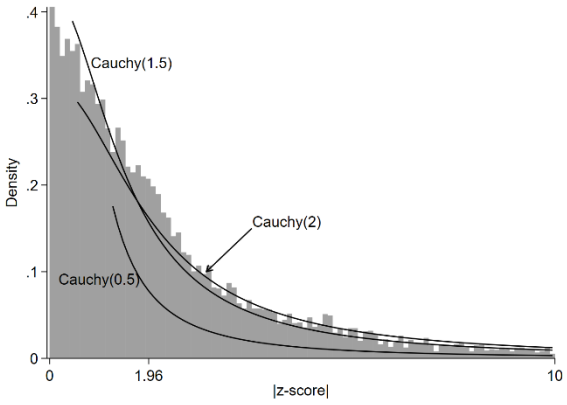
A: Main region of |z|, Student candidates for P[z]



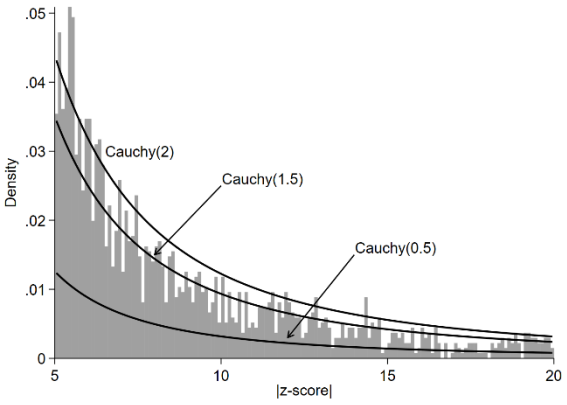
B: Right tail of |z|, Student candidates for P[z]



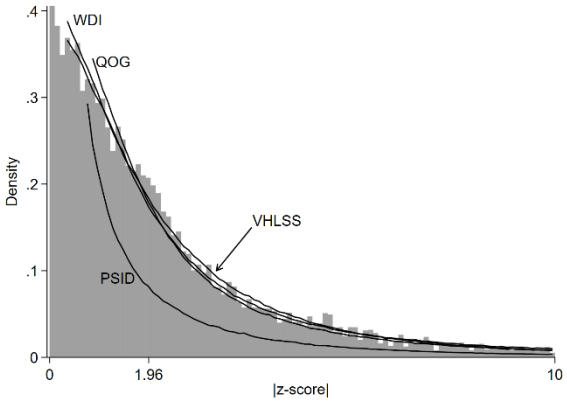
C: Main region of |z|, Cauchy candidates for P[z]



D: Right tail of |z|, Cauchy candidates for P[z]



E: Main region of |z|, empirical candidates for P[z]



F: Right tail of |z|, empirical candidates for P[z]

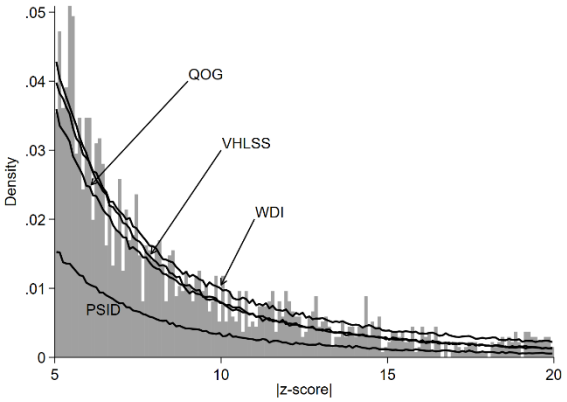


Figure A6: Distributions of z-scores on control variables, against potential candidates for P[z]

Notes: The control variables distribution excludes tests that authors disclose as coming from data-driven model selection techniques, as well as tests coming from research that authors portray as "reverse causal" (as per Gelman and Imbens, 2013).

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

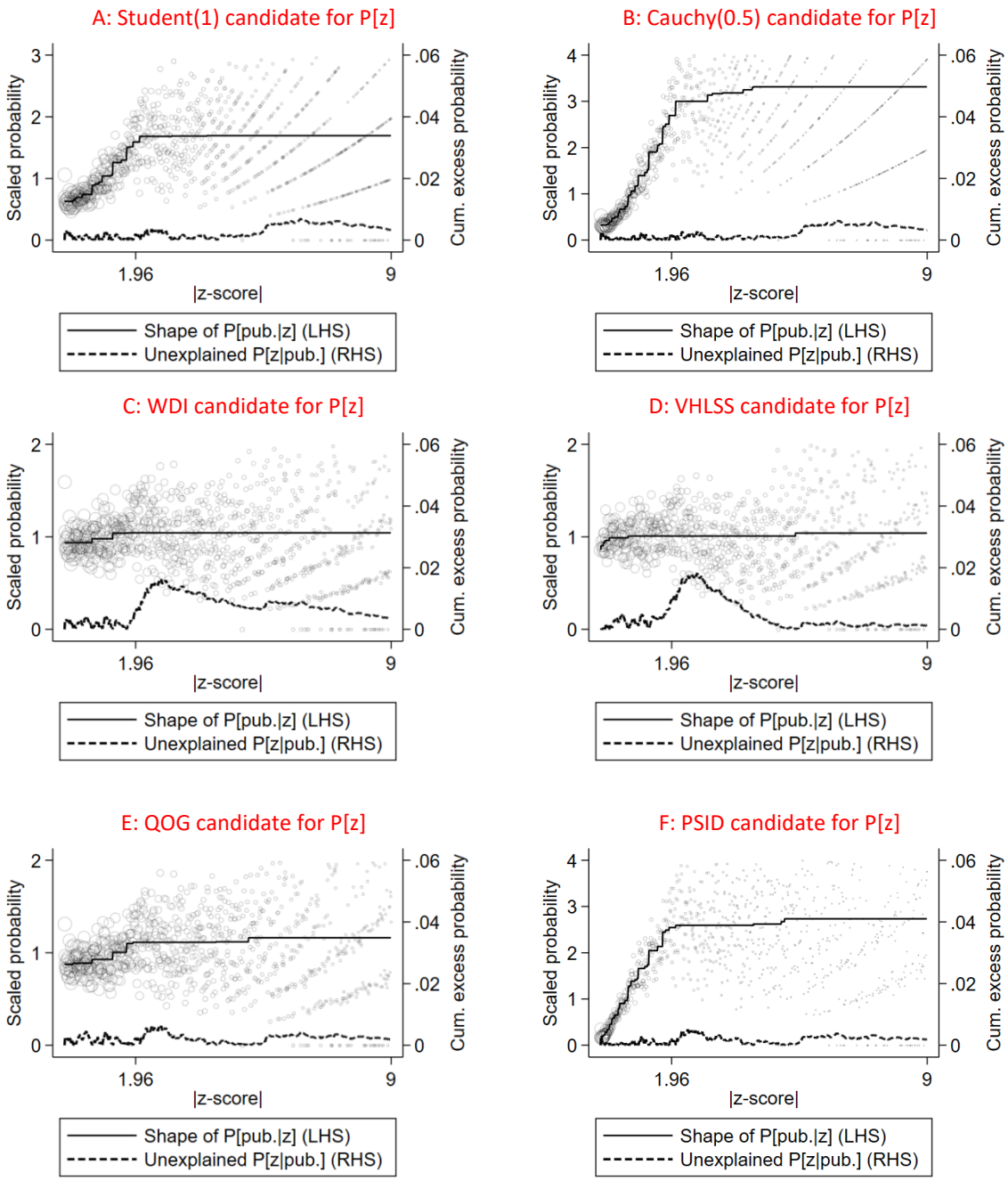


Figure A7: Unexplained variation in P[z|published] and our non-parametric estimates for P[published|z], controls sample

Notes: The graphs exclude tests that authors disclose as coming from data-driven model selection techniques, as well as tests coming from research that authors portray as "reverse causal" (as per Gelman and Imbens, 2013).
Sources: Brodeur et al. (2016), *American Economic Review*, *Journal of Political Economy*, *Quarterly Journal of Economics*

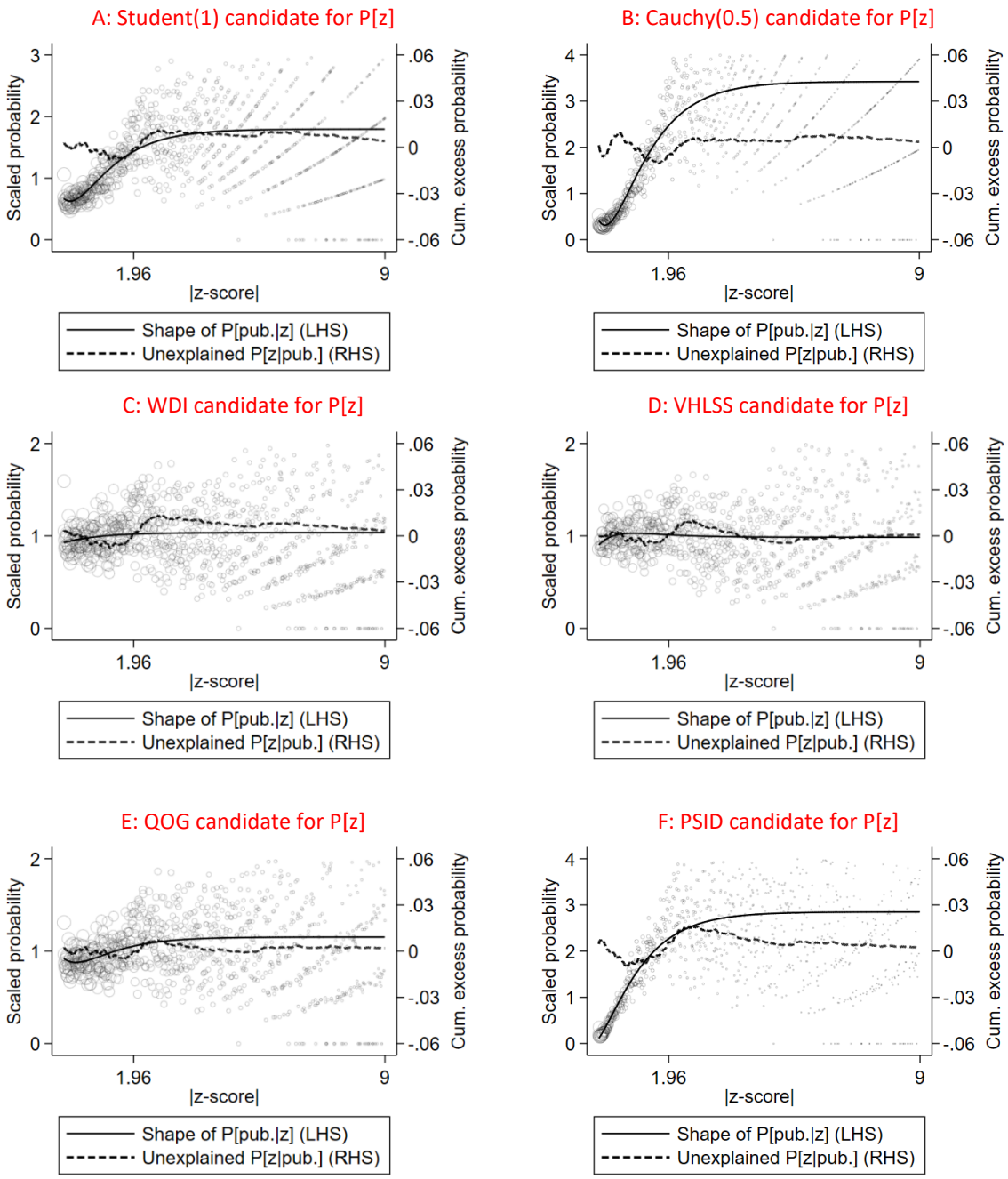


Figure A8: Unexplained variation in P[z|published] and our parametric estimates for P[published|z], controls sample

Notes: The graphs exclude tests that authors disclose as coming from data-driven model selection techniques, as well as tests coming from research that authors portray as "reverse causal" (as per Gelman and Imbens, 2013).
Sources: Brodeur et al. (2016), *American Economic Review*, *Journal of Political Economy*, *Quarterly Journal of Economics*

Table A3: Formal inflation estimates from decompositions

	Maximum cumulated residuals			
	Controls (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)	0.008	0.011	0.027	0.035
Cauchy (0.5)	0.007	0.015	0.019	0.042
WDI	0.015	0.013	0.038	0.039
VHLSS	0.017	0.010	0.030	0.029
QOG	0.006	0.007	0.024	0.029
PSID	0.005	0.016	0.023	0.033

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 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*

4.0 Expected timeline.

- Collect all data: by 31 December 2020
- Complete first draft: by 28 January 2020

5.0 Unplanned analysis

Table 2 in our write-up of the main paper uses a presentation format that differs from the one chosen by Brodeur et al (2016). In particular, we show excess z-scores in the marginally significant zone ($2 < |z| < 4$) as a percentage share of all z-scores in the marginally significant zone. Brodeur et al. (2016), on the other hand, show a measure of "maximum cumulated residuals". We depart from that approach because several of the maxima in the placebo decomposition occur well beyond the marginally significant zone (see Figure A7 above), which is inconsistent with Brodeur et al.'s (2016) description of inflation. Table A3, above, shows the equivalent results that use maximum cumulated residuals. They still show that the decomposition performs well in the placebo test.

References

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.

Bank J, H Fitchett, A Gorajek, B Malin, A Staib (forthcoming), 'Star Wars at Central Banks', Discussion paper jointly release by the Reserve Bank of Australia, Reserve Bank of New Zealand, and the Federal Reserve Bank of Minneapolis. [Note that the author ordering changed before release.]

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