ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Naguib, Costanza

Working Paper

P-hacking and significance stars

Discussion Papers, No. 24-09

Provided in Cooperation with:
Department of Economics, University of Bern

Suggested Citation: Naguib, Costanza (2024) : P-hacking and significance stars, Discussion Papers, No. 24-09, University of Bern, Department of Economics, Bern

This Version is available at:
https://hdl.handle.net/10419/308751

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

https://creativecommons.org/licenses/by/4.0/

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

UNIVERSITÄT
BERN

Faculty of Business, Economics
and Social Sciences

Department of Economics

P-hacking and Significance Stars

Costanza Naguib

24-09

October, 2024

DISCUSSION PAPERS

Schanzeneckstrasse 1
CH-3012 Bern, Switzerland
http://www.vwi.unibe.ch

P-hacking and Significance Stars

Costanza Naguib*

Abstract

In mid-2016, all journals of the American Economic Association (AEA) stopped including significance stars in their regression tables. This policy aimed to reduce the emphasis on statistical significance and shift focus toward the broader economic importance of research findings. This study examines the impact of this change on p-hacking and publication bias. The findings indicate some reduction in the bunching of the reported test statistics just beyond the 5%-significance threshold in the treated journals after 2016. However, the effect is modest.

JEL codes: A11, A14, C13

Keywords: p-hacking, significance stars, publication bias, difference-in-difference

1 Introduction

[1]Increasing pressure on researchers to publish in top-tier journals may lead to selective reporting of positive results, prioritizing statistical significance over comprehensive analysis (see e.g. Abadie (2020), Imbens (2021)). This focus on "significant" results, fueled by incentives tied to publication success, has been extensively documented as contributing to biased research outputs (see Ioannidis and Doucouliagos (2017) or Brodeur et al. (2020)). Editors and reviewers often favor studies with positive findings, further exacerbating this phenomenon (Ashenfelter and Greenstone (2004), Stanley (2008)). Such practices—whether through specification searching or preferential treatment of positive results—risk increasing the prevalence of false positives in the literature, skewing the research presented to policymakers. P-hacking and publication bias may undermine the evaluation of public policies, by presenting an inaccurate picture of policy effectiveness to the decision-makers.

In response to this relevant issue, and following the statement by the American Statistical Association on statistical significance and p-values (7 march 2016), in mid-2016

*University of Bern

[1]I thank Reto Horst and Thibaud Laurent for the excellent research assistance. I am also thankful to Jean-Michel Benkert, Kai Gehring, Blaise Melly and Eric Strobl for the insightful comments.

all the journals of the American Economic Association (AEA) stopped reporting significance stars in their regression tables[2]. By removing the emphasis on whether a result crosses arbitrary significance thresholds, the AEA aimed at alleviating the pressure on researchers to report "significant" findings. Instead, the focus is shifted towards evaluating the magnitude and robustness of estimates. This star-omission policy should encourage more balanced reporting, reducing the incentives for selective reporting and fostering an environment where both positive and null results are deemed worth to publish.

The aim of this paper is to evaluate the impact of the AEA's star-omission policy on p-hacking and publication bias. P-hacking refers to practices, such as specification searching, that researchers may use, consciously or unconsciously, to obtain more favorable p-values, often as a response to the challenges of publishing null results. Publication bias, on the other hand, means that papers reporting null findings are less likely to be published than those showing statistically significant results, as null findings are often perceived to be of lower quality (Imbens (2021), Chopra et al. (2024))[3]. Using a new dataset and a novel identification strategy, I investigate whether this star-omission policy has effectively reduced these distortions in the research process[4]. I exploit a quasi-natural experiment, in which all the journals from the American Economic Association stopped publishing significance stars in mid-2016[5]. Other leading journals in economics such as the *Journal of the European Economic Association* or the *Economic Journal* continue up to date to use significance stars in their published regression tables, hence providing a suitable control group for the analysis.

First, I apply a difference-in-difference approach to determine whether the star-omission policy caused a reduction in the share of statistically significant test statistics being pub-

[2]In the field of psychology, in 2014 the Editor of *Basic and Applied Social Psychology* went further and essentially banned the use of significance levels, including p-values and confidence intervals (Imbens (2021)). However, this policy did not have the desired effects, but rather led some authors to overstate the conclusions of their studies beyond what data would support (Imbens (2021)). In economics, *Econometrica* also has in its style guide for authors: "Please do not use asterisks or bold face to denote statistical significance", source: https://www.econometricsociety.org/publications/econometrica/information-authors/instructions-preparing-articles-publication. Similarly, the style guide of the *American Economic Review*, as well as those of each of the *American Economic Journals* read: "Do not use asterisks to denote significance of estimation results. Report the standard errors in parentheses" (source: https://www.aeaweb.org/journals/aer/style-guide).

[3]According to Imbens (2021): "In empirical studies, estimates with one, two, or three stars are often viewed as superior to those without such adornments".

[4]Differently from Brodeur et al. (2023), I do not aim at disentangling p-hacking from publication bias, but to assess whether the non-reporting of significance stars has an impact on both these phenomena (jointly considered).

[5]The first issue without significance stars appeared either in July 2016 or in August 2016, depending on the issue calendar of each journal. The journals affected from this policy are: *American Economic Review*, *American Economic Journal: Macroeconomics*, *American Economic Journal: Microeconomics*, *American Economic Journal: Economic Policy* and *American Economic Journal: Applied Economics*.

lished. Second, I investigate whether a battery of tests can still detect the presence of p-hacking and publication bias in the treated group after the star-omission policy entered into force. Although the policy appears to have slightly increased the number of published test statistics that do not reject the null hypothesis, this effect is not statistically significant at conventional levels. When focusing on the treated journals, only one of the six tests proposed by Elliott et al. (2022) rejects the null hypothesis of no p-hacking and no publication bias in the post-intervention period, compared to three out of six tests in the pre-intervention period[6]. This suggests that the policy (moderately) reduced the extent of p-hacking and publication bias.

The contribution of this paper is twofold: I am the first to evaluate the impact the removal of the significance stars from the regression tables on the desired goal of switching the focus from statistical significance to the broader economic significance of the results. This implies analyzing whether the intervention caused a reduction in the extent of p-hacking and publication bias in published articles. Second, in order to answer to this question, I put together a unique dataset of test statistics collected from six leading journals in economics for the period 2011-2023, including 50,610 test statistics from 3,768 articles.

Empirical researchers have often overemphasized statistical significance (Imbens (2021)). Omitting significance stars from regression tables in published papers can help reducing p-hacking and publication bias by shifting the focus away from arbitrary thresholds of statistical significance (e.g. $p < 0.05$). Significance stars may encourage researchers and readers to prioritize whether results are labeled as "significant" rather than considering the overall meaning or size of the estimated effects. This emphasis can incentivize p-hacking—looking for different model specifications until significant results are achieved—to obtain the desired stars. By removing these visual cues, the attention moves toward interpreting the actual estimates and their confidence intervals, promoting a more nuanced understanding of the data (Imbens (2021)). This practice can diminish the undue pressure to produce significant results, thereby reducing p-hacking and helping to mitigate publication bias against non-significant findings.

This study contributes to the expanding literature on empirical analyses of p-hacking. In their pioneering work, Brodeur et al. (2016) document the presence of p-hacking and publication bias in a group of journals constituted by three of the Top Five journals in economics: the *American Economic Review*, the *Journal of Political Economy*, and the

[6]Conversely, among the control journals, more tests detect p-hacking in the post-intervention period than before.

Quarterly Journal of Economics. Further, Brodeur et al. (2020) find that the extent of p-hacking and publication bias varies notably by estimation method. Kranz and Pütz (2022) show that the results of Brodeur et al. (2020) are mostly confirmed after accounting for rounding error. Furthermore, Brodeur et al. (2024a) find that the extent of p-hacking and publication bias does not relevantly vary in function of the data availability and replication policies adopted by journals. Additionally, Brodeur et al. (2024b) find that pre-registration and pre-analysis plans usually do not help alleviate the problem of p-hacking, as they are often not detailed enough.

The paper that is closest in spirit to the present one is Blanco-Perez and Brodeur (2020), in which the authors aim to assess the impact of the editorial statement released by the editors of eight health economics journals in February 2015 on the extent of p-hacking and publication bias. In this statement, the editors encouraged authors to submit, and reviewers not to reject, papers with economic relevance but statistically insignificant results[7]. Blanco-Perez and Brodeur (2020) find that this intervention was effective in reducing publication bias and p-hacking. More specifically, the authors find that the editorial statement decreased the share of published tests rejecting the null hypothesis by around 18 percentage points.

The present study expands on this line of research by considering a different policy—the omission of significance stars—that involved leading general-interest journals in economics. I find that the star-omission policy reduced the amount of p-hacking and publication bias in published articles in the treated group (as confirmed by the battery of statistical tests proposed by Eliott et al. (2022)). However, the impact of the policy on the probability that a published test statistic is significant was small (around 1/9 of the effect of the editorial statement studied by Blanco-Perez and Brodeur (2020)) and insignificant at the conventional levels.

2 Data and methods

I consider a treated group constituted by *American Economic Journal: Macroeconomics*, *American Economic Journal: Economic Policy* and *American Economic Journal: Applied Economics*. I exclude from the analysis *American Economic Journal: Microeconomics*, as: i) most papers published there are theoretical in nature and do not include estimated coefficients, and ii) this journal is ranked notably lower than the others, i.e. it is at rank

[7]There is no overlap between the journals affected by the Editorial statement studied by Blanco-Perez and Brodeur (2020) and those impacted by the star-omission policy analyzed in the present paper.

93 according to the RePEc simple impact factor list averaged over the last 10 years[8]. Following Brodeur et al. (2024a), I focus exclusively on journals that are among the first 25 positions in the RePEc simple impact factor list mentioned above. Further, I exclude from the analysis the *American Economic Review* as it is one of the so-called "Top Five" journals in economics. Top five and not top five journals may have reacted differently to the introduction of the star-omission policy[9]. By considering leading journals in economics, but not top fives, my estimated impact can be more easily generalized, as one could expect a similar impact to take place in case another leading journal in economics would decide to apply the same star-omission policy.

As control group, I select the journals immediately above and below each of the three *American Economic Journals* considered, according to the RePEc ranking of the simple impact factor averaged over the last ten years. This is the same ranking used by, e.g. Brodeur et al. (2024a). *American Economic Journal: Macroeconomics* and *American Economic Journal: Applied Economics* are respectively at places 9 and 10 in this ranking, hence I include in the control group the number 8 (*The Economic Journal*). I exclude however number 11 and 12 (*Annual Review of Economics* and *Journal of Economic Literature*) as they rarely, if ever, include any estimated coefficients and for this reason they have also been excluded by Brodeur et al. (2020). Further, since the *American Economic Journal: Economic Policy* is at place 21 in the RePEc ranking, I also include in the control group the *Journal of the European Economic Association* (number 20) and *Economic Policy* (number 22).

To summarize, my treated group includes the following journals: *American Economic Journal: Macroeconomics*, *American Economic Journal: Economic Policy* and *American Economic Journal: Applied Economics*. The control group includes the *Journal of the European Economic Association*, *The Economic Journal* and *Economic Policy*. The period of analysis ranges from 2011 (i.e. five years before the star-omission policy became effective) and 2023, i.e. the last full year available.

I exclude corrigenda, comments and replies to research papers from the analysis. I further exclude papers that do not include any estimated coefficient. Following Brodeur et al. (2020), I only collect estimates from results tables and only for the coefficients of interest, or main results, excluding regression controls, constant terms, balance and ro-

[8]Source: https://ideas.repec.org/top/top.journals.simple10.html, last retrieved on 20th September 2024.

[9]Figures 10 and 11 in the Appendix (based on the random sample of articles collected by Brodeur et al. (2024a) show a notable reduction in bunching of the z-statistics right after the critical threshold of 1.96 after 2016 for the *American Economic Journals*, but not for the *American Economic Review*.

bustness checks, heterogeneity of effects, and placebo tests. I however collect coefficients drawn from multiple specifications of the same hypothesis. Moreover, I collect the estimated coefficients of interaction terms only if such terms are the variable of interest, for example in the case of the interaction between the post-treatment period and the treated dummy in a difference-in-difference setup. If the main findings of a paper are expressed by means of a Figure, e.g. impulse response functions, then I drop the paper from the collection. If the declared aim of an article is to identify the drivers of a certain outcome variable without focusing on a specific explanatory variable or on a couple of explanatory variables of interest, then I drop the article as well[10].

I collect all reported decimal places. If more than one standard error per estimated coefficient is reported (e.g. obtained with different methods of clustering), I only collect the first one. Differently from Brodeur et al. (2024a) I do not collect a random sample of articles, but rather collect all the articles from the selected journals that report at least one estimated coefficient with its standard error (or a t-statistic, or a p-value) in the period 2011-2023.

Further, differently from Brodeur et al. (2020), I do not restrict the analysis to articles that use one of a pre-defined set of estimation methods (DID, RDD, RTC, IV), but I collect results from all methods that produce estimated coefficients and standard errors (or t-statistics, or p-values). Notably, this means that I also include OLS estimates. However, if OLS estimates are only used to present correlations or as a sort of descriptive statistics, and hence they are not in the Results Section of the paper, but rather in the Data description, then I do not collect them. In case of IV estimations, following Brodeur et al. (2020), I only collect the coefficient(s) of the instrumented variable(s) in the second stage.

Data have been coded independently by at least two of the following: the author and two research assistants. We discussed and clarified discordant cases. Additionally, I have checked the data used in the present paper against the data from Brodeur et al. (2024a). For the articles that are included in both samples[11], the same coefficients and standard errors (or t-statistics, or p-values) are collected. Finally, following Brodeur et al. (2020), since all of the test statistics in the sample relate to two-tailed tests and degrees of freedom are not always reported, I treat coefficient and standard error ratios as if they follow an asymptotically standard normal distribution. When articles report t-statistics or p-values, I transform them into equivalent z-statistics.

[10] This is sometimes the case especially in articles from *Economic Policy*.

[11] There is overlap for 211 articles, including 7,336 coefficients.

3 Results

3.1 Descriptive evidence

In this Section, I aim at giving a first insight into the data. From Figure 1, I deduce that the share of reported test statistics that are significant at the 5% level was initially (e.g. in 2011/2012) rather close between the control and the treated group, ranging between around 65 and 70%. After some oscillations, especially in the treatment group, in the years 2015-2018, the journals in the treatment group report in the years 2018-2021 on average a smaller share of significant test statistics than the journals in the control group. However, this difference, of around 5-10%, disappears over time. In 2023, the last year in the sample, the share of published test statistics that are significant at a 5% level is essentially identical in the two groups (around 60%). In the Appendix (Figures 6 and 7), I present the same graph, but with the share of test statistics significant at 1% and at 10%. Those patters are very similar to the ones presented here.

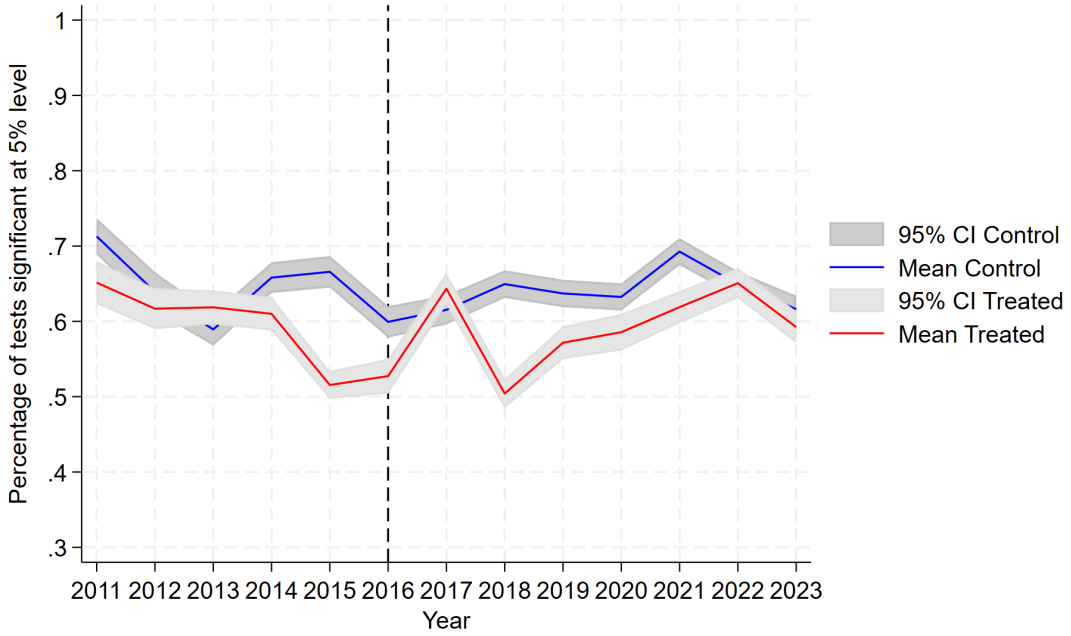

Figure 1: Percentage of tests significant at the 5% level, by year of publication. Treated journals include *AEJ: Applied Economics*, *AEJ: Macroeconomics* and *AEJ: Economic Policy*. Control journals include the *Journal of the European Economic Association*, *The Economic Journal* and *Economic Policy*.

From Figure 2, I get further insights into the changes in the distribution of the test statistics after the star-omission policy entered into force. In the control group (upper

panel) one witnesses a (slightly) larger amount of bunching immediately after the critical value of 1.96 in the period 2017-2023, than in the previous period, 2011-2015.

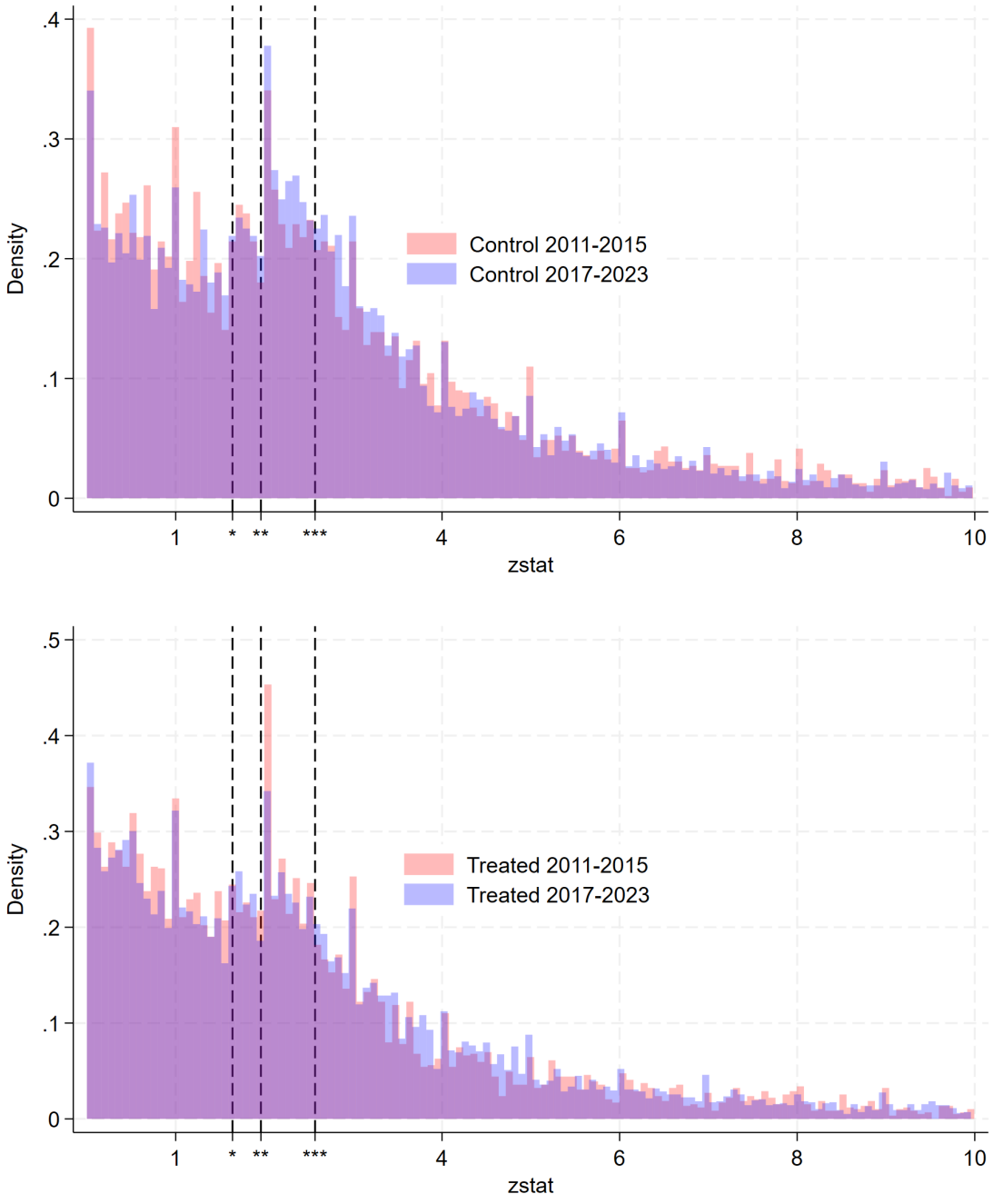

Figure 2: Histogram (125 bins) of the z-statistic values collected from journals in the control group (upper panel) and in the treatment group (lower panel), comparison of the periods 2011-2015 and 2017-2023. In the control group, $N = 6,952$ for period 2011-2015 and $N = 16,420$ for 2017-2023. In the treated group, $N = 7,366$ for 2011-2015 and $N = 12,293$ for 2017-2023. Z-statistics larger than 10 have been trimmed in order to improve graph readability.

On the contrary, from the lower panel of Figure 2, one notices that in the treated group the amount of bunching immediately after the critical value of 1.96 diminished from the 2011-2015 period to the subsequent 2017-2023 period. This provides a first hint that the star-omission policy reduced the amount of p-hacking and publication bias in the treated journals. Figures 4 and 5 in the Appendix show a comparison of the distribution of the z-statistics in the control and in the treated group in each of the years included in the sample. Further, Table 3 in the Appendix provides summary descriptive statistics for the whole sample.

For completeness, in Figure 10 in the Appendix I show the distribution of z-statistics in *American Economic Review* (data are from the random sample collected by Brodeur et al. (2024a)) in the period 2011-2015 and 2017-2020. In this case there is no visual evidence of a reduction in bunching right after the critical value of 1.96 in the post-treatment period. In Figure 11 in the Appendix I replicate Figure 2 using the data collected by Brodeur et al. (2024a, random sample of articles) and I find further confirmation of a reduction in bunching immediately after the critical value of 1.96 after 2016 in the *American Economic Journals*.

3.2 Difference-in-difference results

The aim of this Section is to dig deeper into the question of whether the star-omission policy had a relevant impact on the share of published test statistics that are significant at conventional levels. Following Blanco-Perez and Brodeur (2020), I adopt a difference-in-difference approach. The unit of observation is a test statistic, and the model reads as follows:

$$\begin{aligned} Y_{aijt} = \quad & \alpha_0 + \alpha_1 BeforeAfter_{aijt} + \alpha_2 Treated_{aijt} + \beta BeforeAfter_{aijt} \times Treated_{aijt} \\ & + \lambda X_{aijt} + \varepsilon_{aijt} \end{aligned} \tag{1}$$

where Y_{aijt} is the outcome variable, i.e. a dummy that takes value 1 if a test statistic i reported in article a in journal j and time t is statistically significant at conventional levels and zero otherwise. $BeforeAfter_{aijt}$ is a dummy that takes value 0 if a test statistic i in article a was published before the star-omission policy entered into place and 1 otherwise. $Treated_{aijt}$ is a dummy that equals 1 if a test statistics i was published in article a in one of the treated journals (*AEJ: Applied Economics*, *AEJ: Macroeconomics* and *AEJ: Economic Policy*), whereas it takes value zero if the article was published in one of the control journals (*Journal of the European Economic Association*, the *Economic Journal* or *Economic Policy*).

	(1)	(2)	(3)	(4)	(5)
Treated journals	-0.042	-0.042	-0.042	-0.041	-0.011
	(0.024)	(0.023)	(0.023)	(0.023)	(0.039)
Before/After dummy	0.011	0.011	0.010	0.010	-0.008
	(0.049)	(0.049)	(0.049)	(0.049)	(0.078)
Before/After X Treated	-0.017	-0.025	-0.025	-0.027	-0.016
	(0.029)	(0.054)	(0.054)	(0.055)	(0.072)
Year dummies	YES	YES	YES	YES	YES
Use of stars dummy		YES	YES	YES	YES
Single author dummy			YES	YES	YES
Method of reporting dummies				YES	YES
Primary JEL code dummies					YES
Constant	YES	YES	YES	YES	YES
Adjusted R-squared	0.008	0.007	0.007	0.011	0.068
Observations	50,610	50,610	50,610	50,610	33,318

Table 1: This table shows OLS estimates of equation (1). The dependent variable is a dummy for whether the test statistic is significant at the 5% level. Robust standard errors are in parentheses, adjusted for clustering by article. In estimation (5) the number of observations is smaller because JEL codes are not reported in *The Economic Journal*, neither in *Economic Policy*, hence these two journals are excluded from this specification. The dummy "use of stars" takes value 1 if either significance stars or bold printing are used to indicate statistical significance at the conventional levels.

The interaction $BeforeAfter_{aijt} \times Treated_{aijt}$ represents the effect of the star-omission policy. Hence, the main coefficient of interest here is β. X includes, depending on the different model specifications, control variables such as year dummies, dummies for the primary jel code, dummy for single-authorship, dummy for the use of stars[12] and dummies for the method of reporting (coefficient and standard error or t-statistic, or p-value).

In Table 1, I report the results of the OLS estimation of equation (1) above[13]. From the results in Table 1, I deduce that the star-omission policy had a negative impact by around 2 percentage points, despite not statistically significant at conventional levels, on the probability of published test statistics to be significant at the 5% level. Adding different sets of controls does not relevantly change this result. This is notably less (around 1/9) then the impact of an editorial statement encouraging submission of papers with negative results, as reported by Blanco-Perez and Brodeur (2020). In the Appendix, the same Table produced for significance levels 1% and 10% is reported (see Table 5 and 6). The estimation results are quite similar to the main estimates in both cases.

Further, in Figure 3, I report the results of the event study based on a dynamic version

[12] Around 4% of articles in the control group for the whole period and 5% of articles in the treated group before treatment do not use stars, nor bold, to denote statistical significance at the conventional levels. Figure 8 in the Appendix shows the evolution over time of this variable, by treatment status.

[13] Blanco-Perez and Brodeur (2020) show that using a logit model instead of OLS does not substantially change the results.

of equation (1). In this dynamic version I regress the outcome variable on year dummies, the treated dummy and interactions of the treated dummy with each year dummy, plus the usual control variables (single author dummy, method of reporting dummies, use of star dummy). The equation reads:

$$\begin{aligned} Y_{aijt} = \quad & \alpha_0 + \alpha_1 YearDummies_t + \alpha_2 Treated_{aijt} + \beta YearDummies_t \times Treated_{aijt} \\ & + \lambda X_{aijt} + \varepsilon_{aijt} \end{aligned} \tag{2}$$

where each term is defined as above. The coefficients of the interactions terms between the treated dummy and the year dummies represent the impact of the policy over time and are reported in Figure 3. Time 0 stands for year 2016, time 1 for 2017 and so on. The event study analysis, while reassuring on the absence of relevant pre-trends, essentially shows that there has been no significant change over time in the probability that a published test statistics is statistically significant following the AEA's star-omission policy. Beyond the statistical insignificance of the estimated impacts over time, no trend is discernible, either.

Since, in economics, the median time from first submission to a journal to publication in that journal is around two years (see Hadavand et al. (2021)), in an alternative specification reported in Appendix (see Table 7), I set the start of the treatment period to 2018 instead than in mid-2016. In this case, I find indication of a somewhat stronger effect (around 3-5 percentage points, still insignificant at conventional levels) of the star-omission policy on the probability that a published test statistic is statistically significant at the 5% level. Figure 9 in the Appendix also shows a reduction in bunching immediately after the critical value of 1.96 in the period 2018-2023 with respect to the previous period (2011-2017) among journals in the treatment group.

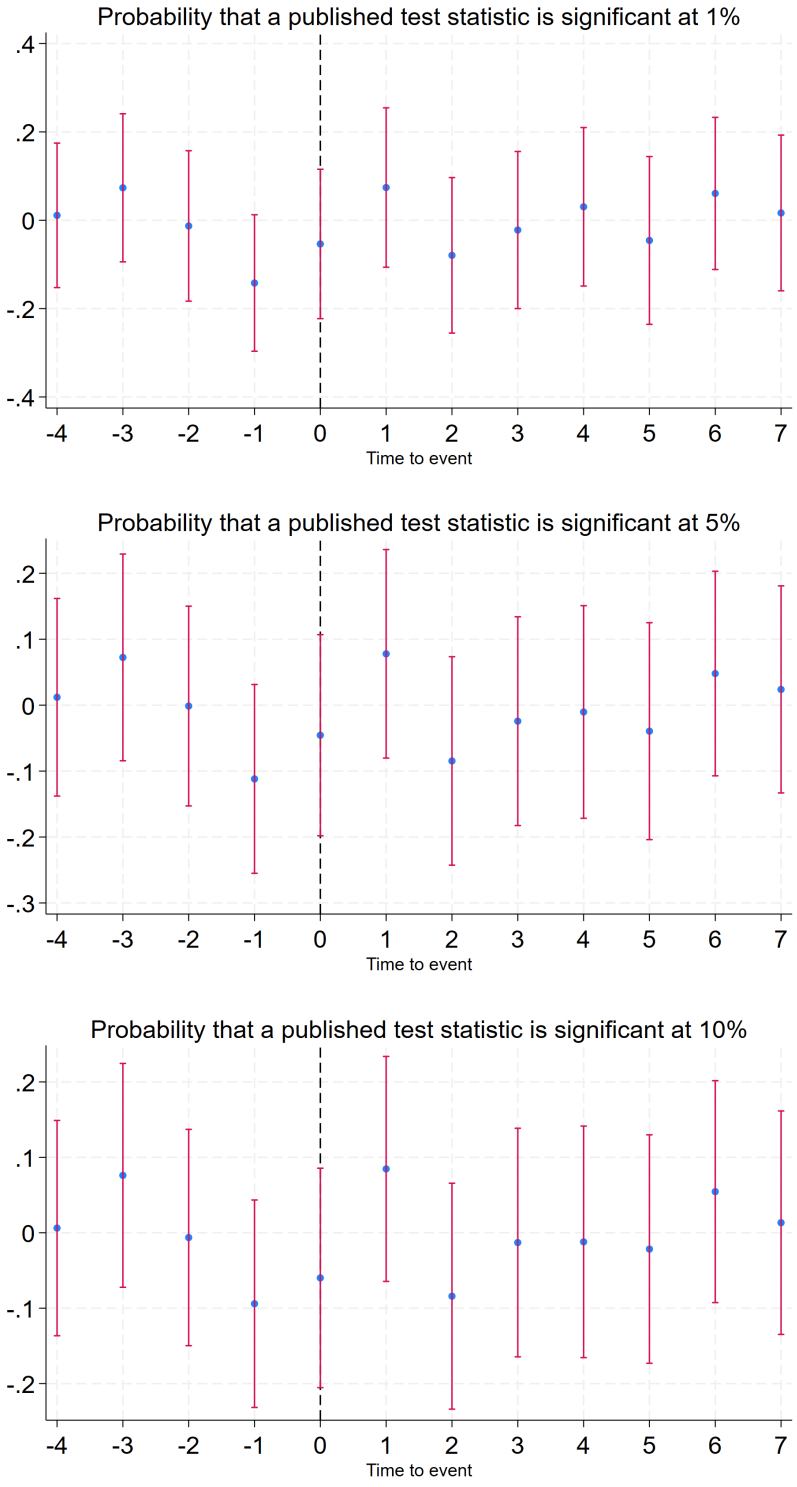

Figure 3: Event study results. The outcome variable is the probability that a published test statistic is significant at the 1% (upper panel), 5% (middle panel) or 10% level (bottom panel). The estimated coefficients of the interaction terms between the treated dummy and the year dummies are reported in the Figure. Control variables are: single author dummy, use of star dummy and method of reporting dummies. Time 0 stands for year 2016, time 1 for 2017 and so on. 95% confidence intervals are reported around the point estimates.

3.3 Results of the tests proposed by Elliott et al. (2022)

In this Section I aim at assessing whether I still detect p-hacking and publication bias in the treated group in the post-treatment period. To this aim, I resort to a series of statistical tests. Following Brodeur et al. (2024a), I present here the results of the battery of tests for p-hacking and publication bias proposed by Elliott et al. (2022). In particular, I report the results for six different tests: binomial, Fisher's, discontinuity, CS1, CS2B and LCM. A comprehensive description of each of these tests can be found in Brodeur et al. (2024a). In all these tests, the null hypothesis is the absence of p-hacking and publication bias. I focus on the results of these tests on the treated group first (Panel A of Table 2). Then, for completeness, I also analyze their results in the control group (Panel B of Table 2). In the first column of Table 2, I present the results of the binomial test, which compares the mass of p-values between 0.045 and 0.05 (test statistics just statistically significant) with the mass of p-values between 0.04 and 0.045 (i.e. those of tests statistics that are slightly more statistically significant). In absence of p-hacking and publication bias, the latter mass should be greater than the former (i.e. the histogram of the p-values should be non-increasing, see Brodeur et al. (2024a)). Based on this test, both before and after the intervention I cannot reject the null hypothesis of absence of p-hacking and publication bias in the treated group.

Name of test	Bin.	Disc.	Fisher	CS1	CS2B	LCM	N. Obs.
Panel A: Treated							
P-value (2011-15)	0.5575	0.0290	1.000	0.9356	0.0818	0.0118	7,964
P-value (2017-23)	0.9741	0.9354	1.000	0.5213	0.1253	0.0754	13,562
Panel B: Control							
P-value (2011-15)	0.9883	0.2648	1.000	0.1812	0.0066	0.4662	7,550
P-value (2017-23)	0.9263	0.0046	1.000	0.0756	0.0083	0.0212	17,694

Table 2: Results of the battery of tests proposed by Elliott et al. (2022), separately for treatment and control group and for the period 2011-2015 and 2017-2023.

The discontinuity test reported in the second column of Table 2 is based instead on a boundary adaptive kernel density estimator that employs local polynomial methods. Under the null hypothesis, the estimated density of the p-values above and below the threshold of p=0.05 should be equal. In the treated group, this null hypothesis is rejected in the pre-treatment sample, but not in the post-treatment one. Similarly to Brodeur et al. (2024a), also here the Fisher test always provides a p-value of 1 and is hence not very informative. The LCM test (column 6 of Table 2) tries to reject the null that the CDF of the curve of the p-values is concave (this is a consequence of the property that the curve

of the p-values needs to be non-increasing). In the treated group, this null hypothesis is rejected in both the 2011-2015 and in the 2017-2023 period, even if the corresponding p-value is larger (0.08 vs 0.01) in the post-intervention period. The CS1 (non-increasingness) and CS2B (which puts bounds on the p-curve and its first and second derivatives) tests are both based on histograms and are more powerful than the more commonly used binomial and Fisher's tests (see Elliott et al. (2022) and Brodeur et al. (2024a)).

Following Elliott et al. (2022), I interpret any p-value in Table 2 lower than 0.1 as evidence of the presence of p-hacking and publication bias. Hence, in the treated group, the CS1 test does not allow to reject the null hypothesis of no p-hacking and no-publication bias in any of the two sample periods. On the other hand, the CS2B test rejects the null in the pre-intervention, but not in the post-intervention period.

To summarize, in the treated group half of the tests proposed by Elliott et al. (2022) suggest the presence of p-hacking and publication bias in the treatment group between 2011 and 2015. However, in the post-intervention period (2017-2023) most of such tests (i.e. five out of six) do not allow us to reject the null hypothesis of absence of p-hacking and publication bias. In the control group, only one of the six text proposed by Elliott et al. (2022), the CS2B, detects the presence of p-hacking and publication bias in the 2011-15 period. However, four out of six tests find evidence of p-hacking in the control group in the 2017-23 period. This is consistent with the increase in bunching immediately after the critical value of 1.96 in the z-statistics after 2016 in the upper Panel of Figure 2. These findings provide indication that there was some reduction in p-hacking following the AEA's star-omission policy and that there was not a generalized reduction in p-hacking and publication bias in the latest years (i.e. in the control group).

In Table 8 in the Appendix, I report the values of the caliper test proposed by Brodeur et al. (2020). This is a local type of analysis (as opposed to the difference-in-difference analysis presented in the previous Section), as only the observations of the z-statistics that are in a certain window around the threshold value of 1.96 are included[14]. Only observations in the treated group are considered here. The marginal effects reported in Table 8 (Appendix) suggest that being in the post-intervention period (represented by the Before/After dummy) is associated with a reduction in the probability that a published test statistic is statistically significant by around 2-4 percentage points. This size of the effect is consistent with the difference-in-difference results presented in Table 1. However, none of these effects are statistically significant at the conventional levels.

[14]I use the same windows as Brodeur et al. (2020).

4 Conclusion

In this paper, I aimed to assess whether the star-omission policy adopted by the AEA journals in mid-2016 impacted the extent of p-hacking and publication bias. Although there are indications that the policy slightly increased the number of published test statistics that do not reject the null hypothesis, this effect is not statistically significant at conventional levels. Focusing solely on the treated journals, I find that only one of the six tests proposed by Elliott et al. (2022) rejects the null hypothesis of no p-hacking and no publication bias in the post-intervention period. In contrast, three out of six tests rejected the null hypothesis in the pre-intervention period. Among the control journals, more tests detect p-hacking in the post-intervention period than before it.

Regarding the magnitude of the star-omission policy's effect, I find a reduction of between 2 and 3 percentage points, depending on the model specification, in the probability that a published test is statistically significant. This is notably less than the 18-percentage-point reduction found by Blanco-Perez and Brodeur (2020) following an editorial statement encouraging the submission of papers with null results.

Authors may have attempted to submit their papers to other journals—where significance stars are reported—before sending them to the treated journals, and this may have mitigated the impact of the star-omission policy. In such cases, indeed, the removal of significance stars would be merely a cosmetic change made after the paper was completed, without reflecting a change in approach. However, it is worth noting that between 25% and 35% of all papers published in the *American Economic Journals* have been transferred with reports from *The American Economic Review*, which also adopted the same star-omission policy in mid-2016. Hence, this effect should be limited[15]. Nevertheless, to observe a more substantial impact on p-hacking and publication bias, a larger number of leading journals should adopt the same star-omission policy. Alternatives, such as issuing an editorial statement, appear to be more effective in achieving the desired goal.

References

1. Abadie, A. (2020). Statistical nonsignificance in empirical economics. American Economic Review: Insights, 2(2), 193-208.

[15]Since 2011, authors have had the option to submit a paper, along with referee reports obtained from *The American Economic Review*, to one of the *American Economic Journals*, streamlining the review process. Such transfer papers account for approximately 25–35% of all published papers, as reported here: https://blogs.worldbank.org/en/impactevaluations/publishing-stats-and-news-aea-journals#: :text=In%202022%20they%20got%2067,percent%20of%20total%20published%20papers.

2. Ashenfelter, O., & Greenstone, M. (2004). Estimating the value of a statistical life: The importance of omitted variables and publication bias. American Economic Review, 94(2), 454-460.

3. Blanco-Perez, C., & Brodeur, A. (2020). Publication bias and editorial statement on negative findings. The Economic Journal, 130(629), 1226-1247.

4. Brodeur, A., Cook, N., & Neisser, C. (2024a). P-hacking, data type and data-sharing policy. The Economic Journal, 134(659), 985-1018.

5. Brodeur, A., Cook, N. M., Hartley, J. S., & Heyes, A. (2024b). Do Preregistration and Preanalysis Plans Reduce p-Hacking and Publication Bias? Evidence from 15,992 Test Statistics and Suggestions for Improvement. Journal of Political Economy Microeconomics, 2(3), 527-561.

6. Brodeur, A., Carrell, S., Figlio, D., & Lusher, L. (2023). Unpacking p-hacking and publication bias. American Economic Review, 113(11), 2974-3002.

7. Brodeur, A., Cook, N., & Heyes, A. (2020). Methods matter: P-hacking and publication bias in causal analysis in economics. American Economic Review, 110(11), 3634-3660.

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

9. Chopra, F., Haaland, I., Roth, C., & Stegmann, A. (2024). The null result penalty. The Economic Journal, 134(657), 193-219.

10. Elliott, G., Kudrin, N., & Wüthrich, K. (2022). Detecting p-hacking. Econometrica, 90(2), 887-906.

11. Hadavand, A., Hamermesh, D. S., & Wilson, W. W. (2021). Publishing economics: How slow? why slow? is slow productive? fixing slow? (No. w29147). National Bureau of Economic Research.

12. Imbens, G. W. (2021). Statistical significance, p-values, and the reporting of uncertainty. Journal of Economic Perspectives, 35(3), 157-174.

13. Ioannidis, J. P., Stanley, T. D., & Doucouliagos, H. (2017). The power of bias in economics research.

14. Kranz, S., & Pütz, P. (2022). Methods matter: P-hacking and publication bias in causal analysis in economics: Comment. American Economic Review, 112(9), 3124-3136.

15. Stanley, T. D. (2008). Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. Oxford Bulletin of Economics and statistics, 70(1), 103-127.

Appendix (for online publication only)

A. Descriptive statistics

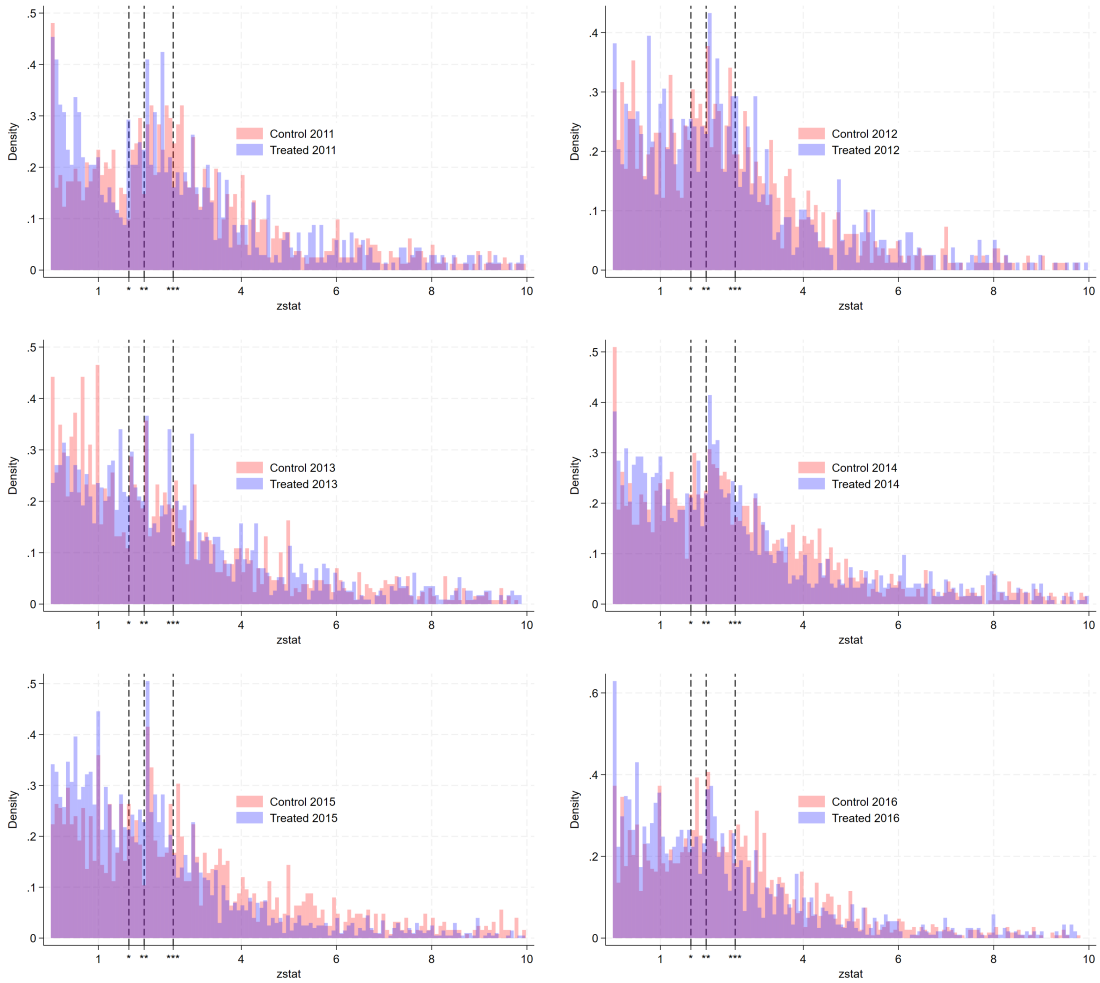

Figure 4: Histogram (125 bins) of the z-statistic values collected, comparison between the control and the treatment group in each of the sample years from 2011 to 2016 (extremes included). Z-statistics larger than 10 have been trimmed in order to improve graph readability.

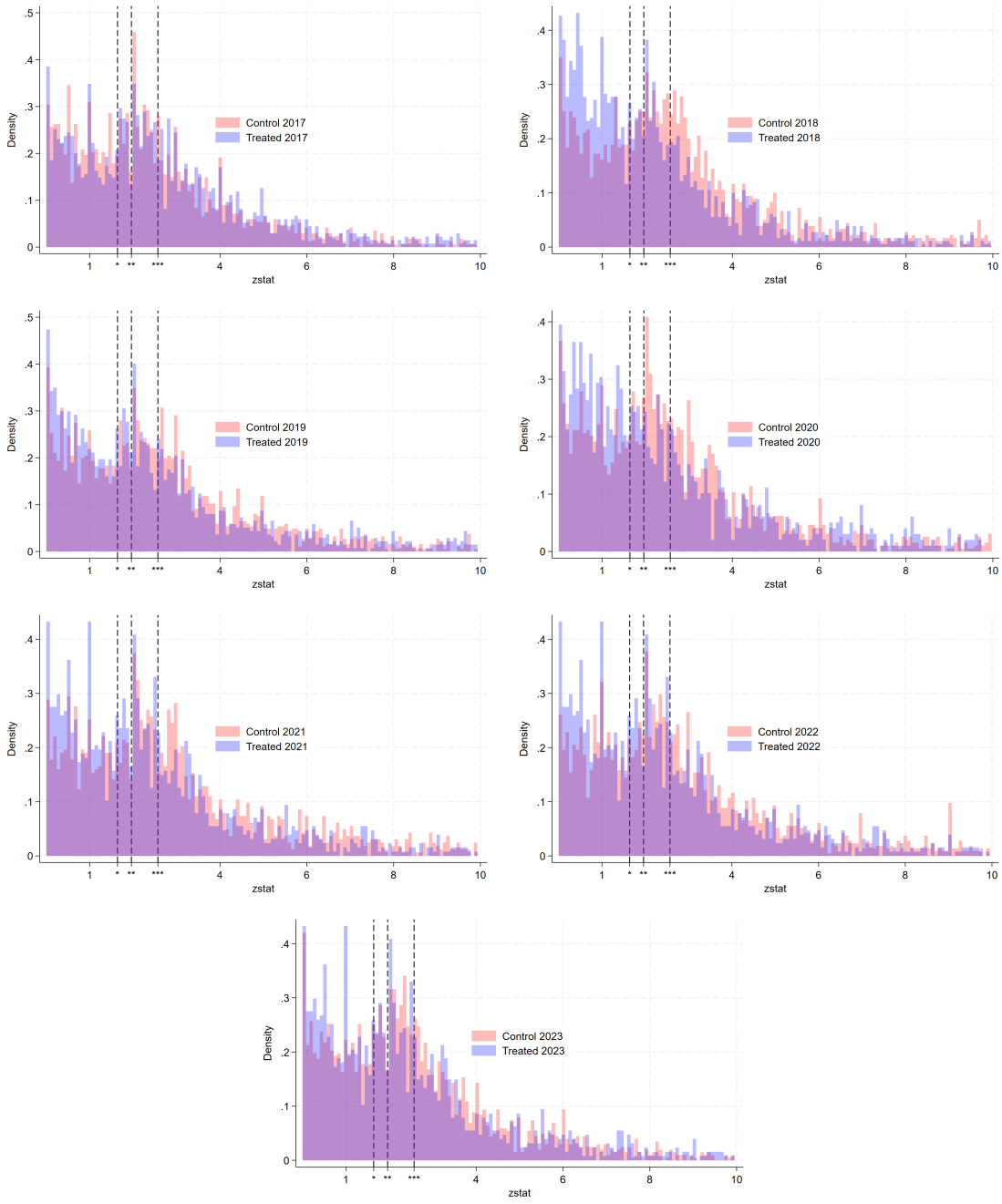

Figure 5: Histogram (125 bins) of the z-statistic values collected, comparison between the control and the treatment group in each of the sample years from 2017 to 2023 (extremes included). Z-statistics larger than 10 have been trimmed in order to improve graph readability.

Proprotion of	Articles	Tests
AEJ: Macro	12.76	3.85
AEJ: Applied	14.64	18.75
AEJ: Policy	16.52	22.94
JEEA	18.71	18.88
EJ	31.76	32.50
Economic Policy	5.61	3.09
Single-authored	19.95	17.44
Using stars	66.43	68.97
Reporting SE	95.52	96.53
Reporting t-stats	3.48	2.50
Reporting p-values	1.01	0.97

Table 3: Descriptive statistics of the treated and control group samples. Period 2011-2023.

	2011-2015	2016-2017	2018-2023
Panel A: Treated Journals			
Significant at 1%	0.4547	0.4556	0.4625
	(0.4980)	(0.4981)	(0.4986)
Significant at 5%	0.5872	0.5890	0.5851
	(0.4924)	(0.4921)	(0.4927)
Significant at 10%	0.6471	0.6524	0.6467
	(0.4779)	(0.4763)	(0.4780)
Panel B: Control Journals			
Significant at 1%	0.5338	0.4703	0.5054
	(0.4989)	(0.4992)	(0.5000)
Significant at 5%	0.6498	0.6080	0.6455
	(0.4771)	(0.4883)	(0.4784)
Significant at 10%	0.7070	0.6772	0.7064
	(0.4552)	(0.4676)	(0.4554)

Table 4: This table reports the percentage of test statistics statistically significant at conventional levels for three categories of articles: (i) published between 2011 and 2015 (ii) published between 2016 and 2017, (iii) published between 2018 and 2023. Standard deviations are in parentheses.

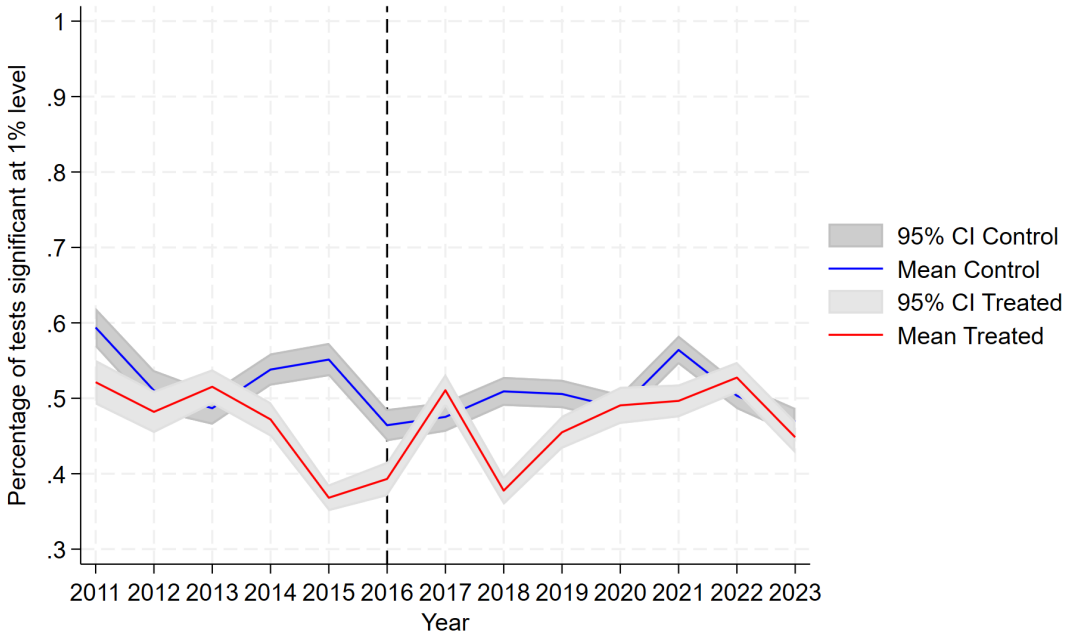

Figure 6: Percentage of tests significant at the 1% level, by year of publication. Treated journals include *AEJ: Applied Economics*, *AEJ: Macroeconomics* and *AEJ: Economic Policy*. Control journals include the *Journal of the European Economic Association*, *The Economic Journal* and *Economic Policy*.

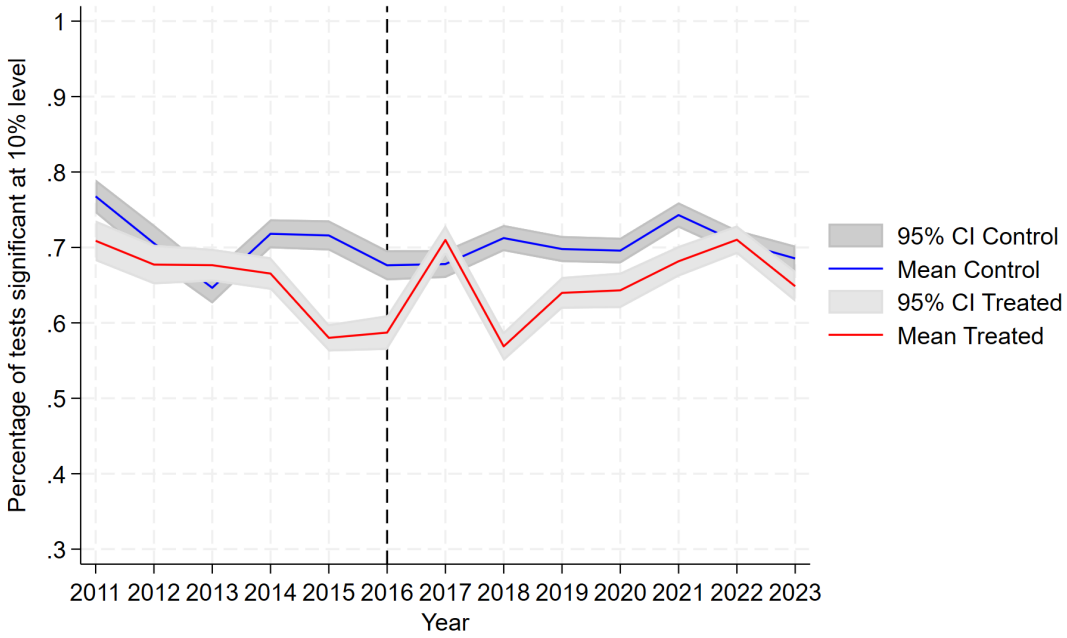

Figure 7: Percentage of tests significant at the 10% level, by year of publication. Treated journals include *AEJ: Applied Economics*, *AEJ: Macroeconomics* and *AEJ: Economic Policy*. Control journals include the *Journal of the European Economic Association*, *The Economic Journal* and *Economic Policy*.

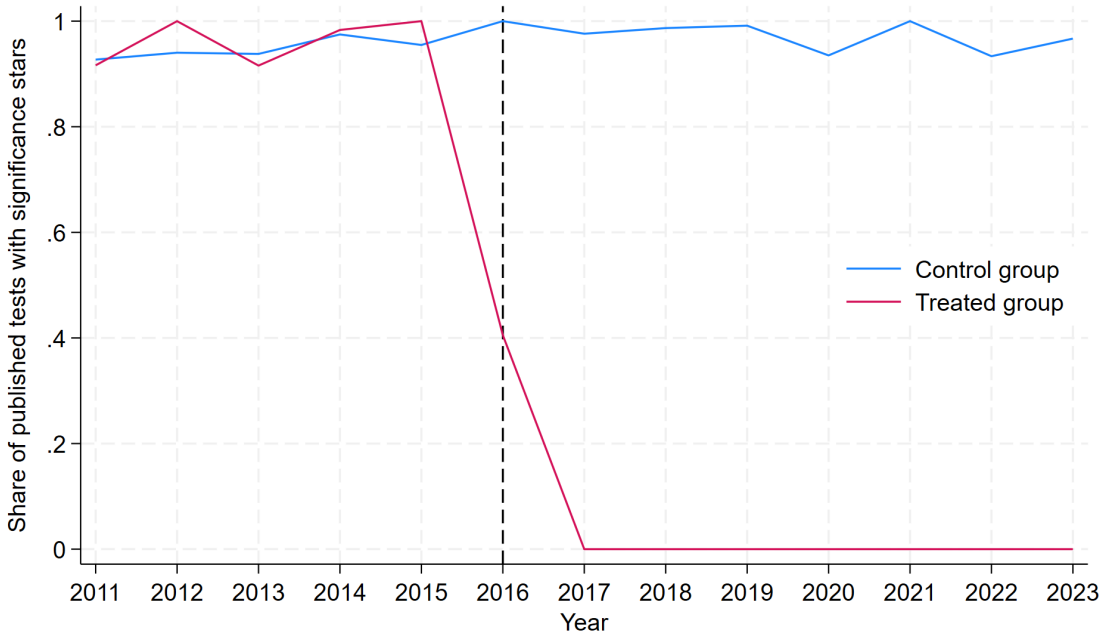

Figure 8: Percentage of tests published with significance stars, by year of publication. Treated journals include *AEJ: Applied Economics*, *AEJ: Macroeconomics* and *AEJ: Economic Policy*. Control journals include the *Journal of the European Economic Association*, *The Economic Journal* and *Economic Policy*. $N = 27,444$ in the control group and $N = 23,166$ in the treatment group.

B. Additional empirical results

	(1)	(2)	(3)	(4)	(5)
Treated journals	-0.053	-0.053	-0.053	-0.052	-0.032
	(0.025)	(0.025)	(0.025)	(0.025)	(0.036)
Before/After dummy	0.005	0.006	0.006	0.005	-0.029
	(0.050)	(0.050)	(0.050)	(0.050)	(0.075)
Before/After X Treated	0.011	-0.012	-0.012	-0.015	0.017
	(0.031)	(0.061)	(0.060)	(0.061)	(0.078)
Year dummies	YES	YES	YES	YES	YES
Use of stars dummy		YES	YES	YES	YES
Single author dummy			YES	YES	YES
Method of reporting dummies				YES	YES
Primary JEL code dummies					YES
Constant	YES	YES	YES	YES	YES
Adjusted R-squared	0.007	0.007	0.007	0.011	0.076
Observations	50,610	50,610	50,610	50,610	33,318

Table 5: This table shows OLS estimates of equation (1). The dependent variable is a dummy for whether the test statistic is significant at the 1% level. Robust standard errors are in parentheses, adjusted for clustering by article. In estimation (5) the number of observations is smaller because JEL codes are not reported in *The Economic Journal*, neither in *Economic Policy*, hence these two journals are excluded from this specification. The dummy "use of stars" takes value 1 if either significance stars or bold printing are used to indicate statistical significance at the conventional levels.

	(1)	(2)	(3)	(4)	(5)
Treated journals	-0.044	-0.044	-0.044	-0.043	-0.029
	(0.022)	(0.022)	(0.022)	(0.022)	(0.040)
Before/After dummy	0.011	0.011	0.010	0.009	-0.011
	(0.046)	(0.046)	(0.046)	(0.046)	(0.080)
Before/After X Treated	-0.014	-0.020	-0.021	-0.022	-0.001
	(0.027)	(0.050)	(0.050)	(0.051)	(0.065)
Year dummies	YES	YES	YES	YES	YES
Use of stars dummy		YES	YES	YES	YES
Single author dummy			YES	YES	YES
Method of reporting dummies				YES	YES
Primary JEL code dummies					YES
Constant	YES	YES	YES	YES	YES
Adjusted R-squared	0.008	0.007	0.007	0.010	0.065
Observations	50,610	50,610	50,610	50,610	33,318

Table 6: This table shows OLS estimates of equation (1). The dependent variable is a dummy for whether the test statistic is significant at the 10% level. Robust standard errors are in parentheses, adjusted for clustering by article. In estimation (5) the number of observations is smaller because JEL codes are not reported in *The Economic Journal*, neither in *Economic Policy*, hence these two journals are excluded from this specification. The dummy "use of stars" takes value 1 if either significance stars or bold printing are used to indicate statistical significance at the conventional levels.

	(1)	(2)	(3)	(4)	(5)
Treated journals	-0.040	-0.042	-0.042	-0.042	-0.001
	(0.020)	(0.022)	(0.022)	(0.022)	(0.034)
Before/After dummy	-0.029	-0.030	-0.031	-0.027	0.031
	(0.037)	(0.038)	(0.038)	(0.038)	(0.053)
Before/After X Treated	-0.025	-0.031	-0.030	-0.029	-0.048
	(0.027)	(0.034)	(0.034)	(0.035)	(0.048)
Year dummies	YES	YES	YES	YES	YES
Use of stars dummy		YES	YES	YES	YES
Single author dummy			YES	YES	YES
Method of reporting dummies				YES	YES
Primary JEL code dummies					YES
Constant	YES	YES	YES	YES	YES
Adjusted R-squared	0.008	0.007	0.007	0.011	0.068
Observations	50,610	50,610	50,610	50,610	33,318

Table 7: This table shows OLS estimates of equation (1). The dependent variable is a dummy for whether the test statistic is significant at the 5% level. Robust standard errors are in parentheses, adjusted for clustering by article. In estimation (5) the number of observations is smaller because JEL codes are not reported in *The Economic Journal*, neither in *Economic Policy*, hence these two journals are excluded from this specification. The dummy "use of stars" takes value 1 if either significance stars or bold printing are used to indicate statistical significance at the conventional levels. Treatment period here is 2018-2023.

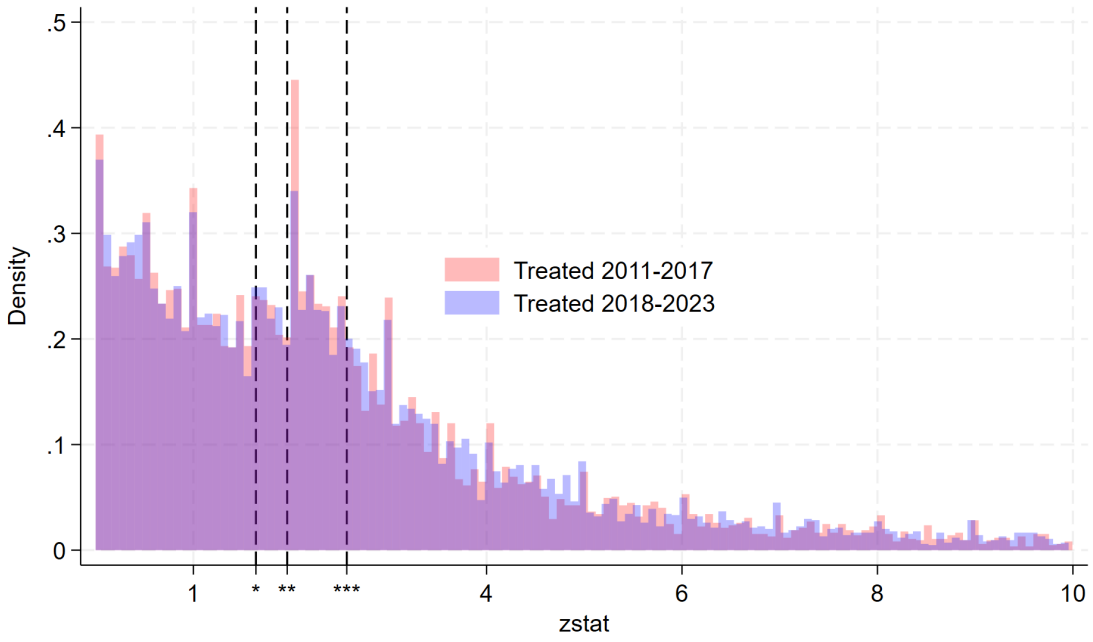

Figure 9: Histogram (125 bins) of the z-statistic values collected from journals in the treatment group, comparison of the periods 2011-2017 and 2018-2023. Z-statistics larger than 10 have been trimmed in order to improve graph readability.

	(1)	(2)	(3)
Before/After	-.0187	-.0406	-.0356
dummy	(.0370)	(.0388)	(.0467)
Journal FE	YES	YES	YES
Year FE	YES	YES	YES
Observations	4,683	3,422	1,981
Threshold	1.96	1.96	1.96
Window	0.5	0.35	0.2

Table 8: Each observation is a test statistic. Only journals in the treated group are kept. I rely on probit models and present the average marginal effects and associated SEs clustered at the journal article level as in Brodeur et al. (2020) and Brodeur et al. (2024a). Observations are unweighted. The dependent variable takes the value one if the test statistic is significant at the 5% level and zero otherwise. The primary independent variable takes value one if the test statistic has been published in the post-intervention period and zero otherwise.

C. Descriptives obtained with the data from Brodeur et al. (2024a)

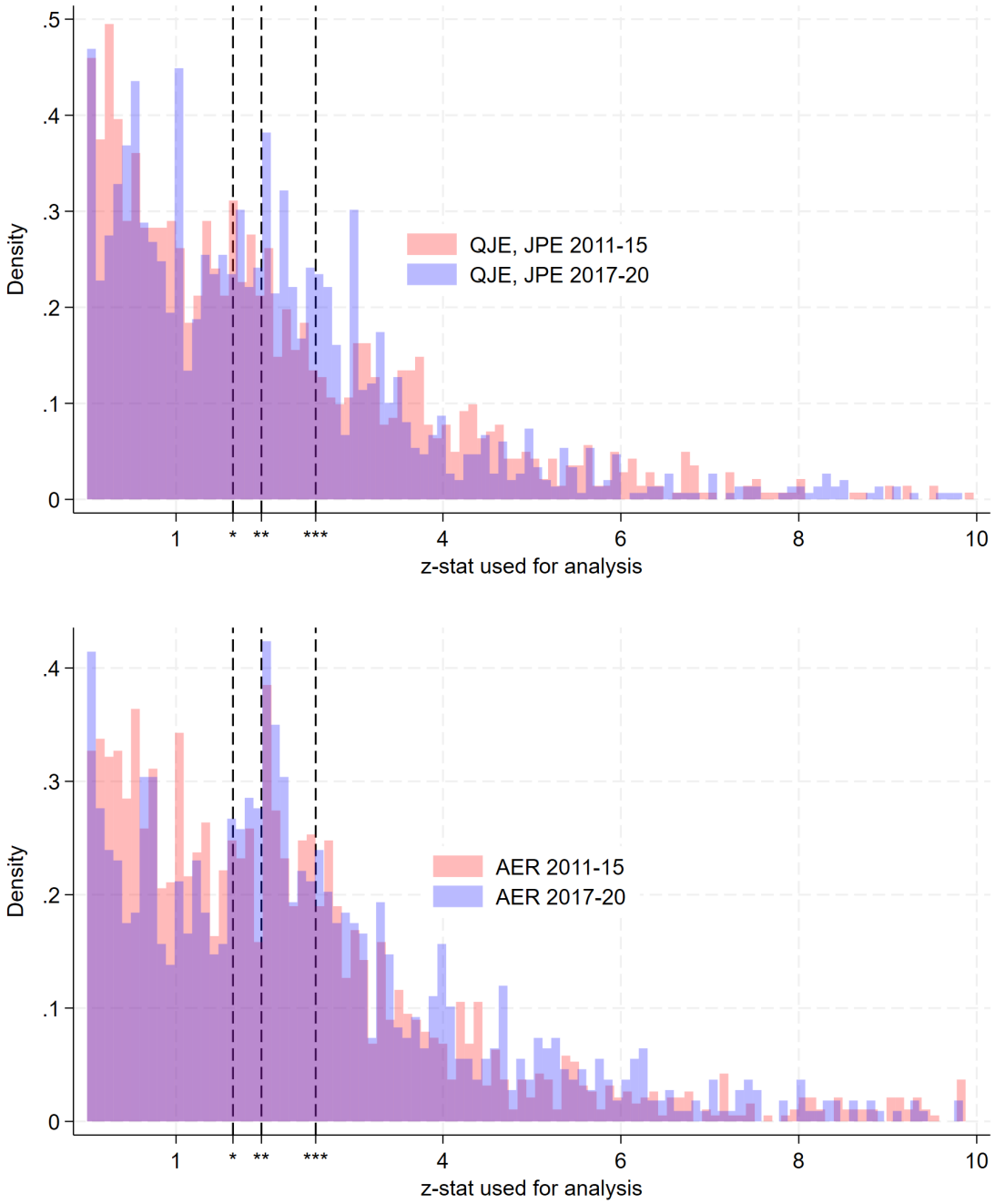

Figure 10: Histogram (100 bins) of the z-statistic values collected from the *Quarterly Journal of Economics* and the *Journal of Political Economy* (control, upper panel), and from *American Economic Review* (treated, lower panel), comparison of the periods 2011-2015 and 2017-2020. Data are from Brodeur et al. (2024a). In the control group, $N = 1,548$ for period 2011-2015 and $N = 1,615$ for 2017-2020. In the treated group, $N = 2,045$ for 2011-2015 and $N = 1,171$ for 2017-2020. Z-statistics larger than 10 have been trimmed in order to improve graph readability.

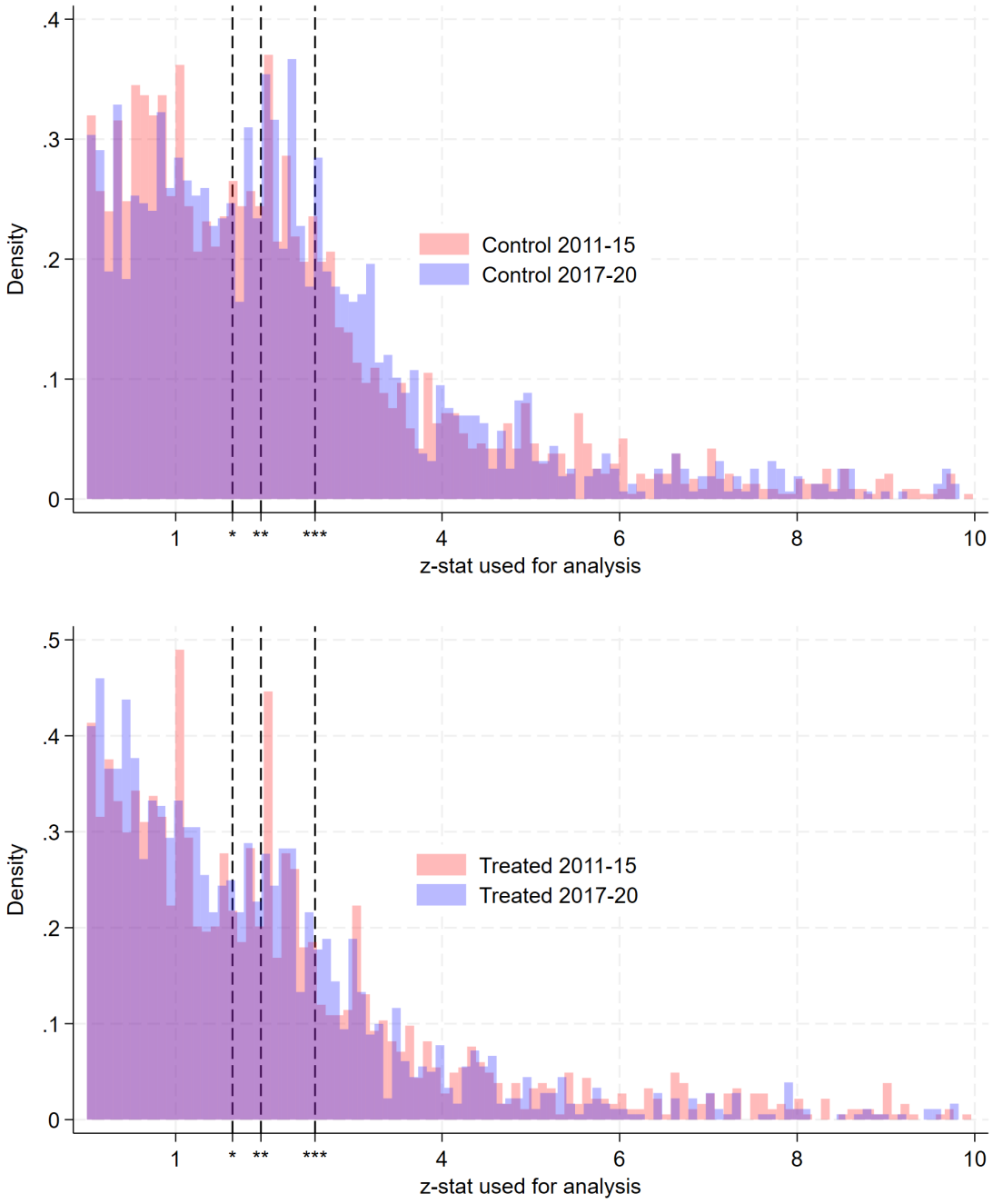

Figure 11: Histogram (100 bins) of the z-statistic values collected from the control journals (*Journal of the European Economic Association*, *Economic Journal*, *Economic Policy*, upper panel), and from the *American Economic Journals* (treated, lower panel), comparison of the periods 2011-2015 and 2017-2020. Data are from Brodeur et al. (2024a). In the control group, $N = 2,486$ for period 2011-2015 and $N = 1,689$ for 2017-2020. In the treated group, $N = 1,894$ for 2011-2015 and $N = 1,881$ for 2017-2020. Z-statistics larger than 10 have been trimmed in order to improve graph readability.