Naguib, Costanza

Working Paper

Does single-blind review encourage or discourage p-hacking?

Discussion Papers, No. 25-04

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

Suggested Citation: Naguib, Costanza (2025) : Does single-blind review encourage or discourage p-hacking?, Discussion Papers, No. 25-04, University of Bern, Department of Economics, Bern

This Version is available at:
https://hdl.handle.net/10419/324321

UNIVERSITÄT
BERN

Faculty of Business, Economics
and Social Sciences

Department of Economics

Does single-blind review encourage or discourage p-hacking?

Costanza Naguib

25-04

July, 2025

DISCUSSION PAPERS

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

Does single-blind review encourage or discourage p-hacking?

Costanza Naguib*

Abstract

In 2011, the American Economic Association (AEA) changed its peer review policy for all their journals, shifting from a double-blind process to a single-blind peer-review process. Under this new system, referees became aware of the authors' identities. In this paper, I explore whether this policy change influenced the prevalence of p-hacking in published papers at *The American Economic Review*.

JEL codes: A11, A14, C13

Keywords: p-hacking, single-blind review, double-blind review, difference-in-difference

1 Introduction

[1] There is an ongoing debate on whether single-blind or double-blind review is more suitable for the peer review of scientific articles in economics. In a single-blind review, the author remains unaware of the referee's identity, whereas in a double-blind review, the referee is also not informed about the author's identity. Proponents of double-blind review argue that it helps mitigate biases from referees. They highlight two primary concerns. First, referees might unfairly reject papers from lesser-known authors or those affiliated with lower-ranked institutions, irrespective of the paper's quality. Second, if gender or ethnic discrimination is present, papers authored by women or individuals with foreign-sounding names might face biased reviews (Blank (1991))[2].

Supporters of single-blind reviewing usually present the following three arguments. First, they claim that referees can often identify the author of a paper through its content or citations, so double-blind systems are seldom truly anonymous.[3] Second, they argue

*University of Bern

[1] I thank Stefan Egli, Reto Horst and Thibaud Laurent for the excellent research assistance.

[2] It is however also possible that referees are stricter on more prolific authors, as they wish to leave space also to newcomers and to prevent already established scholars from publishing marginal papers, as suggested by Card and DellaVigna (2020).

[3] According to Blank (1991), referees were able to identify the authors of around 50% of papers at the end of the 80s. Despite the improvement in search engines, this share seems to have remained fairly stable up to now (see Cressey (2014), Hill and Provost (2003)).

that knowing the author's name and institution provides valuable context that can affect how the paper is read and evaluated. Third, editors often note that double-blind reviewing involves additional administrative effort, requiring more meticulous procedures in the editorial office.

In this paper, I aim to assess whether the single-blind review system encourages or discourages p-hacking. P-hacking refers to practices, such as specification searching, that researchers may use to obtain more favorable p-values. This often occurs in response to the challenges of publishing null results, as null findings are often perceived to be of lower quality (Imbens (2021), Chopra et al. (2024)). Such practices increase the prevalence of false positives in the literature, skewing the research presented to policymakers. Under a single-blind system, young researchers and those from less prestigious institutions may feel pressured to impress referees with eye-catching statistically significant results. If this is the case, it would provide yet another argument against adopting a single-blind system.

In March 2011, the American Economic Association (AEA) decided to switch from a double-blind to a single-blind peer review standard.[4]. At the time, it was the only one among the so-called top-5 journals to use a double-blind system, whereas all the others were using single-blind reviews. This contrast provides a framework for studying a quasi-natural experiment. I aim to analyze whether the policy change by *The American Economic Review* was followed by a change in p-hacking practices[5].

It is worth noting that, as mentioned above, in the age of Google, reviewers are able to identify the authors of a paper from its text or citations with approximately 50% accuracy (Cressey (2014), Hill and Provost (2003)). This means that I actually estimate the impact on p-hacking of an increase in the probability of author identification by the referees from around 50% to 100%.

[4]The statement read: "Upon a joint recommendation of the editors of the American Economic Review and the four American Economic Journals, the Executive Committee has voted to drop the 'double-blind' refereeing process for all journals of the American Economic Association. The change to 'single-blind' refereeing will become effective on July 1, 2011. Easy access to search engines increasingly limits the effectiveness of the double-blind process in maintaining anonymity. Further, it increases the administrative cost of the journals and makes it harder for referees to identify an author's potential conflicts of interest arising, for example, from consulting." Source: https://crookedtimber.org/2011/06/05/should-the-american-economic-review-drop-double-anonymous-review/ and Goldberg (2012). In the present paper I only consider the *American Economic Review* and not the four American Economic Journals, as the latter only started to publish issues in 2009, i.e. two years before the change in the peer-review standard only.

[5]Another change of policy in the period under scrutiny is the implementation of a strict page limit on submissions in 2008 by AER alone among the top-5. While a strict limit on page count may also encourage p-hacking practices, Card and DellaVigna (2014) find that this intervention had essentially no impact on the length of published papers, and authors mostly adjusted their submission by means of purely aesthetic formatting changes in order to meet the page limit. Hence, it is likely that this intervention did not influence the extent of p-hacking.

First, I apply a difference-in-difference approach to determine whether a double-blind review policy is associated with a lower proportion of statistically significant test statistics being published than a single-blind one. The treatment is having a single-blind policy. Since no other Top-5 journal was adopting a double-blind policy in 2011, there is no control group, but only an always treated group and a switcher group. I am hence in the framework of "time-reverse difference-in-difference". As shown by Kim and Lee (2018), the estimation procedure is essentially the same as in the standard DiD framework. Second, I investigate whether a series of statistical tests can detect p-hacking under the single-blind regime, respectively under the double-blind regime in papers published at *The American Economic Review*.

I am the first to evaluate the impact of different peer review standards on the extent of p-hacking. I find that, under a double-blind review system, authors from top institutions tend to report a higher proportion of statistically significant results, whereas authors from non-top institutions, on average, report a lower share of statistically significant findings than they do under a single-blind standard. These findings are robust to a range of sensitivity analyses, including procedures such as de-rounding and weighting.

This study contributes to the growing body of research on p-hacking, building on seminal work by Brodeur et al. (2016), who documented p-hacking and publication bias in three top economics journals (*The American Economic Review* (AER), *The Quarterly Journal of Economics* (QJE), and *The Journal of Political Economy* (JPE)). Brodeur et al. (2020) later demonstrated that these issues vary by estimation method, with Kranz and Pütz (2022) noting that such findings may be relevantly influenced by rounding errors. Subsequent research by Brodeur et al. (2024a, 2024b) showed that neither data availability and replication policies nor pre-registration and pre-analysis plans significantly reduce p-hacking.

Blanco-Perez and Brodeur (2020) assessed the impact of a 2015 editorial statement from eight health economics journals encouraging the publication of statistically insignificant but economically relevant results. This intervention effectively reduced p-hacking and publication bias, lowering the proportion of tests rejecting the null hypothesis by approximately 18 percentage points. Similarly, Naguib (2024) studies the impact of the omission of significance asterisks implemented by the AEA journals in mid-2016 on the extent of p-hacking and publication bias, finding essentially no impact of the policy. Finally, McCloskey and Michaillat (2024) derive critical values for hypothesis testing that are robust to p-hacking.

Studies evaluating the costs and benefits of single-blind vs double-blind peer review

of scientific articles in economics are scarce. The few existing papers provide suggestive evidence of editorial favoritism in the single-blind review process. Blank (1991) analyzes submissions to *The American Economic Review* from 1987 to 1989. In this period and experiment took place, where some papers were randomly assigned to single-blind and others to double-blind review. Blank (1991) finds that the double-blind review process results in lower acceptance rates and more critical referee comments for authors affiliated with mid-ranking top universities (ranks 6–50). However, she does not find any specific adverse effect of single-blind review on female authors. This is further confirmed by Carlsson et al. (2012), studying double and single-blind acceptance decisions for a Swedish conference held in 2008[6]. Nevertheless, Laband and Piette (1994) analyzed more than 1,000 articles published in 28 leading economics journals in 1984 and discovered that articles from journals using double-blind review received more citations over a five-year period, indicating higher quality compared to those published in journals using single-blind review.

1.1 Potential mechanisms

Under a single-blind peer review system, authors affiliated with less prestigious institutions might feel increased pressure to present statistically significant results to enhance their chances of publication. Consequently, they may be more inclined to engage in p-hacking practices compared to when they work under a double-blind review system.

Conversely, in a double-blind review system, authors from reputable institutions may experience heightened pressure to present particularly compelling results, as their identities are anonymized and cannot influence reviewers. Thus, the shift from single-blind to double-blind review creates opposing incentives for different groups of authors. It is not clear, a priori, what the net effect of this transition on the prevalence of p-hacking would be. Notably, Brodeur et al. (2020) find that p-hacking practices are not relevantly influenced by authors' experience levels or their institutions' rankings.

Regarding publication bias, this term refers to the preference exhibited by editors and referees for statistically significant results. Given that editors always know the authors' identities, I anticipate no significant change in their attitudes towards null results between single-blind and double-blind review systems.

However, referees may display greater bias against null findings submitted by authors from less prestigious institutions under a single-blind system. Conversely, referees may also

[6]However, Hengel (2022) finds that, under a single-blind system, women are held to higher writing standards than men.

demonstrate more leniency towards null results when they originate from well-established authors at top institutions in the same system. In summary, switching from a double-blind to a single-blind review policy is expected to:

- Reduce p-hacking practices and publication bias among papers authored by prominent researchers from reputable institutions.
- Increase p-hacking practices and publication bias among papers authored by less-known researchers from lower-ranked institutions.

The overall net impact of these two contrasting effects on p-hacking and publication bias remains unclear a priori.

2 Data Description

I exploit the quasi-natural experiment of *The American Economic Review* (AER) passing from a double-blind to a single-blind peer review standard for their papers in July 2011[7]. I consider data for the period 2005-2015. The analysis period stops in 2015, because in mid-2016 the AEA introduced another policy that may potentially impact the extent of p-hacking, e.g. the omission of significance stars from the regression tables. This intervention is described in detail in Naguib (2024). I compare the extent of p-hacking in the AER with that in comparable top-5 journals in economics such as *The Quarterly Journal of Economics* (QJE) and *The Journal of Political Economy* (JPE), which have both been adopting a single-blind review standard for the full period of analysis[8].

In my dataset, 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 robustness 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

[7] The decision was announced in March 2011, and become effective on 1st July 2011.

[8] As mentioned above, we are hence in the framework of an "reverse-time difference-in-difference setup", where one group was treated (i.e. adopting a single-blind standard) for the whole period of analysis and the other group started being treated only at a certain point in time, see Kim and Lee (2018).

from the collection. 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.

There is notable overlap between my dataset and the one collected by Brodeur et al. (2016), in particular for the years 2005-2011. For those years, I only collect the main results, whereas they also collect robustness checks and similar additional results. For the years 2012-2016, I collect estimated coefficients from all the articles published in the three journals under study, whereas Brodeur et al. (2024a) only collects coefficients from a random sample of articles. In Figure 5 in the Appendix I show the distribution of z-statistics, respectively in my sample and in the sample collected by Brodeur et al. (2024a). Both histograms clearly exhibit what Brodeur et al. (2016) calls a "two-humped camel shape", which suggests the presence of p-hacking and/or publication bias.

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 three research assistants. We discussed and clarified discordant cases. 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. This can be a rough approximation of significance for two reasons. First, the effective number of degrees of freedom may be modest, especially in case of results obtained with cluster-robust standard errors and a limited number of clusters. Second, different journals may adopt different rounding conventions for their reported results. If these conventions differ across journals and/or across time, this may not cancel out in my diff-in-diff setting and hence bias the results.

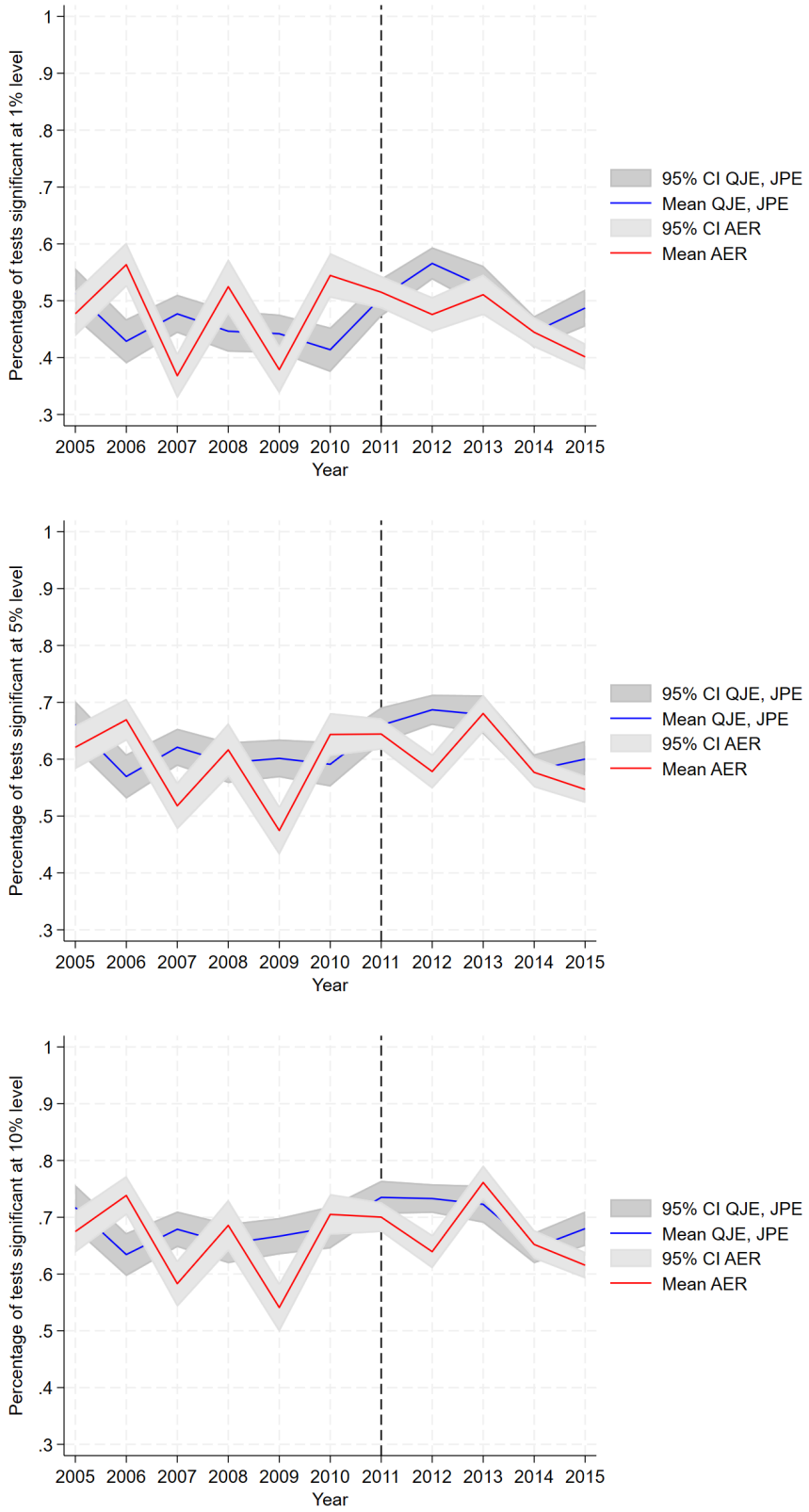

Figure 1: Percentage of tests significant, respectively at the 1% level (upper panel), at the 5% level (middle panel), and at the 10% level (bottom panel), by year of publication.

While I am not able to reconstruct the effective number of degrees of freedom for each reported results and hence I have to accept this source of imprecision in my data, in Appendix D I present the results obtained via two de-rounding methods to tackle the issue of coarse rounding of the reported values.

Table 3 in Appendix A reports descriptive statistics for the variables collected in my dataset. The variable "Top institution" records the share of authors of a certain paper that were affiliated with one of the 20 top institutions as defined by Brodeur et al. (2020)[9] in the year in which the paper was published.

In Figure 1, I provide a descriptive snapshot of the data. Over the entire period, roughly 60–70% of published tests are significant at the 5% level, 40–50% are significant at the 1% level, and 60–75% at the 10% level. Crucially, neither the control nor the treatment group shows a clear upward or downward trend in the share of significant results. The assumption of parallel trends does not look unreasonable in this context.

From Figure 2, I get further insight into the the distribution of test statistics before and after the switch from double-blind to single-blind policy at AER. All the histograms presented in Figure 2 provide suggestive evidence of p-hacking and/or publication bias. Rather than displaying a smooth, monotonic decline as the value of the z-statistics increases, each histogram features a noticeable dip just before the conventional significance thresholds, followed by a pronounced spike immediately after. This pattern is consistent with researchers selectively reporting results that cross significance cutoffs.

I do not observe any clear shift in the distribution of z-statistics at AER following the change in the review standard. This aligns with my initial hypothesis: the new review policy may encourage p-hacking behaviors among some authors while discouraging them among others, resulting in no obvious net effect. Although there appears to be a slight increase in bunching just above the 1.96 threshold during the 2012–2015 period compared to earlier years, Appendix D shows that this is likely driven by coarse rounding of reported statistics. Therefore, it should not be interpreted as direct evidence of a change in p-hacking practices.

[9]These are: Barcelona GSE, Boston University, Brown, Chicago, Columbia, Dartmouth, Harvard, MIT, Northwestern, NYU, Princeton, PSE, TSE, UC Berkeley, UCL, UCSD, UPenn, Stanford, and Yale.

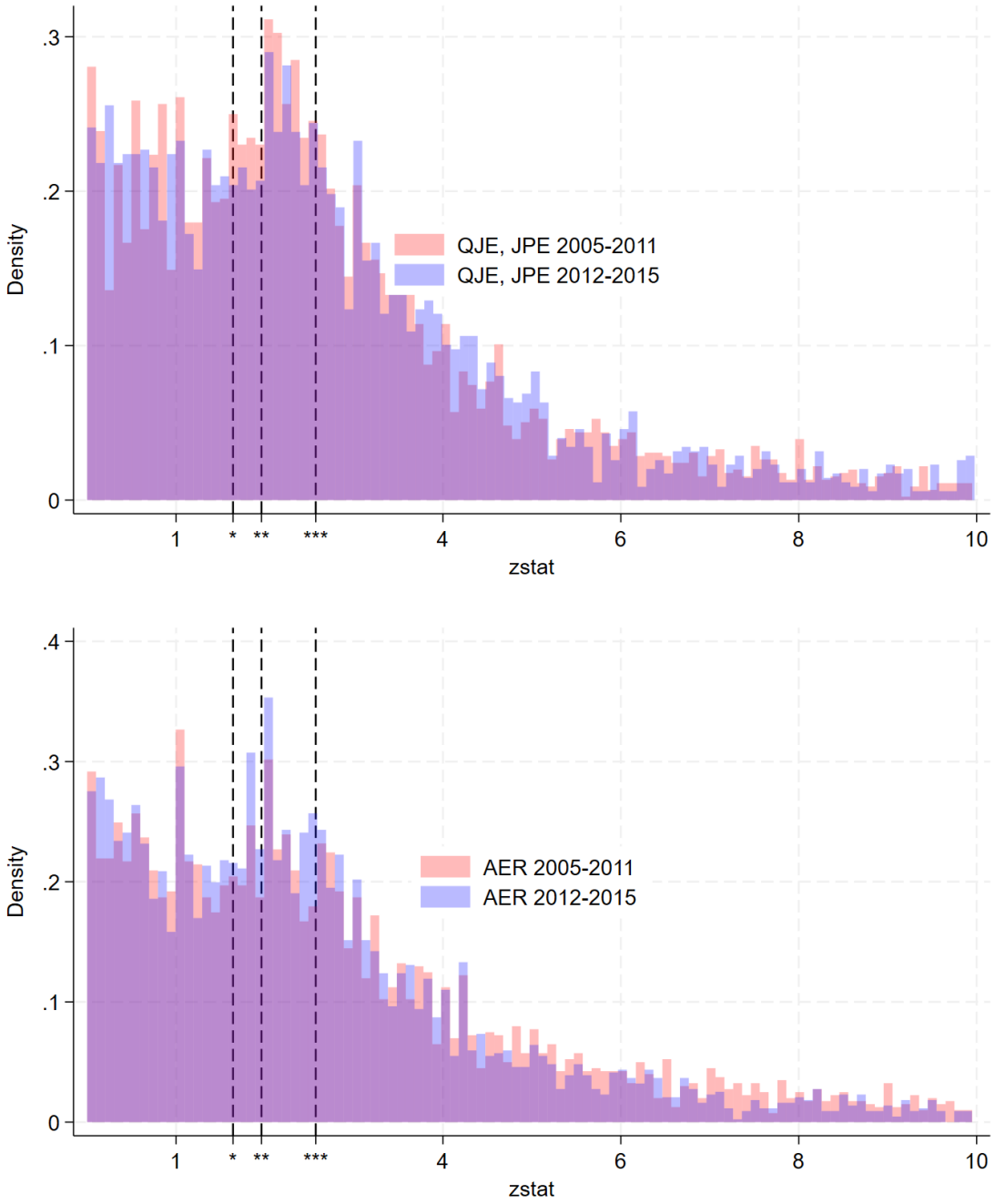

Figure 2: Histogram (100 bins) of the z-statistic values collected from QJE and JPE (upper panel) and AER (bottom panel), comparison of the periods 2007-2011 and 2012-2015. For QJE and JPE, $N = 4,975$ for period 2005-2011 and $N = 3,993$ for 2012-2015. For the AER, $N = 4,483$ for 2005-2011 and $N = 4,785$ for 2012-2015. Z-statistics larger than 10 have been trimmed in order to improve graph readability.

3 Difference-in-difference results

In this Section, I present the results of a difference-in-difference estimation. The unit of observation is a test statistic, and the model reads as follows:

$$\begin{aligned} Y_{aijt} = \quad & \alpha_0 + \alpha_1 Dummy20052011_{aijt} + \alpha_2 DoubleBlind_{aijt} \\ & + \beta Dummy20052011_{aijt} \times DoubleBlind_{aijt} + \varepsilon_{aijt} \end{aligned} \tag{1}$$

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. In this analysis, the treatment condition is defined as having a single-blind review process. As discussed earlier, my setup does not include a traditional untreated control group. Instead, there are two distinct sets of journals: those consistently treated with single-blind review (QJE, JPE), and a journal (AER) that switched from double-blind to single-blind review during the study period.

I estimate the model using a reverse-time difference-in-differences approach (following Kim and Lee (2018)). Practically, this involves redefining the treatment variable so that it equals 1 if the review process is double-blind and 0 if single-blind. Similarly, the time variable equals 1 during the period 2005–2011 and 0 afterward. Thus, my goal is to investigate whether the adoption of a double-blind review standard at AER (during 2005–2011) affected the prevalence of p-hacking. As a comparison group, I use similar journals (QJE and JPE) that maintained a single-blind review standard throughout the entire period.

With reference to equation (1), the interaction $Dummy20052011_{aijt} \times DoubleBlind_{aijt}$ represents the effect of the double-blind policy. Hence, the main coefficient of interest here is β. In Table 1, I report the results of the OLS estimation of equation (1) above[10]. The same estimation results, but considering significance levels of, respectively, 1 and 10% are reported in Appendix B (Tables 4 and 5). The interaction term in Table 1 captures the effect of employing a double-blind review system at AER between 2005 and 2011 on the probability that a published test statistic is significant at the 5% level. The estimates indicate heterogeneity across subsamples: while the interaction is negative and statistically significant for the "Theory model" and "Not top institution" categories, suggesting a reduction in significant results (potentially less p-hacking) under double-blind review, it is positive and significant in the "Single author" and "Top institution" subsamples, indicating an increase in significant results. This mixed evidence is consistent with the hypothesis

[10]Blanco-Perez and Brodeur (2020) show that using a logit model instead of OLS does hardly changes the results.

that double-blind review influences publication behavior differently depending on author characteristics and affiliations. In particular, I find confirmation for the hypothesis that well-known authors affiliated with prestigious universities tend to resort more to p-hacking-type behaviors under a double-blind review standard, whereas lesser known authors feel less pressure to present statistically significant results under the same review standard.

The effects are quantitatively relevant, as the double-review standard decreases the share of significant test statistics respectively by 5pp in papers with a theory model and by 10pp among papers whose authors do not come mostly from a top institution. Conversely, the double review standard increases the share of significant test statistics by around 5pp for papers without a theory model, by around 8pp among papers that are single-authored and by around 6pp for papers where at least half of authors are affiliated with a top institution. Overall, the result is a statistically insignificant reduction in the share of significant test statistics published, as reported in the first column of Table 1. This essentially zero effect is a "precisely estimated" zero, as the standard error is of the same order of magnitude as the relative estimated coefficient.

It appears that, under a double-blind standard, single authors as well as authors coming from top institutions have more incentive to present statistically significant results, as their identity will not be revealed to the referees. Results on the subsamples with and without a theory model are less robust to different specifications as will be detailed in the following.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Overall	Theory	No theory	Single aut	No single aut	Not top	Top inst
Double x '05-11	-0.014	-0.054	0.047	0.081	-0.024	-0.101	0.059
	(0.015)	(0.021)	(0.024)	(0.041)	(0.017)	(0.021)	(0.028)
Constant	YES	YES	YES	YES	YES	YES	YES
Year FEs	YES	YES	YES	YES	YES	YES	YES
Adj R-sq	0.006	0.013	0.016	0.040	0.009	0.012	0.023
Obs	18,236	9,857	8,379	3,329	14907	9,358	8,878
Articles	812	469	351	159	653	404	408

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**. Standard errors in parentheses. Top institutions are the 20 defined by Brodeur et al. (2020). The dummy is equal to one if at least half of the authors belong to a top institution at the time of the article publication.

Since the average time from submission to publication of a paper at the AER is around two years[11], in Table 6 in Appendix B I replicate the estimates of Table 1, this time setting the end of the double review standard at AER in 2013 instead than in 2011. The results

[11]See Hadavand et al. (2024) and the Report of the Editor American Economic Review (2024), who mentions: "The lag from submission to acceptance for articles published in 2023 was 115 weeks".

are essentially unchanged. I still find an overall zero impact on the probability that a published test statistics is significant at the 5% level, and a reduction (by 7pp) in the same probability for authors not at a top institution, together with an increase by around 6pp in the same variable for authors at a top institution. Single-authored papers still exhibit an increase by around 9pp in the probability that a published test statistics is significant at the 5%. However, in this model specifications the effects on the subsamples of papers with and without theoretical model are no longer significant, and their size is notably reduced as well.

Since different articles may include a different number of test statistics, in Appendix C I check the robustness of my results to weighting. I use as weights the inverse of the number of tests reported in each article. In this case (Table 7 in Appendix C) I find an overall reduction in the share of published test statistics that are significant at the 5% under the double-review standard. Results in the subsamples with and without a theory model are mixed. However, I find an (insignificant) increase by 6pp in the outcome variable for single authored papers and a (significant) decrease by around 5pp in the same variable among papers with multiple authors, which is broadly consistent with the unweighted results. Most notably, I still find a significant decrease (by around 17pp) in the probability that a published test statistics is significant for authors not coming from top institutions and conversely an increase by around 5pp in the same variable for authors affiliated with top institutions.

As mentioned in the Data section, my findings might be influenced by coarse rounding of the published coefficients and standard errors (or t-statistics, or p-value). In order to check whether this is the case, in Appendix D.1 I present my main results by applying the de-rounding method used by Brodeur et al. (2016), whereas in Appendix D.2 I replicate the main results by adopting the de-rounding method proposed by Kranz and Pütz (2022). Details on both derounding methods are provided in Appendix D.

In Table 8, where I apply the method suggested by Brodeur et al. (2016), I still find an overall insignificant impact of the double blind review standard on the probability that a published test statistics is significant. Further, the negative impact on the outcome variable for papers with a theory model (minus 4pp) and for authors not affiliated at top institutions (minus 7pp), the positive impact for papers without a theory model (+7pp) and with a single author (+9pp) as well as for authors coming from a top institution (+7pp) are still present, and statistically significant and close in size to the baseline results.

My results are broadly robust to derounding with the method of Kranz and Pütz (2022) as well. Indeed, in Table 10 I find evidence of a positive impact of the double

blind standard on the probability that a published test statistics is significant at the 5% level in the subsamples of single-authored papers (+5pp, insignificant) and of a negative impact on the same outcome in the subsamples of papers with a theory model (minus 8pp), multiple-authored papers (-7pp) and authors not affiliated with a top institution (-11pp). However, in this case the impact in the subsample of papers with authors coming from top institutions is negative and insigificant.

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

In this Section I aim at assessing whether I can detect p-hacking and publication bias under a double-blind, respectively a single-blind, review system. 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 five different tests: binomial, discontinuity, CS1, CS2B and LCM[12]. 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.

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)).

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. The LCM test (column 5 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). 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)). For both these tests, I consider p-values between 0 and 0.15 and I use 30 bins,

[12]I do not report the results of the Fisher test, as it as essentially zero power in most cases (see Elliott et al. (2022a and 2022b).

as in Elliot et al. (2022). Further, 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.

Name of test	Bin.	Disc.	CS1	CS2B	LCM	N. Obs.	N. articles
Panel A: AER, Double-Blind Review (2005-2011)							
Full sample	0.201	0.994	0.647	0.137	0.959	4483	222
Theory model	0.209	0.128	0.027	0.002	1.000	2557	129
No theory model	0.434	0.570	0.050	0.024	0.994	1926	93
Single author	0.412	0.594	0.000	0.000	1.000	975	51
Not single author	0.238	0.590	0.238	0.131	0.968	3508	171
Top institution	0.314	0.890	0.000	0.000	1.000	1924	92
Not top institution	0.292	0.082	0.474	0.607	1.000	2559	130
Panel B: AER, Single-Blind Review (2012-2015)							
Full sample	0.939	0.748	0.102	0.070	0.568	4785	200
Theory model	0.959	0.605	0.161	0.013	0.750	3144	131
No theory model	0.678	0.608	0.000	0.000	1.000	1641	74
Single author	0.685	0.528	0.004	0.000	1.000	754	32
Not single author	0.943	0.589	0.006	0.000	0.591	4031	168
Top institution	0.849	0.207	0.026	0.003	0.947	1886	74
Not top institution	0.900	0.172	0.243	0.040	0.957	2899	126

Table 2: P-values by Subsample: AER, Double-Blind vs. Single-Blind Review

From Table 2 I deduce that, under double-blind review (Panel A, 2005–2011), I can detect p-hacking in the subsamples of papers with and without a theory model, as well as single-authored and with authors coming from a top institutions (in all these cases, only hte CS1 and CS2B tests are able to reject the null of no p-hacking and no publication bias). No test rejects the null in the subsample of papers with multiple authors, as well as in the overall sample, and only the discontinuity test rejects the null among papers whose authors are not affiliated with top institutions. In Panel B (AER under a single-blind review standard), the null hypothesis is rejected by at least one test in all the subsamples as well as in the overall full sample. Similarly to panel A, only the CS1 and CS2B tests are able to detect p-hacking. It seems that it is easier to detect p-hacking under the single-blind regime. However, a word of caution is necessary, as some forms of selective reporting might not be detectable by the tests presented here and difference in significance do not correspond to significant differences (see Gelman and Stern (2006)).

In Appendix D I present the results of the tests reported in Table 2 when the two above-mentioned methods for derounding are applied. My results are broadly consistent to derounding. In Table 9 I apply the derounding method used in Brodeur et al. (2016)

and I find that, under double-blind review, I can detect p-hacking in the subsamples of papers without a theory model, both single-authored and with multiple authors, and with authors coming from top institutions. On the other hand, under single-blind review standard I can detect p-hacking in the subsamples of papers with and without a theory model, single-authored and with authors coming from top institutions.

In Table 11, where I apply the derounding method proposed by Kranz and Pütz (2022) I find evidence of p-hacking/publication bias under double-blind review standard in the full sample of papers, as well as in the subsamples of both single-authored and multi-authored papers, of papers without a theory model and with authors coming from top institutions. Under a single-blind review standard I am able to detect the presence of p-hacking in all the six subsamples considered, but not in the full sample.

5 Conclusion

This study explores the impact of single-blind versus double-blind peer review systems on the prevalence of p-hacking in economics journals, focusing on *The American Economic Review*'s (AER) transition to a single-blind review policy in 2011. Using a difference-in-differences approach and a series of statistical tests, the research provided suggestive hints that the change in the review system did not relevantly influence overall the extent of p-hacking in published papers at AER.

However, results notably differ across subgroups. In particular, the hypothesis according to which authors coming from top institutions engage more in p-hacking-type practices and authors coming from non top-institutions engage less in them under double-blind review and vice versa under a single-blind review system is confirmed empirically.

The results have broader implications for the ongoing debate between single-blind and double-blind review systems. Although proponents of double-blind reviews argue that they reduce biases and promote fairness, it might have the unintended consequence of increasing incentives for p-hacking for some groups of authors, in particular those who are single authors and come from top institutions. At the same time, a double-blind review system appears to reduce incentives for p-hacking-type practices across authors who are not affiliated with top universities and who coauthor papers with others.

References

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

2. Blank, R. M. (1991). The effects of double-blind versus single-blind reviewing: Experimental evidence from the American Economic Review. The American Economic Review, 1041-1067.

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

4. 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.

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

6. 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.

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

8. Card, D., & DellaVigna, S. (2020). What do editors maximize? Evidence from four economics journals. Review of Economics and Statistics, 102(1), 195-217.

9. Card, D., & DellaVigna, S. (2014). Page limits on economics articles: Evidence from two journals. Journal of Economic Perspectives, 28(3), 149-168.

10. Carlsson, F., Löfgren, Å., & Sterner, T. (2012). Discrimination in scientific review: A natural field experiment on blind versus non-blind reviews. The Scandinavian Journal of Economics, 114(2), 500-519.

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

12. Cressey, D. (2014, July 14). Journals weigh up double-blind peer review. Nature. Retrieved from https://www.nature.com/news/journals-weigh-up-double-blind-peer-review-1.15564

13. Elliott, G., Kudrin, N., & Wüthrich, K. (2022a). The Power of Tests for Detecting p-Hacking. arXiv preprint arXiv:2205.07950.

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

15. Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. The American Statistician, 60(4), 328-331.

16. Goldberg, P. (2012): "Report of the Editor: American Economic Review," American Economic Review: Papers & Proceedings, 102, 653–665.

17. Hadavand, A., Hamermesh, D. S., & Wilson, W. W. (2024). Publishing economics: How slow? Why slow? Is slow productive? How to fix slow?. Journal of Economic Literature, 62(1), 269-293.

18. Hengel, E. (2022). Publishing while female: Are women held to higher standards? Evidence from peer review. The Economic Journal, 132(648), 2951-2991.

19. Hill, S., & Provost, F. (2003). The myth of the double-blind review? Author identification using only citations. ACM SIGKDD Explorations Newsletter, 5, 179-184.

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

21. Kim, K., & Lee, M. J. (2019). Difference in differences in reverse. Empirical Economics, 57, 705-725.

22. 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.

23. Laband, D. N., & Piette, M. J. (1994). Does the" blindness" of peer review influence manuscript selection efficiency?. Southern Economic Journal, 896-906.

24. Erzo F.P. Luttmer, (2024). Report of the Editor American Economic Review. AEA Papers and Proceedings 2024, 114: 734–750.

25. McCloskey, A., & Michaillat, P. (2024). Critical values robust to p-hacking. Review of Economics and Statistics, 1-35.

26. Naguib, C. (2024). P-hacking and Significance Stars. Discussion paper series. University of Bern.

Appendix (for online publication only)

A. Descriptive statistics

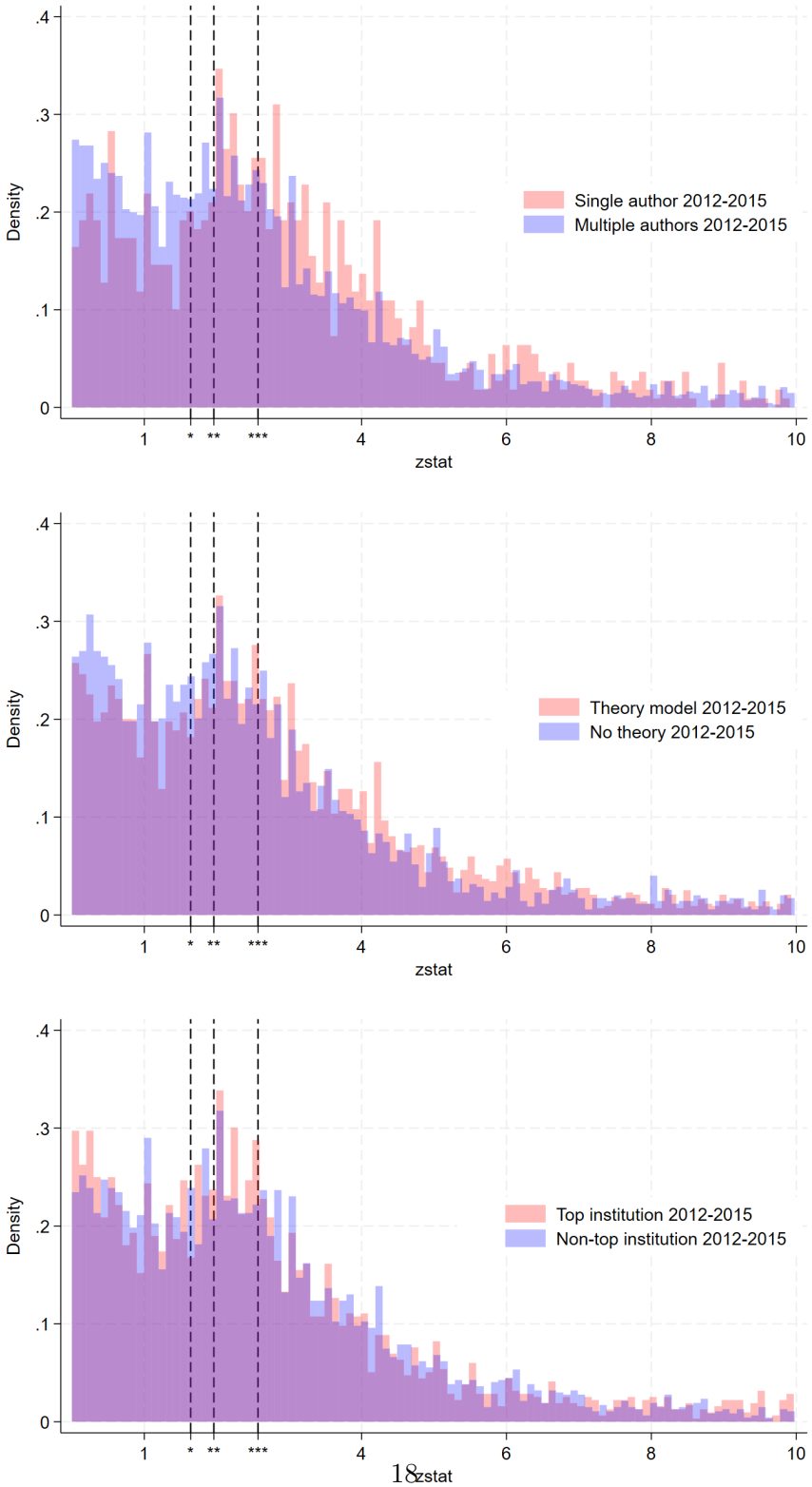

Figure 3: Histogram (100 bins) of the z-statistic values collected from journals in the AER, QJE and JPE for the period 2012-2015, i.e. when all these journals were adopting **single-blind review**, by subgroups. Z-statistics larger than 10 have been trimmed in order to improve graph readability.

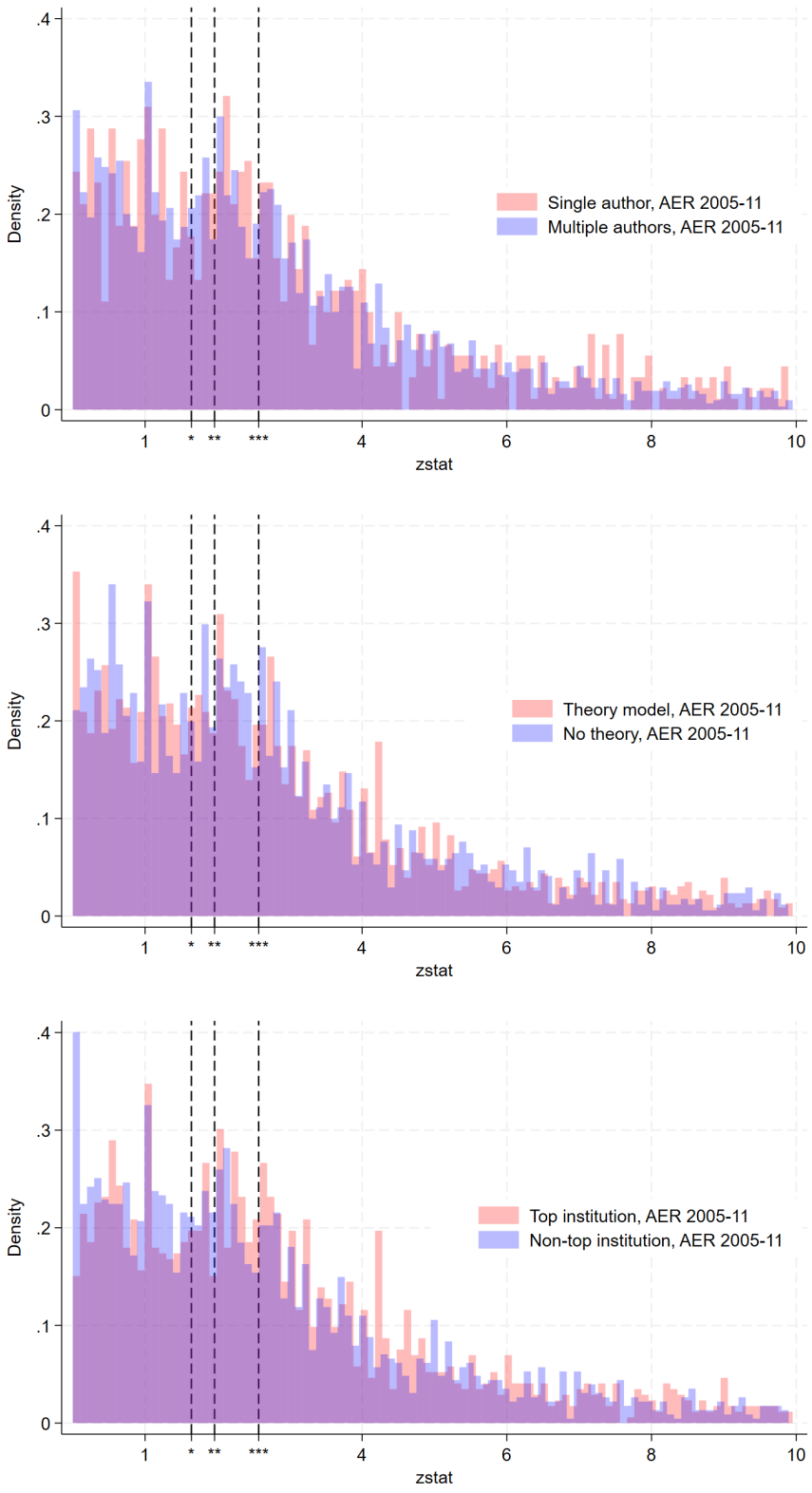

Figure 4: Histogram (100 bins) of the z-statistic values collected from journals in the AER for the period 2005-2011, i.e. under a regime of **double-blind review**, by subgroups. Z-statistics larger than 10 have been trimmed in order to improve graph readability.

	% of Articles	% of Tests	N. Articles	N. Tests
AER	51.97%	50.82%	422	9268
QJE	32.88%	33.75%	267	6154
JPE	15.15%	15.43%	123	2814
Single-authored	19.58%	18.26%	159	3329
With a theoretical model	57.14%	54.05%	464	9857
Share of authors from top inst	50.25%	48.68%	408	8878

Table 3: Descriptive statistics of the treated and control group samples. Period 2005-2015.

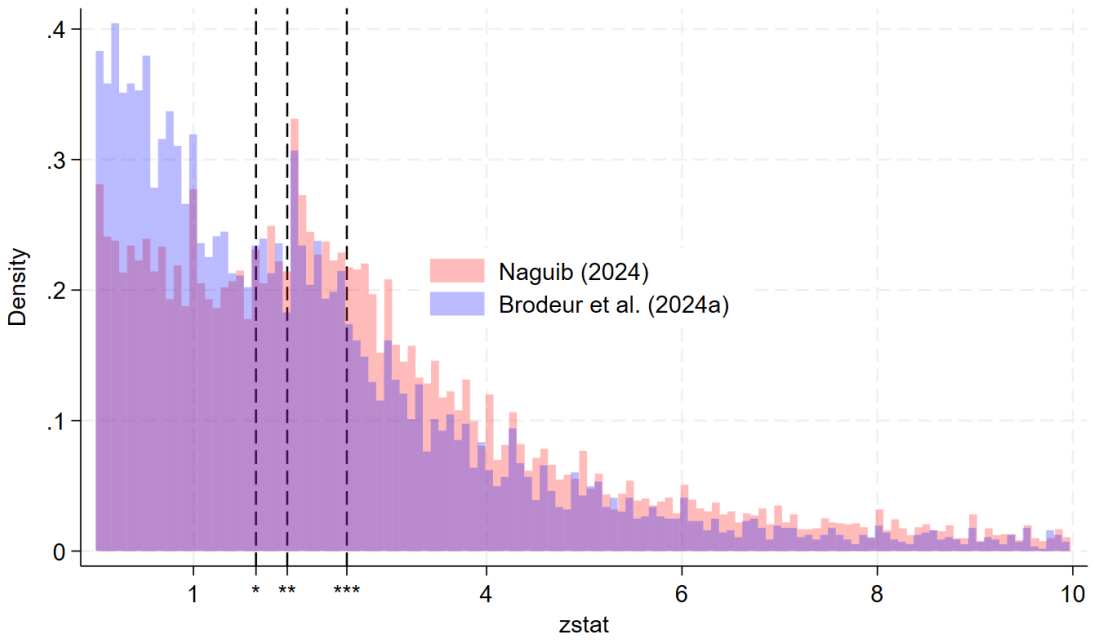

Figure 5: Histogram (125 bins) of the z-statistic values collected from QJE, JPE and AER, for the period 2005-2015 by Naguib (2024) and Brodeur et al. (2024a). Z-statistics larger than 10 have been trimmed in order to improve graph readability. In the data collected by Naguib (2024) $N = 18,236$, whereas in the data collected by Brodeur et al. (2024a) $N = 7,658$. * = 1.65, ** = 1.96, and *** = 2.58.

B. Additional empirical results

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Overall	Theory	No theory	Single aut	No single aut	Not top	Top inst
Double x 2005-11	0.032	-0.004	0.104	0.064	0.032	-0.055	0.107
	(0.016)	(0.022)	(0.024)	(0.042)	(0.017)	(0.021)	(0.024)
Constant	YES	YES	YES	YES	YES	YES	YES
Year FEs	YES	YES	YES	YES	YES	YES	YES
Adj R-sq	0.005	0.010	0.017	0.037	0.007	0.010	0.013
Obs	18,236	9,857	8,379	3,329	14907	9,358	8,878
Articles	812	469	351	159	653	404	408

Table 4: 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**. Standard errors in parentheses. Top institutions are the 20 defined by Brodeur et al. (2020). The dummy is equal to one if at least half of the authors belong to a top institution at the time of the article publication.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Overall	Theory	No theory	Single aut	No single aut	Not top	Top inst
Double x 2005-11	-0.024	-0.062	0.026	0.075	-0.034	-0.104	0.053
	(0.015)	(0.020)	(0.023)	(0.039)	(0.016)	(0.020)	(0.022)
Constant	YES	YES	YES	YES	YES	YES	YES
Year FEs	YES	YES	YES	YES	YES	YES	YES
Adj R-sq	0.006	0.011	0.016	0.038	0.009	0.010	0.018
Obs	18,236	9,857	8,379	3,329	14907	9,358	8,878
Articles	812	469	351	159	653	404	408

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 **10% level**. Standard errors in parentheses. Top institutions are the 20 defined by Brodeur et al. (2020). The dummy is equal to one if at least half of the authors belong to a top institution at the time of the article publication.

The median lag in economics from paper submission to publication is around two years. For this reason, in this Section I present some of the baseline estimates by setting the starting date of the change in the peer-review standard to 2013, instead than to 2011. Indeed, the first papers subject to the double review standard at AER were most likely not published until 2013. I focus here on the 5% significance threshold only.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Overall	Theory	No theory	Single aut	No single aut	Not top	Top inst
Double x **'05-13**	-0.003	-0.005	0.005	0.091	-0.006	-0.074	0.063
	(0.016)	(0.023)	(0.025)	(0.045)	(0.018)	(0.022)	(0.025)
Constant	YES	YES	YES	YES	YES	YES	YES
Year FEs	YES	YES	YES	YES	YES	YES	YES
Adj R-sq	0.006	0.012	0.016	0.040	0.009	0.010	0.016
Obs	18,236	9,857	8,379	3,329	14907	9,358	8,878
Articles	812	469	351	159	653	404	408

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 **5% level**. Standard errors in parentheses. Top institutions are the 20 defined by Brodeur et al. (2020). The dummy is equal to one if at least half of the authors belong to a top institution at the time of the article publication.

C. Robustness checks with weighting

In this Section, I replicate the main tables and figures of the paper using article weights, in order to avoid that articles with more tests have a disproportionate influence on the results. Following Brodeur et al. (2016), to obtain the results with article weights I associate to each test statistic the inverse of the total number of tests that are reported in the same article. The result is that each article contributes in the same way to the distribution.

Since I focus on the main results of each paper, for most of the papers in my sample I collect coefficients from one table only. For this reason I refrain from reporting results obtained with article and table weigths. Indeed, they would be essentially identical to the ones obtained with article weights, which are reported in the following.

Similar to Brodeur et al. (2016) the "camel shape" of z-statistics is more evident for the weighted than for the unweighted distributions. This supports the hypothesis that researchers are likely to report more estimated coefficients if their results are statistically significant and conversely they only report a few if other specifications would fail to yield statistically significant results. Weighted distributions give less weight to articles and tables in which many tests are reported.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Overall	Theory	No theory	Single aut	No single aut	Not top	Top inst
Double x 2005-11	-0.050	-0.027	-0.081	0.060	-0.054	-0.168	0.045
	(0.019)	(0.025)	(0.029)	(0.045)	(0.021)	(0.026)	(0.028)
Constant	YES	YES	YES	YES	YES	YES	YES
Year FEs	YES	YES	YES	YES	YES	YES	YES
Adj R-sq	0.017	0.022	0.031	0.039	0.022	0.034	0.034
Obs	18,236	9,857	8,379	3,329	14907	9,358	8,878
Articles	812	469	351	159	653	404	408

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**, **with article weights**. Standard errors in parentheses. Top institutions are the 20 defined by Brodeur et al. (2020). The dummy is equal to one if at least half of the authors belong to a top institution at the time of the article publication.

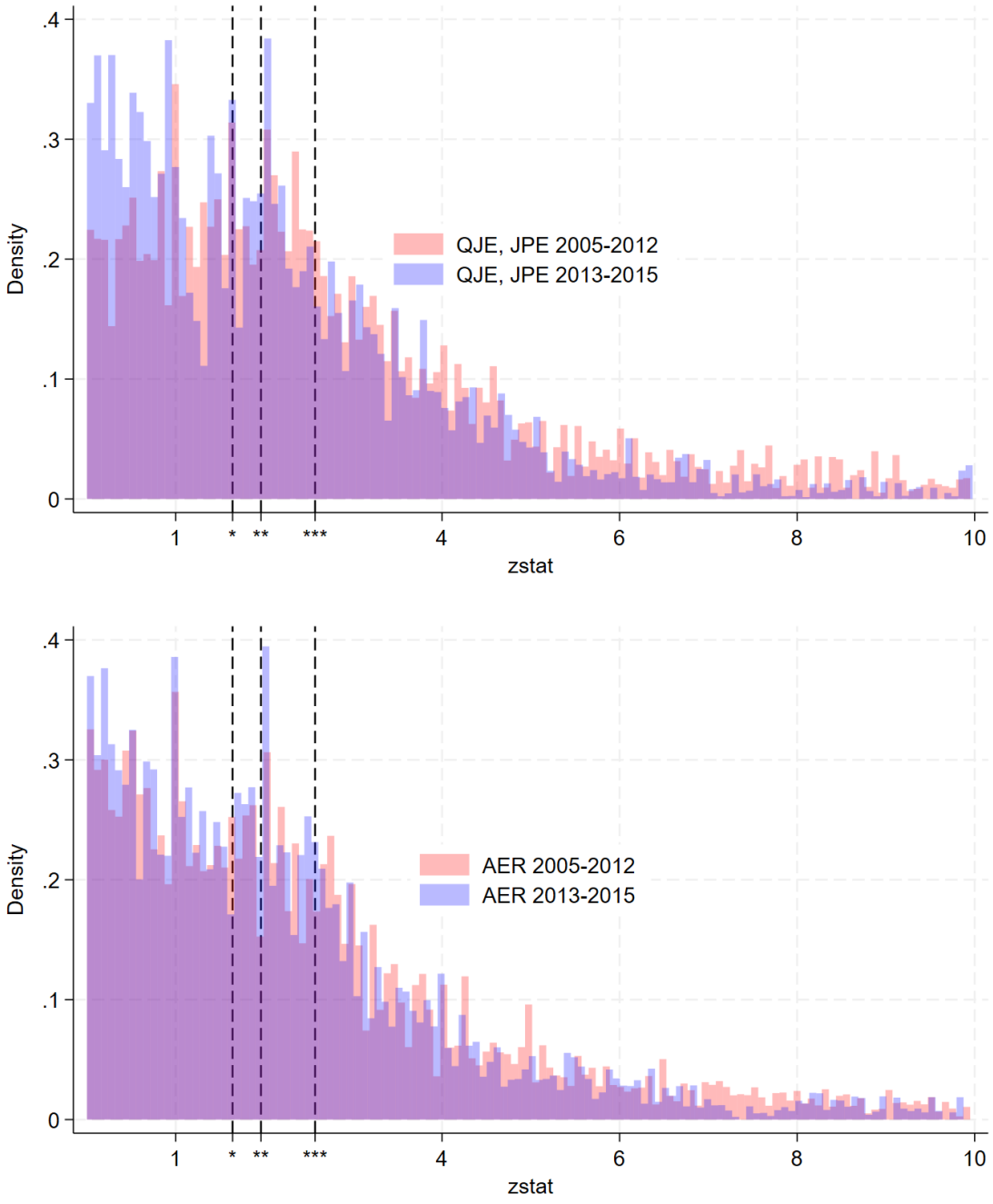

Figure 6: Histogram (125 bins) of the z-statistic values collected from journals in the control group (QJE and JPE, upper panel) and in the treatment group (AER, lower panel), with **article weights**, comparison of the periods 2005-2012 and 2013-2015. Z-statistics larger than 10 have been trimmed in order to improve graph readability. * = 1.65, ** = 1.96, and *** = 2.58.

D. Robustness to de-rounding

D.1 De-rounding with the method of Brodeur et al. (2016)

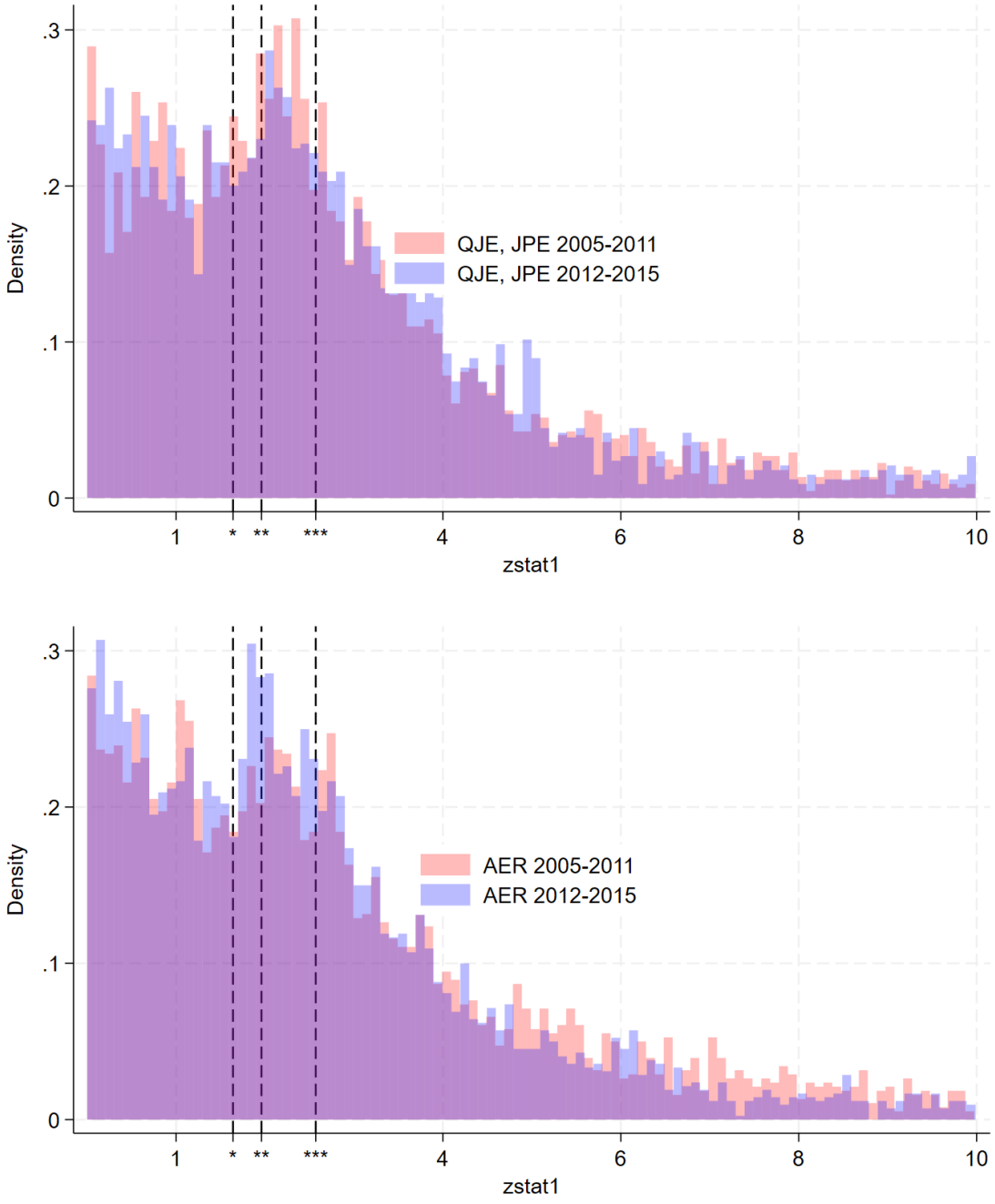

Figure 7: Histogram (125 bins) of the z-statistic values collected from journals in the control group (upper panel) and in the treatment group (lower panel), with **derounding as in Brodeur et al. (2016)**, comparison of the periods 2005-2012 and 2013-2015. Z-statistics larger than 10 have been trimmed in order to improve graph readability. * = 1.65, ** = 1.96, and *** = 2.58.

	(1) Overall	(2) Theory	(3) No theory	(4) Single aut	(5) No single aut	(6) Not top	(7) Top inst
Double x 2005-11	0.002	-0.042	0.067	0.091	-0.005	-0.070	0.074
	(0.016)	(0.021)	(0.024)	(0.041)	(0.017)	(0.022)	(0.029)
Constant	YES	YES	YES	YES	YES	YES	YES
Year FEs	YES	YES	YES	YES	YES	YES	YES
Adj R-sq	0.006	0.011	0.017	0.039	0.009	0.009	0.023
Obs	17,467	9,332	8,135	3,304	14,163	8,875	5,884
Articles	771	439	340	156	615	381	390

Table 8: 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**. Standard errors in parentheses. Top institutions are the 20 defined by Brodeur et al. (2020). **De-rounding as in Broduer et al. (2020)**. The dummy is equal to one if at least half of the authors belong to a top institution at the time of the article publication.

Name of test	Bin.	Disc.	CS1	CS2B	LCM	N. Obs.	N. articles
Panel A: AER, Double-Blind Review (2005-2011)							
Full sample	0.409	0.266	0.801	0.510	1	4236	207
Theory model	0.441	0.461	0.861	0.614	1	4236	207
No theory model	0.500	0.608	0.519	0.017	1	1853	87
Single author	0.324	0.444	0.005	0.000	1	971	50
Not single author	0.553	0.260	0.894	0.091	1	3265	157
Top institution	0.932	0.757	0.448	0.065	1	1818	86
Not top institution	0.092	0.434	0.107	0.178	1	2418	121
Panel B: AER, Single-Blind Review (2012-2015)							
Full sample	0.998	0.295	0.489	0.108	1	4576	189
Theory model	0.957	0.276	0.037	0.012	1	2955	122
No theory model	0.994	0.501	0.595	0.048	1	1621	72
Single author	0.928	0.808	0.008	0.000	1	742	31
Not single author	0.995	0.583	0.447	0.113	1	3834	158
Top institution	0.986	0.769	0.001	0.000	1	1816	69
Not top institution	0.976	0.454	0.813	0.056	1	2760	120

Table 9: P-values by Subsample: AER, Double-Blind vs. Single-Blind Review, De-rounding with the method of Brodeur et al. (2016).

D.2 De-rounding with the method of Kranz and Pütz (2022)

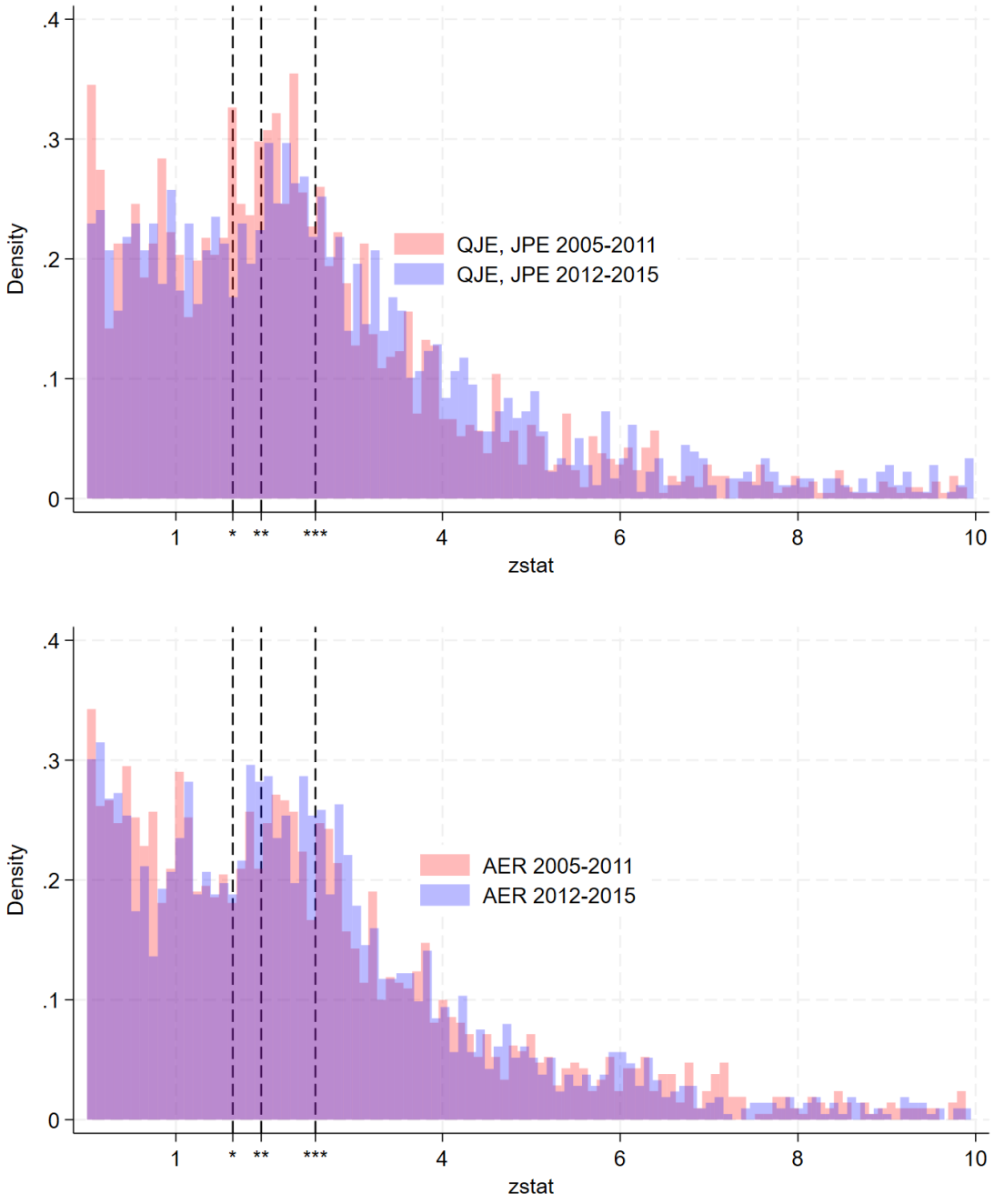

Figure 8: Histogram (125 bins) of the z-statistic values collected from journals in the control group (upper panel) and in the treatment group (lower panel), with **derounding as in Kranz and Pütz (2022)**, comparison of the periods 2005-2012 and 2013-2015. Z-statistics larger than 10 have been trimmed in order to improve graph readability. * = 1.65, ** = 1.96, and *** = 2.58.

	(1) Overall	(2) Theory	(3) No theory	(4) Single aut	(5) No single aut	(6) Not top	(7) Top inst
Double x 2005-11	-0.060	-0.081	-0.029	0.054	-0.067	-0.106	-0.084
	(0.022)	(0.031)	(0.034)	(0.053)	(0.025)	(0.031)	(0.043)
Constant	YES	YES	YES	YES	YES	YES	YES
Year FEs	YES	YES	YES	YES	YES	YES	YES
Adj R-sq	0.013	0.013	0.020	0.062	0.014	0.016	0.033
Obs	8,654	4,569	4,085	1,805	6,849	4,611	2,734
Articles	654	370	290	134	520	326	328

Table 10: 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**. Standard errors in parentheses. Top institutions are the 20 defined by Brodeur et al. (2020). **De-rounding as in Kranz and Pütz (2022)**. The dummy is equal to one if at least half of the authors belong to a top institution at the time of the article publication.

Name of test	Bin.	Disc.	CS1	CS2B	LCM	N. Obs.	N. articles
Panel A: AER, Double-Blind Review (2005-2011)							
Full sample	0.326	0.271	0.083	0.049	1.000	2219	174
Theory model	0.351	0.497	0.117	0.131	1.000	1273	100
No theory model	0.500	0.915	0.000	0.000	1.000	946	74
Single author	0.726	0.059	0.000	0.000	1.000	576	45
Not single author	0.243	0.927	0.579	0.043	1.000	1643	129
Top institution	0.685	0.789	0.082	0.000	1.000	866	74
Not top institution	0.221	0.921	0.145	0.188	1.000	1353	100
Panel B: AER, Single-Blind Review (2012-2015)							
Full sample	0.932	0.968	0.115	0.267	1.000	2260	153
Theory model	0.828	0.747	0.000	0.000	0.994	1386	98
No theory model	0.928	0.782	0.000	0.000	0.997	874	59
Single author	0.746	0.527	0.108	0.081	1.000	416	25
Not single author	0.934	0.344	0.021	0.000	1.000	1844	128
Top institution	0.968	0.752	0.175	0.001	1.000	820	59
Not top institution	0.721	0.303	0.027	0.005	1.000	1440	94

Table 11: P-values by Subsample: AER, Double-Blind vs. Single-Blind Review, derounding with the method by Kranz and Pütz (2022).