de Gendre, Alexandra; Salamanca, Nicolás

Working Paper

Reproduction Report on "Unpacking p-Hacking and Publication Bias"

I4R Discussion Paper Series, No. 305

Provided in Cooperation with:
The Institute for Replication (I4R)

Suggested Citation: de Gendre, Alexandra; Salamanca, Nicolás (2026) : Reproduction Report on "Unpacking p-Hacking and Publication Bias", I4R Discussion Paper Series, No. 305, Institute for Replication (I4R), s.l.

This Version is available at:
https://hdl.handle.net/10419/342310

INSTITUTE for REPLICATION

No. 305
DISCUSSION PAPER SERIES

Reproduction Report on "Unpacking p-Hacking and Publication Bias"

Alexandra de Gendre

Nicolás Salamanca

This paper received a response:

Brodeur, A., S. Carrell, D. Figlio, and L. Lusher. 2026. Response to de Gendre and Salamanca's Comment on "Unpacking p-Hacking and Publication Bias". *I4R Discussion Paper Series* No. 306. Institute for Replication.

July 2026

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 305

Reproduction Report on "Unpacking p-Hacking and Publication Bias"

Alexandra de Gendre[1], Nicolás Salamanca[1]

[1]The University of Melbourne/Australia

JULY 2026

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

Reproduction Report on
"Unpacking p-Hacking and Publication Bias"

Alexandra de Gendre (The University of Melbourne)
and
Nicolás Salamanca (The University of Melbourne)

May 2026

Abstract

Brodeur, Carrell, Figlio and Lusher (2023) present the first descriptive evidence on p-hacking and publication bias throughout the publication process in economics. Using data from the Journal of Human Resources, they show that 1) p-values in initial submissions present humping around the 10 and 5% significance thresholds, 2) the distribution of desk rejections shows more humping than that of papers sent for review, 3) favorable reviewer recommendations are positively related to statistical significance and 4) overall, the p-value distribution of initial and final versions of accepted drafts appear similar. We first conduct a mostly successful computational reproduction of those results. Next, we conduct a series of robustness replication exercises and show that, while the authors' main results are robust to random removals of small amounts of data and to bandwidth choices, they are sensitive to removing papers that contribute many estimates, to some key modelling choices in density discontinuity tests, and to some conceptually equivalent variations of their tests. Last, we complement the authors' results by providing statistical tests of their main hypotheses. Our results support the claims that 1) the distribution of p-values of desk-rejected papers displays greater bunching than that of papers sent for review, and 2) the distributions of p-values of initial submissions and final drafts are similar. Overall, our findings lend support to the authors' main claims, suggesting that after excluding desk-rejected papers, the peer review process does not seem to affect p-hacking or publication bias.

1 Introduction

In the main results, Brodeur, Carrell, Figlio and Lusher (2024, henceforth BCFL) present the first descriptive evidence on p-hacking and publication bias throughout the publication process in economics from a random sample of papers submitted to *The Journal of Human Resources*. Their final sample contains 705 manuscripts submitted between 2013 and 2018 and handled across 28 editors: 171 desk rejections, 210 rejections after receiving reviews, 162 drafts of eventually accepted and published manuscripts, and 20,206 tests of main hypotheses. They analyze data at each stage of the review process, from initial submission until acceptance, by a) testing for evidence of p-hacking and publication bias at each stage using a battery of five tests that jointly test p-hacking and publication bias proposed by Elliott, Kundrin and Wütirch (2022) (henceforth EKW), b) visually inspecting differences in the distributions of *p*-values across stages in the publication process, and c) characterizing observable differences between point estimates right below and right above relevant conventional significance thresholds using Caliper tests (see e.g., Gerber and Malhotra, 2008; Brodeur, Cook and Heyes, 2020). The paper's main findings, all summarized on page 2976, are that:[1]

1. "the distribution of test statistics among submitted articles displays a hump around 10 and 5% significance thresholds",
2. "the distribution of desk rejections displays greater bunching than those sent for review",
3. "this [last] result is partially explained by author characteristics",
4. "favorable reviewer recommendations are positively related to statistical significance", and

[1] Brodeur, Carrell, Figlio and Lusher (2023) also produce additional results and investigate other questions related to p-hacking and publication bias. Namely, they first visually inspect the distribution of p-values at each stage of the publication process to detect evidence of p-hacking and publication bias. They also collect survey data from published authors in economics to explore the potential reasons behind p-hacking. Their robustness exercises mostly focus on their Caliper tests and whether their results are affected by the inclusion of "ambiguous results" as coded by them, to only including the first set of main estimates for every paper, and to the use of wider bandwidths. They also explore the robustness of some of their exercises to de-rounding p-values (see e.g., Brodeur et al. 2016; Bruns et al. 2019; Kranz and Putz, 2022). In this report we largely ignore these additional analyses and focus on the article's main claims.

5. “the distribution of statistics from the final draft of accepted manuscripts is similar to its initial draft counterpart”.

In this comment, we first conduct computational reproducibility exercises. We focus on only a subset of BCFL’s results since part of their data is confidential and not publicly available. We first reproduce their Table 2 using the authors’ original data and code in *R*, find no coding errors, and can perfectly reproduce the vast majority estimates. However, we found discrepancies in two key results in Table 2 related to initial submissions: in the heaping tests around the 10% significance threshold, we find a *p*-value of 0.067 (instead of BCFL’s *p*-value of <0.001) for the Binomial test, and of 0.277 (instead of BCFL’s *p*-value of 0.004) for the density discontinuity test. These discrepancies could potentially alter one of the paper’s main claims. We also performed a successful near-enough computational reproduction (i.e., a reproduction where we expected to reproduced their results very closely, but not exactly) of BCFL’s Caliper tests without covariates (Tables 4 through 8, Column 1).

We further computationally reproduce the density discontinuity tests in Table 2 using BCFL’s data but our own code in *Stata*, instead of using their code in *R*. We failed to computationally reproduce all discontinuity tests and identified the reason to be tiny differences in the optimal bandwidths selected in *R* and *Stata*, likely due to differences in their underlying numerical optimization algorithms. These tiny differences in bandwidth often lead to moderate differences in density discontinuity test *p*-values that, while mostly inconsequential, could have altered one of the article’s main findings.

Next, we conduct several robustness reproduction exercises. We find BCFL’s results in Table 2 to be very robust to randomly dropping 1% of articles, but somewhat sensitive to dropping the 1% of articles that contribute the most estimates. We also find their density discontinuity tests to be sensitive to the specific modelling choices, and to using $|z|$-statistics rather than *p*-values as the running variable. We find BCFL’s caliper tests without covariates to be very robust to alternative bandwidth choices, but less robust to a conceptually equivalent test using regression discontinuities in $|z|$-statistics rather than in *p*-values.

Finally, we reanalyze BCFL's data in an effort to provide formal statistical tests to complement the authors' findings of differences in heaping across the editorial process. Most of their main findings are supported by graphical evidence, which prevent any inference, and on results from Caliper tests, which we cannot reproduce with the available data. However, we reconstruct the joint variance-covariance matrix of all density discontinuity estimates in BCFL's Table 2 and, using this matrix, formally test all of the authors' hypotheses regarding heaping differences between subsamples. Our results support the claims that 1) the distribution of p-values of desk-rejected papers displays greater bunching than that of papers sent for review, and 2) the distributions of p-values of initial submissions and final drafts are similar.

In the conclusions, we combine the evidence reproduced in this comment with the evidence presented in BCFL to revisit their conclusions. Overall, our findings lend support to the authors' main claims, suggesting that after excluding desk-rejected papers, the peer review process does not seem to affect p-hacking or publication bias.

2 Computational Reproducibility

In this section we describe the completeness and quality of the replication package provided with the paper and describe our computational reproduction.

2.1 Completeness of the replication package

We used the replication package available here. The replication package contains a detailed *Read Me* and is logically organized, making it easy to understand which code is meant to produce which table and figure. The package is organized into code and data subdirectories. The code directory contains clearly labeled analysis code for each table and figure in the paper. Within each table-specific subdirectory there is code and additional datasets in *.csv* format to be used as input into the tests coded in *R*. The data directory contains three additional datasets in *Stata* format.

[TABLE 1 HERE]

For confidentiality reasons, the replication package does not provide raw data, cleaning code for raw data, nor several analysis files. We therefore can only work from partial analysis data and code. Given these restrictions, we could not reproduce much of the evidence from the paper, and more specifically, the part of Table 1 which contains summary statistics on confidential regressors, Tables 3 through 8 and Appendix Tables A5 through A34, all of which are Caliper tests using confidential regressors.[2]

We focus on the computational reproducibility of Table 2 and accompanying Appendix Table A1, and of Appendix Tables A3 and A4. We also perform a "near-enough computational reproducibility" of the first column in Tables 4 through 8, which present results from various Caliper tests without covariates. For our robustness reproducibility exercises, we focus on the results in Tables 2 and A1, and on Column 1 of Tables 4 through 8. Given the confidentiality of the data, we cannot attempt Recreate Reproducibility, Direct nor Conceptual Replication exercises.[3]

With the evidence we can reproduce we cannot test the robustness of all the authors' main claims. Specifically, we cannot test the robustness of the evidence used to claim that the differences in heaping in the distribution of desk rejections versus that of articles sent for review is partially explained by author characteristics. We also cannot test the robustness of the evidence that favorable reviewer recommendations are positively related to statistical significance using the authors' own methods, though we can use different methods and the available data to test evidence related to this claim (see Section 5).

2.2 Using the original data, code and statistical software

Tables 2 and 3 below shows our computational reproduction of BCFL's Table 2 (with number of observations reported in Appendix Table A1 in the original paper) and

[2] We also reproduce the figures in the main text of the paper. However, these figures generally do not provide core results. Rather, they serve to give a visual description of the data for support, but not formal testing, of the authors' main claims. We therefore focus on reproducing tables. Whenever estimates in tables were copy-pasted from figures, we reproduce the figure to obtain the estimate for the table.

[3] See the Institute for Replication's template for a replication report in economics (accessible here) for relevant definitions of *Computational Reproducibility*, *Recreate reproducibility*, *Robustness Reproducibility*, *Direct Replicability*, *Conceptual Replicability*.

Appendix Table A3 (and number of observations from Appendix table A4) respectively. Overall, we can reproduce the vast majority of the original estimates, with the following six exceptions:

1. Table 2, Row 1 (*Initial Submissions*), Columns 3 and 4 (*10% significance*) (also in Table A1, Row 1, Columns 4-5): we find that the p-values of the Binomial test to be 0.067 (instead of 0.000) and that of the Discontinuity test to be 0.277 (instead of 0.004), changing the significance of the test. The number of observations (reported in Table A1, row 1, column 6) does however reproduce.
2. Table 2, Row 2 (*Desk rejections*), Column 6: we find a slightly larger p-value (0.270) than the one reported by Brodeur et al. (0.255), although this difference does not change the significance of the test.
3. Table 2, Row 7 (*Initial Submission of Accepted Manuscripts*), Columns 1 and 2 (5% significance) (also in Table A1, Row 7, Columns 1-3): those estimates are not computationally reproducible. We first note that, for all tables, the code for all results printed in the tables at the 5% significance only produces figures. We work under the assumption that the table values were copied from these figures. However, estimates in figures produced for Row 7 do not match the numbers reported in the table. We believe this error is likely caused by a file that was wrongfully overwritten in the replication package. Our own reconstruction of the replication code reproduced these numbers accurately.
4. Table 2, Row 10 (*Accepted Manuscripts*) (also in Table A1, Row 10): those estimates are not computationally reproducible. The data file for reproducing this row seems to be missing from the package. We deduced that this is likely the same data file as the one used for producing Row 8 (Published Version) since the numbers in both rows are identical for all columns and the number of articles in both publication process stages is the same (162 papers, as reported on page 2978). However, the corresponding rows in Table A1 (which allegedly reproduces Table 2 and just adds the number of observations) are supposed to be identical, but they are not. For these same missing dataset issues, we cannot computationally reproduce the entire Row 10 of Table A1 (Accepted Manuscripts).

5. Table A3, Row 4 (*Recommend rejection*), Columns 1 and 2 (*5% significance*) (also in Table A4, Row 4, Columns 1 through 3): those estimates are not computationally reproducible. As in Exception #3 above, the code for these results only produces figures and we work under the assumption that the table values were copied from these figures. Yet, the code for producing the figure providing estimates for Row 4 does not run. Our own reconstruction of the replication code however reproduced these numbers accurately.
6. Table A3, Row 6 (*Recommend accept as is or with minor edits*), Column 2: we find a tiny difference in the third decimal of the p-value (0.666) compared to the one reported by Brodeur et al. (0.665). This difference is inconsequential for the conclusions of the test and is likely due to a difference in rounding at the third decimal digit (or an unwillingness to report 'the number of the Beast' in a Top 5 publication).

[TABLE 2 HERE]

[TABLE 3 HERE]

However, the differences documented in the first exception above *could change the main conclusions of the paper*. The five tests of p-hacking and publication bias at the 10% confidence level now provide ambiguous evidence: three of these tests would reject the null of no p-hacking at the 1% significance level, one test would reject the null but only at the 10% significance level (Binomial test) and one test would not reject the null (Discontinuity test). This weakens the authors' main finding that p-values in initial submissions present heaping around the 10 and 5% significance thresholds.[4]

Table 4 below shows our "near-enough computational reproduction" of BCFL's Column 1 in Tables 4 through 8, which show Caliper tests without covariates across subsamples. We could not attempt an exact reproduction of these results since the samples used in the paper exclude observations with missing confidential covariates,

[4] We also uncovered one seemingly inconsequential oddity. In the paper, the authors report that their "[…] final analytic sample contains 705 manuscripts […]" (pg. 2978). However, in the replication package, the dataset "a_initial_submissions.csv", which should contain all these manuscripts, only has 529 unique study identifiers (as marked by the variable *id*). However, this difference between the reported number of manuscripts in the paper and the number of unique manuscripts in the dataset does not seem to affect the reproducibility of most estimates in Tables 2, A1, A3 and A4.

which we do not have access to. However, there are only very few observations with missing covariates in all subsamples (e.g., 29 out of 2,056 observations for the desk rejection analysis) so we would expect our near-enough computational reproduction to be very close to the original results. This reproduction was done in *Stata* using the paper-by-estimate level data and the referee-by-estimate level data files available in the replication package and using BCFL's own code provided for use with the confidential data. Overall, we found Brodeur et al.'s results from this analysis to be *Computationally Reproducible from Analysis Data*.

[TABLE 4 HERE]

2.3 Using *Stata* as alternative statistical software

We also used *Stata* as an alternative statistical software to conduct further computational reproducibility tests. We focus on the estimates of discontinuity tests presented in Table 2, Columns 2 and 4. We use Stata's *rddensity* command (Cattaneo, Jansson and Ma, 2018, 2020) which is the companion of *R*'s *rddensity* package provided by the same team.

We failed to computationally reproduce the discontinuity test estimates using *Stata*'s *rddensity* instead of *R*'s *rddensity*. These differences are not cause by different default options being used in both versions of the *rddensity* package; we confirmed that all estimation options used are the same in both packages. We have traced down the issue to differences in the behavior of the *rdbwselect* subpackage used. This subpackage estimates slightly different optimal bandwidths in *R* and *Stata*, even when using the same preferred procedure to do so (e.g., choosing bandwidths based on the Mean Squared Error of the sum of the two density estimators below and above the threshold). This is likely due to software differences in the numerical optimization algorithms. When we manually impose the bandwidths at both sides of the threshold (through the option *h = c(#1,#2)* in *R* and the option *h(#1 #2)* in *Stata*), both packages yield the same results.

Table 5 below shows the results of this exercise. The slight differences in optimal bandwidth choice between both softwares lead to moderate differences in density discontinuity test *p*-values that are mostly inconsequential, could nevertheless alter some of the article's conclusions. For example, while BCFL's density discontinuity test using *R*

finds no evidence of heaping at the 10% level in the distribution of papers that passed the des rejection stage (*p*-value = 0.184), our reproduction using *Stata* find marginally significant evidence of heaping (*p*-value = 0.080). Somewhat unnervingly, the difference in *p*-values is driven by a wider optimal bandwidth in *Stata* than in *R* by only one observation on each side of the discontinuity.

[TABLE 5 HERE]

3 Robustness Reproduction

After reading the paper and conducting our computational reproducibility, we conducted a series of robustness reproduction exercises. Our code for these exercises is available at (hyperlink).

3.1 The influence of individual papers

In the original data, a few papers contribute many observations while most papers contribute few observations, or even a single one. In this section we explore to what extent the findings in BCFL's Table 2 are driven by only a few papers.

Individual papers can have a large influence on BCFL's tests for at least four reasons. First, removing papers that contributes many observations might create "holes" in the distribution of *p*-values that affect density fits. Second, articles with multiple estimates get disproportionally large weights in BCFL's Table 2 since none of their tests here weight observations by the inverse of the number of estimates in each article. Third, since binomial and discontinuity tests are local to significance thresholds, removing papers that disproportionately contribute estimates around these thresholds could meaningfully altern their conclusions. And fourth, all of BCFL's tests rely on only a few observations. One can see this, for example, in the sample sizes in Table A1 and using for benchmark EKW's exercises with random samples of *p*-values (which are always larger than BCFL's samples, sometimes by an order of magnitude), which always fail to reject the null hypothesis. EKW interpret this as "suggesting that the tests may be

underpowered in small samples" (pg. 897). Reliance on few observations can make the tests sensitive to tiny changes in the sample.[5]

We start by exploring the sensitivity of findings in BCFL's Table 2 to removing a random set of one percent of papers. We do this 1,000 times for each subsample and conduct all tests in Table 2 with the *p*-values from the remaining 99 percent of papers. In Tables 6 and 7 we show the 10^{th}-90^{th} percentile range across these 1,000 random paper removal exercises for all EKW tests, where numbers in bold mark instances where the authors' conclusions would have been different in at least 10% of cases. For instance, in Table 6 the BCLF *p*-value for heaping at the 10% significance threshold (column 4) for papers that failed to publish (last row) is 0.046 and our remove-one-random-paper 10^{th}-90^{th} percentile range is [0.042,0.059]. Because the upper end of the interval indicates significance at the 10% but not the 5% level as the BCFL original estimate does, this cell is in bold. Yet this cell and the one right below it are the only ones in bold. We conclude that all of BCFL's original estimates are robust to removing random small amounts of data from their sample.

[TABLE 6 HERE]

[TABLE 7 HERE]

We then explore the robustness of findings to excluding the paper(s) that contribute the 1% most estimates to each subsample. For example, from the initial submission subsample we exclude six out of 529 papers. These six papers have, in order of most to least main estimates: 217, 196, 196, 180, 162 and 140. In comparison, the average estimates per paper in this subsample is 24 and the median is 14. Table 6 shows the results of this exercise. We see several instances where BCFL's conclusions would have changed based on these estimates. Often, the changes only warrant slight reassessments of the evidence base (e.g., the binomial test *p*-value around the 5% threshold in desk rejections goes from 0.005 to 0.094) and in some instances the

[5] In an ideal scenario, we would have assessed the tests' sensitivity to removing highly leveraged observations as suggested by Broderick, Giordano and Meager (2023). However, leveraged statistics used for these exercises are usually regression-based. Such statistics are not appropriate in density tests (where observations around the threshold are naturally overweighted) or in the EKW tests (where the behavior of p-values across the (0,1) interval is assessed as a whole).

evidence for heaping becomes stronger. However, in general the evidence for heaping becomes weaker, and in many instances significantly so (e.g., the LCM test *p*-value for papers that passed the desk rejection goes from 0.038 to 0.121).

[TABLE 8 HERE]

3.2 Density tests under alternative modelling choices

It could also be that the findings of BCFL's density discontinuity tests are driven by the exact modelling choices made by the authors, such as the bandwidth or the polynomial fit of the running variable. We have already presented some evidence that the inclusion or exclusion of only a few observations near thresholds can produce moderate and sometimes consequential changes to these tests (see Section 2.3). Below we explore robustness to other modelling choices central to density discontinuity tests and regressions discontinuity designs. We conduct our exercises using the *R* command *rddensity* and varying the choice of polynomial fit, restricted or unrestricted polynomial fit parameters across the threshold, and the algorithm for automatic bandwidth selection at each side of the threshold. We focus on these modelling choices due to their key role in density discontinuity tests and our own previous experience with the pivotal role of bandwidth choice in these designs. We do not have an *a priori* reason to prefer one modelling specification over another here.

Table 9 presents results from density discontinuity tests under different assumptions, with Columns (1) and (6) reprinting the original BCFL results. We bold the cells for which alternative modelling choices would have altered conclusions. In virtually all subsamples (rows) there is at least one bolded cell, suggesting that arbitrary changes in modelling choices could have had a meaningful impact. For example, when looking at manuscripts that were rejected by reviewers (row 4), we find that three out of four tests under alternative modelling assumptions would have found evidence of heaping at the 10% significance level even though BCFL find a *p*-value of 0.380. Conversely, for manuscripts where reviewers recommended against outright rejection (row 5), three out

of four alternative modelling choices find no evidence of heaping at the 5% significance level even though BCFL find a *p*-value of 0.028.

The choice of order for the polynomial is the one that most commonly alters BCFL's conclusions. It is therefore important to reassess whether there is good *a priori* ground for this choice. Our reading of the literature is that there are no grounds for making this choice as long as they are low-order polynomials, and that results should be robust across alternatives (see e.g., Cattaneo, Idrobo & Titiunik, 2019). The fact that BCFL's results are sensitive to this choice is surprising, and we believe it might have two root causes. First, the small number of observations local to each threshold could result few bins with large variation in densities, which is hard to fit even with flexible methods. Second, the presence of heaps nearby (e.g., around the 0.05 threshold when testing for density discontinuities at the 0.10 threshold) could throw off even lower-order local polynomial fits.

[TABLE 9 HERE]

3.3 $|z|$-statistics rather than *p*-values as the running variable

Given the potential sensitivity of the heaping tests to the way local polynomials are fitted, it is reasonable to ask whether the choice of running variable for these tests could also matter. In particular, one could have also estimated heaping tests using the absolute value of *z*-statistics, which can be inferred as $|z| = \Phi^{-1}(1 - \frac{p}{2})$ where $\Phi^{-1}(\cdot)$ is the inverse normal function and p is the *p*-value. While $|z|$ is a rank-inverted bijection of p and therefore results from both tests should be equivalent, modelling the distribution of *p*-values around the 5% and 10% significance thresholds could be complicated by 1) the large mass of very small values nearby both thresholds and 2) that the distribution is compressed in $(0,1)$. Modelling the distribution of $|z|$-statistics alleviates both issues. See Figure 1 below for an illustration of this using the BCFL's distribution of *p*-values and our derived $|z|$-statistics in initial submissions.

[FIGURE 1 HERE]

Tables 10 and 11 show the results from estimating BCFL's discontinuity tests with $|z|$-statistics as a running variable, including the authors' original results for reference. For both set of results we also report the density discontinuity point estimates and effective number of observations, which allows for a better comparison and interpretation of results. We mark in bold the cells for which the conclusions differ between both analyses, in that the combination of *p*-value and point estimate imply different heaping patterns. Note that since $|z|$-statistics are a rank-inverted bijection of *p*-values, they are inconsistent if the point estimates in both tests *have the same sign*.[6] Our exercise shows that changing the running variable often alters the test's conclusions. For example, in Table 11 for papers with negative reviewer recommendations the test using *p*-values as the running variable indicates an excess mass of estimates right below the significance threshold (point estimate = 3.339, *p*-value = 0.001) whereas the test using $|z|$-statistics as the running variable indicates an excess mass right above the significance threshold (point estimate = 2.403, *p*-value = 0.016). Conclusions are altered in seven out of eleven cases when testing both discontinuities at the 5% and 10% significance threshold, although for the latter threshold we further observe sign inconsistencies across tests in three of these cases. The differences between the effective number of observations further suggests that using $|z|$-statistics as the running variable allows for simpler and better local polynomial fits, which in turn allows for broader bandwidths, a larger number of effective observations, and ultimately more powerful tests.

[TABLE 10 HERE]

[TABLE 11 HERE]

3.4 Caliper tests: varying bandwidths and alternative approaches

[6] To see this, take the example of density discontinuity tests around the 5% significance thresholds. A negative point estimate using *p*-values as the running variable implies "too many" estimates with *p*-values right below 0.05. The same would be implied by a positive point estimate using $|z|$-statistics as the running variable (i.e., "too many" estimates with $|z|$-statistics right above 1.96).

Our successful near-enough computational reproduction of BCFL's Caliper tests without covariates means we can also further probe the robustness of these results; even though it was impossible for us to exactly reproduce the original estimates.

We first show the robustness of all the Caliper tests to narrowing or broadening the bandwidth for inclusion of estimates in the estimation sample. Tables 12 and 13 show these results for the 10% and 5% significance thresholds. The first two columns show the effect to narrowing the bandwidth to 0.1 and 0.2 around the normal distribution critical values (so we use the range of observations with a $|z|$-statistic in the range of e.g., [1.55,1.75] around the 10% critical value of 1.65), Column (3) uses BCFL's preferred 0.3 bandwidth, and the remaining columns widen the bandwidth to up to 0.6 (so using a range as wide as e.g., [1.05,2.25] around 1.65). Overall, the main conclusions are very robust to variation in bandwidth, with the potential exception of the narrowest bandwidth of 0.1, though we believe the large changes in size and significance in these tests are due to the small number of observations included (e.g., 685, compared to 2,056 in the author's preferred bandwidth of 0.3).

[TABLE 12 HERE]

[TABLE 13 HERE]

We then test the robustness of these Caliper tests to an alternative approach to estimate the same relationship between statistical significance of estimates and paper outcomes using regression discontinuities. Our alternative approach estimates

$$y_p = \theta 1\left[\text{p-value}_{pe} > \tau\right] + f\left(\text{p-value}_{pe}\right) + \gamma' X_{pe} + \varepsilon_{pe} ,$$

where y_p is an outcome for paper p (e.g., desk rejection, reviewer recommends reject), p-value_{pe} is the *p*-value of estimate e in paper p, $1\left[\text{p-value}_{pe} > \tau\right]$ is a dummy variable that takes the value of one if p-value_{pe} is above the threshold τ, $f\left(\text{p-value}_{pe}\right)$ is a flexible polynomial function of p-value_{pe} fitted separately above and below τ, X_{pe} is a set of additional controls, and ε_{pe} is model error which we cluster at the paper level. The key estimate $\hat{\theta}$ captures the discontinuous change in outcome y_p for a paper with an estimate

right above the significance threshold τ, which can be estimated using standard regression discontinuity methods.

Note that $\hat{\theta}$ estimates, at its core, the same relationship as the coefficients in Caliper tests—that is, the relationship between statistical significance of estimates and paper outcomes local to a small bandwidth around critical values. Our approach, however, offers two additional advantages. First, it removes researcher degrees of freedom by allowing for a data-driven approach to determining the relevant bandwidth around significance thresholds. Second, it flexibly accounts for small differences in *p*-values around significance thresholds. This brings the analysis closer to the ideal experiment of comparing two identical papers that differ only in that one has an estimate that just reached statistically significant while the other has an estimate that just missed statistical significance.[7,8]

Table 14 shows the regression discontinuity estimates from these tests, which broadly show the same pattern as the Caliper tests: there is little evidence that statistical significance is related to paper outcomes. The one exception in both cases are estimates that just miss the 10% significance threshold, which are more likely to be desk rejected. However, while BCFL's Caliper test estimate provides strong evidence of this relationship (*p*-value = 0.004), evidence from our regression discontinuity estimate is much weaker (*p*-value = 0.080). All remaining Caliper and regression discontinuity estimates show no evidence of a relationship, with estimate magnitudes at least an order of magnitude smaller.

[TABLE 14 HERE]

[7] One unfortunate issue is that our approach reverses the role of outcome and key dependent variable in the regression compared to a Caliper test. This means we cannot directly compare coefficient estimates across tests. Since both tests pursue the same core hypothesis, however, we can compare the strength of the evidence they provide against the null of no relationship between statistical significance and paper outcomes in terms of their *t*-statistics and *p*-values.

[8] As BCFL, $\hat{\theta}$ should still not be interpreted as the causal effect of statistical significance since, in the presence of p-hacking and publication bias, papers with estimates right above the threshold might still differ to those right below in ways unobserved to the researcher (e.g., flimsier robustness analyses, writing style, author's position and/or training).

Finally, Table 15 shows the robustness of our regression discontinuity approach to using $|z|$-statistics rather than *p*-values as the running variable. As we argue in Section 3.3, it might be easier to properly model $|z|$-statistics with local polynomials. Other than that, both approaches are conceptually identical and should recover the same core parameter. Yet Table 15 shows no evidence that statistical significance at the 10% significance threshold is related to desk rejections (*p*-value = 0.843). This casts further doubts on the authors' finding that the distribution of desk rejections displays greater bunching than those sent for review.

[TABLE 15 HERE]

Overall, we find that BCFL's results are generally robust. However we also find that the discontinuity tests are sensitive to specific modelling choices such as the polynomial choice and the format of the running variable.

4 Extension: Statistical tests of Brodeur et al.'s claims

Beyond reproducibility exercises, we also extend the original analyses and interpretations from BCFL in one important way. We reconstruct the joint variance-covariance matrix of density discontinuity tests across subsamples to be able to explicitly test specific hypotheses they discuss in the paper (e.g., whether there are differences in heaping across subsamples).

BCFL's core analysis is a visual inspection of the *p*-value histograms, which the authors use to describe differences in their heaping tests for the distribution of *p*-values at the 5 and 10% significance thresholds at different stages of the publication process. One shortcoming of this approach, however, is that it limits the authors' ability to conduct inference on differences in heaping across distributions[9]. While their graphical evidence is compelling, visual inspection cannot formally take into consideration the magnitude of

[9] For example, in comparing desk rejected and not desk rejected submissions, they write that "[w]hile both distributions still display heaping at significance thresholds, the peak for desk rejections is much more pronounced" (pg. 2987). The authors then go on to analyze how desk rejection correlates with marginal statistical significance using Caliper tests.

the heaping across distributions and the uncertainty in these magnitude estimates, and does not allow one to conclude that the differences in heaping are economically and statistically meaningful. We complement BCFL's graphical analyses by using their provided data to construct such statistical tests.

We start by constructing a joint variance-covariance matrix of the heaping estimates in each subsample by 1) constructing a dataset of all appended subsamples, 2) estimating all the heaping estimates using this appended dataset, and 3) bootstrapping this procedure 1,000 times resampling entire papers with replacement. In Table 16 we present the results of this exercise. Columns (1) and (5) show the size and sign of each heaping estimate at the 5 and 10% significance thresholds. Each estimate should be interpreted as the additional mass in the density function of *p*-values found right above each significance threshold (e.g., the additional mass of estimates right above the 5% threshold compared to right below). While absolute values of these estimates are difficult to interpret due to normalization, the magnitudes are comparable across tests so that one can e.g., assess whether differences in heaping are most severe at the desk rejection stage of the review process. Columns (2) and (6) show the corresponding bootstrap-based standard errors of these estimates, which differ slightly from their analytical counterparts. Columns (3)-(4) and (7)-(8) show the 95% percentile-based bias-corrected confidence intervals based on bootstrapped estimates (and are therefore not necessarily symmetric around the estimate). All these quantities can offer finite-sample improvements to their analytical counterparts.

[TABLE 16 HERE]

Using Table 16 we can enrich some of BCFL's analyses by visual inspection. For example, we see no evidence of heaping around the 5% thresholds for desk rejected papers, but we do see strong evidence of heaping right below the 5% threshold for not desk rejected papers. Our finding using formal statistical testing complements BCFL's visual inspection: the authors state that "[…] the distribution of desk rejections displays greater bunching than those sent for review" (pg. 2976), and we find this difference is statistically significant specifically at the 5% threshold (see Test #2 below). We also see that papers recommended for rejection by reviewers have an excess mass right above the 10% threshold, whereas papers recommended against outright rejection have an excess mass right below the 5% threshold. These differences are also clear to

see when comparing all rejections to accepted manuscripts. Our results add nuance to BCFL's conclusion that "[r]ecommendations from anonymous reviewers […] are positively associated with statistical significance" (pg. 2976) and are consistent with a pattern in which estimates that just miss the 90% significance threshold are disproportionately associated with rejection recommendations, and estimates that just make the 95% significance threshold are disproportionately associated with non-rejection recommendations.

More importantly, having constructed the joint variance-covariance matrix of all the heaping tests, we can formally test BCFL's main hypotheses and test some additional hypotheses of interest. To do so, we test joint significance or joint equality of sets of coefficients in Table 16.

Table 17 describes our tests and presents our results. Test #1 pairs with BCFL's descriptive finding that "the distribution of test statistics among submitted articles displays a hump around 10 and 5% significance thresholds" (pg. 2976). The result of our test does not support this claim, since we cannot reject the null hypothesis that heaping at both thresholds is significantly different from zero (*p*-value = 0.438).

Test #2 pairs with BCFL's descriptive finding that "the distribution of desk rejections displays greater bunching than those sent for review" (pg. 2976). We only find weak evidence to support the claim as is (*p*-value = 0.123). However, in further analyses we do see evidence of differences in heaping at the 5% threshold between both distributions (*p*-value = 0.044), while finding no evidence of differences at the 10% threshold (*p*-value = 0.588). Therefore, our results support the authors' claim.

Test #3 pairs with BCFL's finding that favorable reviewer recommendations are positively related to statistical significance (pg. 2976, paraphrased) by first testing for systematic differences in heaping across the distributions of rejections, recommendation of no outright rejections, and acceptances as is or with minor edits. We find no evidence to support the claim, since we cannot reject the null of no differences in heaping across all three distributions (*p*-value = 0.233). In further analysis we find no evidence of differences across these distributions at the 5% significance threshold (*p*-value = 0.706) and only weak evidence at the 10% threshold (*p*-value = 0.109).

Test #4 pairs with BCFL's descriptive finding that "the p-value distribution of initial submissions and final drafts are similar" (pg. 2976), at least in terms of heaping, and

cannot reject the null of no differences in heaping (*p*-value = 0.699), which supports the claim.[10]

Finally, Test #5 is a formal omnibus tests of heaping across all stages of the publication process that rejects the null of no heaping at the 5 and 10% significance thresholds (*p*-value = 0.003).

[TABLE 17 HERE]

In sum, our results using formal statistical testing probe the authors' descriptive findings based visual inspection method. Our results align with the authors' two main claims, however we find no support that there is heaping in initial submissions, and scant evidence that heaping differs between desk rejections and not desk rejections, or by referee recommendations.

5 Conclusion

Brodeur et al. (2023) are the first to investigate *p*-hacking and publication bias across all stages in the publication process using unique data from the *Journal of Human Resources*. This is a seminal paper in the literature on research transparency, because no previous study has been able to discuss at what stage of the publication process does publication bias occur, which limits the scope for designing policies to tackle publication bias.

The authors make five key claims (pg. 2976):

1. the distribution of test statistics among submitted articles displays a hump around 10 and 5% significance thresholds,
2. the distribution of desk rejections displays greater bunching than those sent for review,
3. this last result is partially explained by author characteristics,

[10] We also estimate a Kolmogorov-Smirnov test of equality between both distributions. This is a better test of the literal claim that "distributions are similar". And even though this test is not powerful in detecting differences in the tails of distributions, it is very powerful in detecting differences due to lumpiness and heaping, which is more relevant here. The Kolmogorov-Smirnov test provides weak evidence that the distributions are not equal (p-value = 0.072), with the directional statistics providing strong evidence that the initial submissions contain smaller *p*-values than the final ones (*p*-value = 0.036) and no evidence that initial submissions contain larger values than the final one (*p*-value = 0.837). We note, however, that these conclusions are based on approximate *p*-values; exact *p*-values for the test could not be computed because there are only 4,970 unique values out of 9,077 observations.

4. favorable reviewer recommendations are positively related to statistical significance, and
5. the distribution of statistics from the final draft of accepted manuscripts is similar to its initial draft.

In this paper we reproduce and probe the robustness of the descriptive evidence supporting their claims. Given the confidential nature of their data and using the limited data that was posted in their replication package, we focused on their key tables. Importantly, we could not reproduce or assess any of the evidence supporting Claim #3. From the evidence we could reproduce, however, we draw several conclusions.

We find that most of BCFL's main results are nearly perfectly computationally reproducible, and also very robust. Our formal statistical tests lend further support to some of the authors' main findings based on descriptive graphical evidence. However, we did uncover three issues. First, the few estimates that were not computationally reproducible could have meaningfully altered the authors' conclusions. Second, their density discontinuity test results were often not robust to reasonable alternative modeling choices. And third, formal statistical tests failed to support some of BCFL's descriptive findings based on graphical evidence.

With our findings we hope to open a discussion following BCFL's work. We view our findings as an invitation to rely the least possible on graphical evidence and density discontinuity tests for drawing conclusions. Graphical evidence will always have a degree of subjectivity and many researcher degrees of freedom. Our exercises have uncovered several ways in which density discontinuity tests are highly sensitive to small and seemingly innocuous researcher decisions. Researcher degree of freedom is emerging as an important hidden source of variation in research findings (see Huntington-Klein et al., 2021, 2025). All these problems are likely compounded by the small number of observations in BCFL's data; across all their density discontinuity tests, their maximum number of effective observations is 383 observations, and some tests run with as few as 99 effective observations.

The good news is that the battery of tests from EKW contains arguably better tests for *p*-hacking and publication bias. The bad news is that these tests often provide conflicting results to those from the density discontinuity tests. Facing conflicting

evidence, which tests should one rely on? On a priori grounds, we would place most weight on EKW's Least Concave Majorant (LCM) test since it can detect p-hacking and publication bias even when p-hacking does not induce an increasing p-curve (which could occur if e.g., everyone p-hacks all their estimates so the p-curve displays no heaps) (see EKW, pg. 893). Our second-preferred tests would be EKW's histogram-based test for 2-monotonicity and bounds on the p-curve and the first two derivatives (CS2B) since it exploits additional restrictions on the shape of a p-curve based on two-sided t-tests (which directly applies to BCFL) to increase power (see EKW, pg. 839). After that, we would weight in the evidence from density discontinuity tests. Finally, we would consider evidence form EKW's histogram-based test using the Cox and Shi (2020) chi-squared test.[11]

Reassessing BCFL's evidence using our preferred tests (and as reproduced in our Table 2), we would have reached the following conclusions regarding evidence of *p*-hacking and publication bias:

1. there is strong evidence of *p*-hacking and publication bias in the distribution of initial submissions,
2. there is no evidence of *p*-hacking and publication bias for desk rejected papers, but some evidence that it remains in non-desk rejected papers,
3. there is strong evidence of *p*-hacking and publication bias in papers across all reviewer recommendations,
4. there is little evidence that *p*-hacking and publication bias differs between initial and accepted versions of papers, between accepted and rejected papers, or between papers published elsewhere and papers that failed to publish.

We view our results as painting a more dismal view of the publication process in the Journal of Human Resources (and, more likely, in Economics). They suggest that p-hacking and publication bias persist throughout the publication process. We do not interpret this as evidence on the causal role of peer review, but our findings are consistent with BCFL's descriptive observation that the distributions of test statistics in initial

[11] For this application, we would disregard the results from the Binomial test, since it almost always rejects the null with a p-value smaller than 0.001. We also disregard Fisher's tests since BCFL also do so on the grounds that it almost always produced a p-value equal to 1 (as also shown in EKW's application). However, these exclusions are not on *a priori* grounds.

submissions and final accepted manuscripts are similar. We do not believe this to be the final word on the matter. Rather, BCFL's efforts and our own highlight the importance of a critical and constructive discussion of the publication process in economics and underline the importance of data availability and data transparency to enable it.

References

Broderick, T., Giordano, R. and Meager, R., 2020. An automatic finite-sample robustness metric: when can dropping a little data make a big difference?. *arXiv preprint arXiv:2011.14999*.

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

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

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

Bruns, S.B., Asanov, I., Bode, R., Dunger, M., Funk, C., Hassan, S.M., Hauschildt, J., Heinisch, D., Kempa, K., König, J. and Lips, J., 2019. Reporting errors and biases in published empirical findings: Evidence from innovation research. *Research Policy*, *48*(9), p.103796.

Cattaneo, M.D., Idrobo, N. and Titiunik, R., 2019. *A practical introduction to regression discontinuity designs: Foundations*. Cambridge University Press.

Cattaneo, M.D., Jansson, M. and Ma, X., 2018. Manipulation testing based on density discontinuity. *The Stata Journal*, *18*(1), pp.234-261.

Cattaneo, M.D., Jansson, M. and Ma, X., 2020. Simple local polynomial density estimators. *Journal of the American Statistical Association*, *115*(531), pp.1449-1455.

Cox, G. and Shi, X., 2023. Simple adaptive size-exact testing for full-vector and subvector inference in moment inequality models. *The Review of Economic Studies*, *90*(1), pp.201-228.

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

Gerber, A. and Malhotra, N., 2008. Do statistical reporting standards affect what is published? Publication bias in two leading political science journals. *Quarterly Journal of Political Science*, *3*(3), pp.313-326.

Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J.R., Burli, P., Chen, N., Grieco, P., Ekpe, G., Pugatch, T. and Saavedra, M., 2021. The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, *59*(3), pp.944-960.

Huntington-Klein, N., Portner, C.C. and McCarthy, I., 2025. *The sources of researcher variation in economics* (No. w33729). National Bureau of Economic Research.

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

Tables and Figures

Table 1 – Summary of the Computational Reproducibility of BCFL

	Fully	Partial	No
Raw data provided			X
Cleaning code provided			X
Analysis data provided		X	
Analysis code provided		X	
Reproducible from raw data			X
Reproducible from analysis data		X	

Table 2 – Computational reproduction of BCFL (Tables 2 and A1)

Threshold:	5% significance			10% significance					
Test:	Bin. (1)	Discont. (2)	Obs. (2')	Bin. (3)	Discont. (4)	Obs. (4')	CS1 (5)	CS2B (6)	LCM (7)
Initial submissions	<0.001	0.743	372	**0.067**	**0.277**	215	<0.001	<0.001	0.004
Desk reject stage									
Desk rejections	0.005	0.599	117	0.014	0.337	67	0.524	**0.270**	0.338
Not desk rejections	<0.001	0.003	255	0.403	0.141	148	<0.001	<0.001	0.038
Reviewer stage									
Recommend rejection	<0.001	0.380	196	0.906	0.001	113	0.005	0.003	0.077
Recommend against outright rejection	<0.001	0.028	238	0.813	0.472	127	0.002	<0.001	0.248
Recommend accept as is or with minor edits	0.001	0.162	103	0.028	0.848	71	<0.001	<0.001	0.703
Accepted manuscripts									
Initial submission	<0.001♠	<0.001♠	149♠	0.012	0.531	79	0.464	0.001	0.404
Published version	<0.001	0.031	139	<0.001	0.970	66	<0.001	<0.001	0.348
Peer review									
All rejections	<0.001	0.485	223	0.466	0.005	136	0.009	0.003	0.029
Accepted manuscripts	♣	♣	♣	♣	♣	♣	♦	♦	♦
After rejection									
Published elsewhere	<0.001	0.878	126	0.376	0.052	90	0.023	0.001	0.374
Failed to publish	0.001	0.678	97	0.671	0.046	46	0.145	0.004	0.210

This table presents the computational reproducibility of Brodeur et al. (2023), Table 2 and A1. Columns 1 through 7 reproduce estimates from Table 2. Columns (2') and (4') add the number of observations from Table A1. These estimates were reproduced using their posted analysis data and code in R. For Columns (1), (2) and (2') the estimates printed are produced in figures in the code provided by Brodeur et al. For cells marked with ♠, the figure produced did not match the estimates reported on the table, but our own reconstruction of the code reproduced the numbers accurately. The data and code for replicating the row 'Accepted manuscripts' was missing from the package, but it seems in BCFL's Table 2 cells marked with ♣ were just copies of those in the row 'Published version' above, which would make sense. However, cells marked with ♦ differed between both rows in BCFL's Table 2. Cells marked in **bold** *indicate subsamples and tests for which our reproduction differs from the original study.*

Table 3 – Computational reproduction of BCFL (Tables A3 and A4)

Threshold:	5% significance			10% significance					
Test:	Bin. (1)	Discont. (2)	Obs. (2')	Bin. (3)	Discont. (4)	Obs. (4')	CS1 (5)	CS2B (6)	LCM (7)
Initial submissions	0.980	0.514	305	0.940	0.687	201	0.207	0.025	1.000
Desk reject stage									
Desk rejections	0.896	0.532	91	0.295	0.912	55	0.657	0.395	1.000
Not desk rejections	0.956	0.831	214	0.987	0.646	146	0.054	0.002	1.000
Reviewer stage									
Recommend rejection	0.681♠	0.397♠	163♠	0.930	0.456	104	0.444	0.077	1.000
Recommend against outright rejection	0.786	0.694	192	0.975	0.849	116	0.738	0.505	1.000
Recommend accept as is or with minor edits	0.580	**0.666**	97	0.957	0.525	67	0.132	0.001	0.999
Accepted manuscripts									
Initial submission	0.954	0.176	128	0.909	0.902	68	0.197	0.110	1.000
Published version	0.363	0.313	131	0.070	0.307	56	0.638	0.037	1.000
Peer review									
All rejections	0.912	0.076	177	0.851	0.387	133	0.122	0.024	1.000
Accepted manuscripts	♣	♣	♣	♣	♣	♣	♣	♣	♣
After rejection									
Published elsewhere	0.386	0.165	108	0.625	0.767	88	0.390	0.053	1.000
Failed to publish	0.996	0.673	69	0.932	0.413	45	0.599	0.741	1.000

This table presents the computational reproducibility of Brodeur et al. (2023), Table A2 and A3. Columns 1 through 7 reproduce estimates from Table A2. Columns (2') and (4') add the number of observations from Table A3. These estimates were reproduced using their posted analysis data and code in R. For Columns (1), (2) and (2') the estimates printed are produced in figures in the code provided by Brodeur et al. For cells marked with ♠, no figure was produced in their code but our own reconstruction of the code reproduced the numbers accurately. The data and code for replicating the row 'Accepted manuscripts' was missing from the package, but it seems in BCFL's Table 2 cells marked with ♣ were just copies of those in the row 'Published version' above, which would make sense. Cells marked in **bold** *indicate subsamples and tests for which our reproduction differs from the original study.*

Table 4 – Near-enough computational reproduction of BCFL (Tables 4 through 8, Column 1)

Threshold:	Caliper tests with $\|z\| \in [1.35,1.95]$	
	10% significance	5% significance
	(1)	(2)
Desk rejected (vs not desk rejected)	0.121***	0.002
	(0.042)	(0.044)
Obs.	2,056	2,066
Weak R&R (vs Reject)	0.043	0.065*
	(0.035)	(0.036)
Strong R&R (vs Reject)	0.052	0.090
	(0.050)	(0.057)
Obs.	3,240	3,224
Accepted version (vs initial version)	0.014	0.032
	(0.044)	(0.049)
Obs.	1,543	1,593
Accepted (vs rejected)	-0.050	0.044
	(0.040)	(0.044)
Obs.	2,014	1,992
Published (vs never published)	-0.031	0.030
	(0.050)	(0.050)
Obs.	1,269	1,236

This table presents the computational reproducibility of Brodeur et al. (2023), Column 1 of Tables 4 through 8. These estimates reproduce Caliper tests, as described in the original paper. These estimates were reproduced using their posted analysis data and code in Stata. We cannot reproduce their numbers exactly since the provided datasets lack confidential covariates, and the final samples they use exclude observations with missing covariates, which we cannot identify. However, there seems to be few of these observations in all samples. Cells marked in **bold** *indicate subsamples and tests for which our reproduction meaningfully differs from the original study.*

Table 5 – Computational reproduction of BCFL (Table 2, Columns 2 and 4) using *Stata* instead of *R*

	Density discontinuity test *p*-values			
Threshold:	5% significance		10% significance	
Software:	R	Stata	R	Stata
	(1)	(2)	(3)	(4)
Initial submissions	0.743	0.138	0.004	0.184
Desk reject stage				
Desk rejections	0.599	0.369	0.337	0.411
Not desk rejections	0.003	<0.001	0.141	**0.080**
Reviewer stage				
Recommend rejection	0.380	0.226	0.001	0.001
Recommend against outright rejection	0.028	0.005	0.472	0.872
Recommend accept as is or with minor edits	0.162	0.348	0.848	0.836
Accepted manuscripts				
Initial submission	<0.001	<0.001	0.531	0.551
Published version	0.031	0.035	0.970	0.899
Peer review				
All rejections	0.485	0.279	0.005	0.004
Accepted manuscripts	0.031	-	0.970	-
After rejection				
Published elsewhere	0.878	0.554	0.052	0.054
Failed to publish	0.678	0.721	**0.046**	**0.055**

Columns (1) and (3) of this table reproduce the p-values from discontinuity tests in Brodeur et al. (2023), Table 2, Columns 2 and 4. These p-values come from local polynomial density tests on p-curves at the 5 and 10% significance level, estimated in R using the package rddensity with all the default options. Columns (2) and (4) show our own computational reproducibility of these same p-values using their analysis files and the Stata package rddensity with the same default options. Cells marked in **bold** *indicate subsamples and tests for which our reproduction meaningfully differs from the original study.*

Table 6 – Sensitivity of BCFL (Table 2, Columns 1 through 4): 10^{th} – 90^{th} percentile range across 1,000 re-estimations randomly removing 1% of papers

	10^{th}-90^{th} percentiles in the p-value distribution of density discontinuity tests across 1,000 random paper removals			
Threshold:	5% significance		10% significance	
Test:	Binomial	Discontinuity	Binomial	Discontinuity
	(1)	(2)	(3)	(4)
Initial submissions	[<0.001,<0.001]	[0.535,0.821]	[0.055,0.085]	[0.211,0.332]
Desk reject stage				
Desk rejections	[0.003,0.006]	[0.528,0.639]	[0.012,0.018]	[0.287,0.388]
Not desk rejections	[<0.001,<0.001]	[0.001,0.004]	[0.369,0.435]	[0.101,0.170]
Reviewer stage				
Recommend rejection	[<0.001,<0.001]	[0.294,0.488]	[0.873,0.922]	[0.001,0.002]
Recommend against outright rejection	[0.001,0.003]	[0.118,0.337]	[0.021,0.036]	[0.715,0.906]
Recommend accept as is or with minor edits	[<0.001,<0.001]	[0.015,0.037]	[0.763,0.858]	[0.423,0.570]
Accepted manuscripts				
Initial submission	[<0.001,<0.001]	[0.026,0.042]	[<0.001,<0.001]	[0.867,0.981]
Published version	[<0.001,0.001]	[<0.001,<0.001]	[0.008,0.015]	[0.483,0.605]
Peer review				
All rejections	[<0.001,<0.001]	[0.361,0.577]	[0.431,0.500]	[0.003,0.007]
Accepted manuscripts				
After rejection				
Published elsewhere	[<0.001,0.001]	[0.646,0.745]	[0.617,0.671]	[0.042,0.059]
Failed to publish	[<0.001,<0.001]	[0.744,0.939]	[0.336,0.416]	**[0.039,0.071]**

This table presents the sensitivity of findings to random removals of 1% of papers from each subsample. For each subsample, we conduct 1,000 random draws removing 1% of papers each draw, and then re-estimate the density discontinuity tests in Brodeur et al. (2023), Table 2. Finally, we compute the 10th and 90th percentiles in the p-value distribution across all 1,000 draws and report it here. Cells marked in **bold** *indicate subsamples and tests for which one of the ends of the 10th-90th percentile range falls outside the significance level of the estimate in the original study.*

Table 7 – Sensitivity of BCFL (Table 2, Columns 5 through 7): 10th – 90th percentile range across 1,000 re-estimations randomly removing 1% of papers

Test:	10th-90th percentiles in the p-value distribution of EKW tests across 1,000 random paper removals		
	CS1 (1)	CS2B (2)	LCM (3)
Initial submissions	[<0.001,<0.001]	[<0.001,<0.001]	[0.004,0.005]
Desk reject stage			
Desk rejections	[0.349,0.568]	[0.225,0.336]	[0.329,0.359]
Not desk rejections	[<0.001,<0.001]	[<0.001,<0.001]	[0.036,0.047]
Reviewer stage			
Recommend rejection	[0.003,0.008]	[0.001,0.005]	[0.074,0.091]
Recommend against outright rejection	[<0.001,<0.001]	[<0.001,<0.001]	[0.690,0.742]
Recommend accept as is or with minor edits	[0.001,0.003]	[<0.001,<0.001]	[0.240,0.276]
Accepted manuscripts			
Initial submission	[<0.001,0.001]	[<0.001,<0.001]	[0.340,0.368]
Published version	[0.301,0.499]	[0.001,0.002]	[0.397,0.430]
Peer review			
All rejections	[0.007,0.016]	[0.002,0.005]	[0.028,0.035]
Accepted manuscripts			
After rejection			
Published elsewhere	[0.118,0.180]	[0.002,0.007]	[0.204,0.230]
Failed to publish	[0.012,0.036]	[<0.001,0.002]	[0.364,0.405]

This table presents the sensitivity of findings to random removals of 1% of papers from each subsample. For each subsample, we conduct 1,000 random draws removing 1% of papers each draw, and then re-estimate the EKW monotonicity and bounds tests in Brodeur et al. (2023), Table 2. Finally, we compute the 10th and 90th percentiles in the p-value distribution across all 1,000 draws and report it here. Cells marked in **bold** *indicate subsamples and tests for which one of the ends of the 10th-90th percentile range falls outside the significance level of the estimate in the original study.*

Table 8 – Sensitivity of BCFL (Table 2): Removing the 1% of papers contributing most estimates to each subsample

	EKW tests excluding the 1% of papers with most estimates						
Threshold:	5% significance		10% significance				
Test:	Binomial (1)	Discontinuity (2)	Binomial (3)	Discontinuity (4)	CS1 (5)	CS2B (6)	LCM (7)
Initial submissions	<0.001	0.700	**0.330**	**0.073**	0.005	0.002	0.047
Desk reject stage							
Desk rejections	**0.094**	0.489	0.034	0.474	0.470	0.217	0.894
Not desk rejections	<0.001	**0.019**	0.789	**0.037**	0.002	<0.001	**0.121**
Reviewer stage							
Recommend rejection	<0.001	0.773	0.918	0.001	**0.023**	0.009	**0.100**
Recommend against outright rejection	<0.001	**0.076**	0.964	0.209	0.007	<0.001	0.228
Recommend accept as is or with minor edits	**0.066**	0.293	**0.214**	0.885	0.008	<0.001	0.980
Accepted manuscripts							
Initial submission	0.022	0.005	**0.111**	0.801	0.731	**0.021**	0.887
Published version	<0.001	0.016	<0.001	0.822	<0.001	<0.001	0.489
Peer review							
All rejections	<0.001	0.912	0.536	0.008	**0.059**	0.010	**0.121**
Accepted manuscripts	-	-	-	-	-	-	-
After rejection							
Published elsewhere	0.003	0.489	0.544	0.096	0.041	0.001	0.921
Failed to publish	<0.001	0.669	0.620	**0.058**	0.221	**0.014**	0.210

This table presents the sensitivity of findings to removing 1 paper at a time from each subsample. For each subsample, we remove 1 paper and re-estimate all tests in Brodeur et al. 2023, Table 2. This table reports the largest p-values found for each subsample and test across all removals. Cells marked in bold indicate instances where the maximum p-value across removals would have overturned the conclusion in the original study.

Table 9 – Sensitivity of BCFL (Table 2, Columns 2 and 4): Density discontinuity tests under alternative modelling choices

	Density discontinuity test *p*-values									
Threshold:	5% significance					10% significance				
Modelling choice:										
Fit polynomial	Quad	Lin	Lin	Quad	Lin	Quad	Lin	Lin	Quad	Lin
Fit across thresholds	Unrest	Unrest	Unrest	Unrest	Rest	Unrest	Unrest	Unrest	Unrest	Rest
BW selector across threshold	Comb	Comb	Each	Each	Comb	Comb	Comb	Each	Each	Comb
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Initial submissions	0.743	0.625	0.159	0.401	**<0.001**	0.277	0.504	0.958	**0.086**	0.111
Desk reject stage										
Desk rejections	0.599	0.196	0.927	0.161	**<0.001**	0.337	0.932	0.394	0.571	0.166
Not desk rejections	0.003	**0.780**	**0.249**	<0.001	0.041	0.141	0.215	0.801	0.106	**0.095**
Reviewer stage										
Recommend rejection	0.380	**0.016**	**0.086**	**<0.001**	0.663	0.001	0.012	**0.656**	0.013	0.037
Recommend against outright rejection	0.028	**0.850**	**0.874**	0.045	**0.193**	0.472	**0.008**	0.127	0.997	**0.005**
Recommend accept as is or with minor edits	0.162	0.119	0.696	0.688	0.301	0.848	0.199	0.347	0.753	0.939
Accepted manuscripts										
Initial submission	<0.001	**0.963**	**0.744**	0.002	**0.293**	0.531	**0.020**	0.188	0.546	0.168
Published version	0.031	**0.588**	**0.563**	0.001	**0.214**	0.970	**0.098**	0.141	0.322	0.393
Peer review										
All rejections	0.485	0.629	**0.090**	**<0.001**	0.880	0.005	**0.191**	**0.257**	0.005	**0.999**
Accepted manuscripts	-	-	-	-	-	-	-	-	-	-
After rejection										
Published elsewhere	0.878	0.107	0.368	**0.041**	**0.080**	0.052	**0.950**	**0.960**	0.055	0.061
Failed to publish	0.678	**0.025**	**<0.001**	**0.001**	0.851	0.046	**0.114**	**0.190**	**0.103**	0.073

Columns (1) and (6) of this table reproduce the p-values from Brodeur et al. (2023), Table 2, Columns 2 and 4. The remaining columns reproduce these results under alternative model specifications, summarized in the top panel of the table. Fit polynomial refers to the degree of the polynomial for the local fit at the threshold, with options linear and quadratic configured. Fit across thresholds refers to whether the local polynomial fit parameters are restricted to be the same above and below the threshold. BW selector across thresholds refers to the bandwidth selection algorithm and we vary the option as whether bandwidths are based on MSE of each density estimator above and below the threshold separately (Each), or whether they are based on the median bandwidth as chosen by the MSE at each side separately, the MSE of difference of two density estimators jointly, or the MSE of the sum of the two density estimators jointly (Combined). All other options in rddensity are kept at the default values in R. Cells marked in **bold** *indicate instances where the p-value based on alternative modelling choices would have overturned the conclusion in the original study.*

Table 10 – Sensitivity of BCFL (Table 2, Columns 2 and 4): Density discontinuity tests using $|z|$-statistics as the running variable (5% significance threshold)

Running variable:	Density discontinuity tests at the 5% significance threshold					
	p-values			$\|z\|$-statistics		
	Estimate	p-value	Eff. Obs.	Estimate	p-value	Eff. Obs.
	(1)	(2)	(3)	(4)	(5)	(6)
Initial submissions	-0.328	0.743	383	1.887	**0.059**	3,126
Desk reject stage						
Desk rejections	0.525	0.599	140	3.093	**0.002**	967
Not desk rejections	-3.012	0.003	282	-0.425	**0.671**	2,473
Reviewer stage						
Recommend rejection	-0.878	0.380	232	-1.587	0.113	1,789
Recommend against outright rejection	-2.202	0.028	265	-0.149	**0.881**	1,909
Recommend accept as is or with minor edits	-1.398	0.162	144	1.528	0.126	1,766
Accepted manuscripts						
Initial submission	-3.689	<0.001	180	0.965	**0.335**	1,973
Published version	-2.158	0.031	184	0.650	**0.516**	1,733
Peer review						
All rejections	-0.698	0.485	244	2.103	**0.036**	1,772
Accepted manuscripts	-	-	-	-	-	-
After rejection						
Published elsewhere	0.153	0.878	164	1.448	0.148	1,076
Failed to publish	-0.416	0.678	114	-1.487	0.137	1,061

Columns (1) through (3) reproduce density discontinuity tests in Brodeur et al. (2023), Table 2, Column 2 and add the estimate and effective number of observations underlying these tests. Columns (4) through (6) replicate these tests using the corresponding |z|-statistics rather than the p-values as the running variable. The threshold used for the tests is 1.96. Note that for both tests to provide consistent evidence, the sign of their statistics need to be opposite since z-statistics are a rank-inverted bijection of p-values. All these are estimated in R using the package rddensity with all the default options. Cells marked in **bold** *indicate instances where the evidence from both version of this test is conflicting based on the p-value or the sign of the estimate.*

Table 11 – Sensitivity of BCFL (Table 2, Columns 2 and 4): Density discontinuity tests using $|z|$-statistics as the running variable (10% significance threshold)

Running variable:	Density discontinuity tests at the 10% significance threshold					
	p			|*z*|		
	Estimate	p-value	Eff. Obs.	Estimate	p-value	Eff. Obs.
	(1)	(2)	(3)	(4)	(5)	(6)
Initial submissions	1.088	0.277	316	1.437	0.151	4,345
Desk reject stage						
Desk rejections	0.961	0.337	129	2.353	**0.019**	1,077
Not desk rejections	1.474	0.141	263	0.375	0.708	4,036
Reviewer stage						
Recommend rejection	3.339	0.001	198	**2.403**	**0.016**	2,716
Recommend against outright rejection	-0.720	0.472	201	0.958	0.338	4,410
Recommend accept as is or with minor edits	-0.192	0.848	177	-0.787	0.432	1,522
Accepted manuscripts						
Initial submission	0.626	0.531	162	-1.776	**0.076**	1,790
Published version	-0.037	0.970	145	-1.841	**0.066**	1,858
Peer review						
All rejections	2.829	0.005	229	**3.861**	**<0.001**	2,031
Accepted manuscripts	-	-	-	-	-	-
After rejection						
Published elsewhere	1.947	0.052	157	**4.457**	**<0.001**	1,237
Failed to publish	1.996	0.046	99	0.272	**0.785**	1,428

Columns (1) through (3) reproduce density discontinuity tests in Brodeur et al. (2023), Table 2, Column 4 and add the estimate and effective number of observations underlying these tests. Columns (4) through (6) replicate these tests using the corresponding |z|-statistics rather than the p-values as the running variable. The threshold used for the tests is 1.65. Note that for both tests to provide consistent evidence, the sign of their statistics need to be opposite since z-statistics are a rank-inverted bijection of p-values. All these are estimated in R using the package rddensity with all the default options. Cells marked in **bold** *indicate instances where the evidence from both version of this test is conflicting based on the p-value or the sign of the estimate.*

Table 12 – Sensitivity of BCFL (Tables 4 through 8, Column 1): Caliper tests using different bandwidths (10% significance threshold)

	Caliper tests around the 10% significance threshold					
Bandwidth around 1.65:	0.1	0.2	0.3	0.4	0.5	0.6
	(1)	(2)	(3)	(4)	(5)	(6)
Desk rejected (vs not desk rejected)	0.037	0.130***	0.121***	0.132***	0.124***	0.110***
	(0.066)	(0.045)	(0.042)	(0.037)	(0.034)	(0.032)
Obs.	685	1,380	2,056	2,757	3,411	3,990
Weak R&R (vs Reject)	-0.053	0.023	0.043	0.040	0.036	0.032
	(0.054)	(0.045)	(0.035)	(0.030)	(0.028)	(0.027)
Strong R&R (vs Reject)	0.003	0.050	0.052	0.071	0.098**	0.081**
	(0.075)	(0.055)	(0.050)	(0.045)	(0.042)	(0.041)
Obs.	1,069	2,130	3,240	4,277	5,289	6,251
Accepted version (vs initial version)	-0.053	-0.003	0.014	0.000	0.010	0.019
	(0.066)	(0.049)	(0.044)	(0.037)	(0.036)	(0.035)
Obs.	511	1,041	1,543	2,098	2,660	3,132
Accepted (vs rejected)	-0.062	-0.053	-0.050	-0.048	-0.035	-0.044
	(0.059)	(0.044)	(0.040)	(0.034)	(0.033)	(0.031)
Obs.	648	1,347	2,014	2,721	3,343	3,922
Published (vs never published)	-0.074	-0.038	-0.031	-0.016	-0.019	-0.012
	(0.072)	(0.054)	(0.050)	(0.044)	(0.040)	(0.037)
Obs.	411	848	1,269	1,697	2,057	2,402

This table presents the robustness of Brodeur et al.'s Caliper tests to using different bandwidths around the 10% significance critical value of 1.65. Cells marked in **bold** *indicate subsamples and tests for which our reproduction meaningfully differs from the original study.*

Table 13 – Sensitivity of BCFL (Tables 4 through 8, Column 1): Caliper tests using different bandwidths (5% significance threshold)

	Caliper tests around the 5% significance threshold					
Bandwidth around 1.96:	0.1	0.2	0.3	0.4	0.5	0.6
	(1)	(2)	(3)	(4)	(5)	(6)
Desk rejected (vs not desk rejected)	0.062	-0.005	0.002	-0.017	-0.002	-0.001
	(0.072)	(0.051)	(0.044)	(0.040)	(0.038)	(0.035)
Obs.	711	1,372	2,066	2,662	3,270	3,870
Weak R&R (vs Reject)	-0.014	0.043	0.065*	0.044	0.045	0.052*
	(0.046)	(0.038)	(0.036)	(0.033)	(0.031)	(0.029)
Strong R&R (vs Reject)	0.079	0.151**	0.090	0.060	0.049	0.066
	(0.085)	(0.064)	(0.057)	(0.055)	(0.053)	(0.048)
Obs.	1,119	2,133	3,224	4,157	5,116	6,096
Accepted version (vs initial version)	0.009	0.040	0.032	0.009	0.009	0.018
	(0.066)	(0.056)	(0.049)	(0.045)	(0.040)	(0.037)
Obs.	561	1,093	1,593	2,080	2,544	2,982
Accepted (vs rejected)	-0.004	0.059	0.044	0.057	0.040	0.038
	(0.069)	(0.051)	(0.044)	(0.040)	(0.036)	(0.032)
Obs.	691	1,316	1,992	2,567	3,197	3,800
Published (vs never published)	0.078	0.019	0.030	0.025	0.048	0.041
	(0.079)	(0.059)	(0.050)	(0.045)	(0.043)	(0.039)
Obs.	421	800	1,236	1,578	1,968	2,352

This table presents the robustness of Brodeur et al.'s Caliper tests to using different bandwidths around the 5% significance critical value of 1.96. Cells marked in **bold** *indicate subsamples and tests for which our reproduction meaningfully differs from the original study.*

Table 14 – Sensitivity of BCFL (Tables 4 through 8, Column 1): Regression discontinuity caliper-like tests using *p*-values as the running variable

	Regression discontinuity tests using *p*-values as the running variable	
Threshold:	10% significance	5% significance
	(1)	(2)
Outcome:		
Desk rejected (vs not desk rejected)	**-0.168***	-0.024
	(0.096)	(0.081)
Effective obs.	2,241	3,665
Weak R&R (vs Reject)	0.027	0.010
	(0.068)	(0.054)
Effective obs.	4,314	5,622
Strong R&R (vs Reject)	-0.064	-0.047
	(0.058)	(0.054)
Effective obs.	4,780	7,163
Accepted version (vs initial version)	0.041	-0.050
	(0.094)	(0.084)
Effective obs.	1,974	3,676
Accepted (vs rejected)	0.041	-0.024
	(0.075)	(0.067)
Effective obs.	2,048	3,966
Published (vs never published)	0.060	-0.030
	(0.102)	(0.100)
Effective obs.	1,803	2,515

This table presents a robustness check of Brodeur et al.'s Caliper tests, instead using regression discontinuity tables. In these regressions the dependent variable is the paper status (i.e., accepted), the running variable are estimates' p-values, and treatment is coded as one if the running variable is over the 10% and 5% significance threshold in Columns (1) and (2). These estimates were produced using the rdrobust package in Stata, and report robust and bias-corrected estimates with CER-optimal bandwidths at both sides of the threshold, controlling for the statistics reported in each paper, correcting for mass points in the running variable, clustering standard errors at the paper level, and weighting observations by the inverse of the number of estimates in each paper. Cells marked in **bold** *indicate estimates for which the original study findings would have been overturned by our analyses.*

Table 15 – Sensitivity of BCFL (Tables 4 through 8, Column 1): Regression discontinuity caliper-like tests using $|z|$-statistics as the running variable

	Regression discontinuity tests using *p*-values as the running variable	
Threshold:	10% significance	5% significance
	(1)	(2)
Outcome:		
Desk rejected (vs not desk rejected)	**0.017**	-0.002
	(0.085)	(0.078)
Effective obs.	3,056	3,861
Weak R&R (vs Reject)	-0.085	0.008
	(0.064)	(0.052)
Effective obs.	4,468	5,283
Strong R&R (vs Reject)	-0.056	0.066
	(0.057)	(0.045)
Effective obs.	4,440	6,103
Accepted version (vs initial version)	-0.056	0.041
	(0.084)	(0.087)
Effective obs.	2,984	3,229
Accepted (vs rejected)	-0.062	-0.044
	(0.064)	(0.086)
Effective obs.	3,101	3,014
Published (vs never published)	-0.080	0.071
	(0.091)	(0.088)
Effective obs.	2,618	2,333

This table presents a robustness check of Brodeur et al.'s Caliper tests, instead using regression discontinuity tables. In these regressions the dependent variable is the paper status (i.e., accepted), the running variable are estimates' t-statistics, and treatment is coded as one if the running variable is over the 10% and 5% significance critical values, in Columns (1) and (2). These estimates were produced using the rdrobust package in Stata, and report robust and bias-corrected estimates with CER-optimal bandwidths at both sides of the threshold, controlling for the statistics reported in each paper, correcting for mass points in the running variable, clustering standard errors at the paper level, and weighting observations by the inverse of the number of estimates in each paper. Cells marked in **bold** *indicate estimates for which the original study findings would have been overturned by our analyses.*

Table 16 – BCFL density discontinuity tests, with point estimates and bootstrapped standard errors and percentile confidence intervals

Threshold:	Density discontinuity tests (w/ bootstrapped statistics)							
	5% significance				5% significance			
	Estimate	s.e.	95% percentile CI		Estimate	s.e.	95% percentile CI	
			Lower	Upper			Lower	Upper
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1) Initial submissions	-1.482	(1.862)	-5.383	1.952	1.328	(1.218)	-0.958	3.880
Desk reject stage								
(2) Desk rejections	0.898	(1.463)	-1.669	3.793	0.822	(1.199)	-1.570	3.166
(3) Not desk rejections	-3.681	(1.618)	-7.059	-0.784	1.753	(1.246)	-0.582	4.196
Reviewer stage								
(4) Rejection	-1.211	(2.110)	-6.314	2.118	3.427	(1.385)	0.438	5.835
(5) Against outright rejection	-2.797	(1.652)	-5.864	0.798	-0.162	(1.720)	-3.818	2.790
(6) Accept as is or with minor edits	-0.938	(1.731)	-3.926	2.185	-0.207	(1.298)	-2.722	2.362
Accepted manuscripts								
(7) Initial submission	-3.499	(1.214)	-6.316	-1.435	0.597	(1.310)	-1.893	3.256
(8) Published version	-2.112	(1.254)	-4.415	0.264	-0.127	(1.244)	-2.295	2.673
Peer review								
(9) All rejections	-1.083	(1.663)	-5.416	1.565	2.875	(1.121)	0.918	5.197
(10) Accepted manuscripts	-	-	-	-	-	-	-	-
After rejection								
(11) Published elsewhere	0.591	(1.243)	-2.098	2.699	1.926	(1.177)	-0.294	4.184
(12) Failed to publish	-0.358	(2.153)	-3.147	0.756	1.915	(1.233)	0.261	4.611

Columns (1), (2), (5) and (6) of this table reproduce the test statistics with bootstrapped standard errors corresponding to Brodeur et al. (2023), Table 2, Columns 2 and 4. These underlying tests are local polynomial density discontinuity tests on p-curves at the 0.05 and 0.10 thresholds. Brodeur et al. (2023) estimate these in R using the package rddensity with all the default options whereas in this table we use Stata's rddensity with the same default options. In addition, we jointly bootstrap the estimates at each stage 1,000 times drawing manuscripts with replacement in order to construct a joint variance-covariance matrix. Byproduct of this are the bootstrapped standard errors in Columns (2) and (4) and the 95% percentile bias-corrected and accelerated confidence intervals in Columns (3) and (4) and (7) and (8).

Table 17 – Hypothesis tests using the joint variance-covariance matrix of BCFL's density discontinuity tests constructed using bootstrapped estimates

Test #	Null hypothesis	Restriction tested in Table 16	Test statistic	p-value
1	There is no heaping in the distributions of initial submissions	Coefficients in Columns (1) and (5) of Row (1) are jointly equal to zero	$\chi^2(2) = 1.65$	0.438
2	Heaping remains the same for desk rejections and not desk rejections	Coefficients in Rows (2) and (3) are equal in Column (1) and in Column (5)	$\chi^2(4) = 4.19$	0.123
3	Heaping remains the same across reviewer recommendations	Coefficients in Rows (4), (5) and (6) are equal in of Column (1) and in Column (5)	$\chi^2(4) = 5.44$	0.245
4	Heaping remains the same for initial submissions and published versions	Coefficients in Rows (7) and (8) are equal in Column (1) and in Column (5)	$\chi^2(2) = 0.72$	0.699
5	There is no heaping at any stage in the publication process	Coefficients in all rows and columns are jointly equal to zero	$\chi^2(22) = 43.89$	0.003

These tests are all based on the joint variance-covariance matrix of local polynomial density discontinuity tests on p-curves at the 0.05 and 0.10 thresholds across samples underlying Table 10. The matrix was constructed by jointly bootstrapping estimates at each stage of the publication process 1,000 times drawing manuscripts with replacement. We used Stata's rddensity with the default options as the core estimation command, bootstrap to construct the matrix, and test for producing chi-squared statistics and p-values.

Figure 1 – Distribution of *p*-values and |z|-statistics from BCFL's initial submissions