UNPACKING P-HACKING AND PUBLICATION BIAS*

Abel Brodeur[a], Scott Carrell[b†], David Figlio[c], Lester Lusher[d]

[a]University of Ottawa and IZA

[b]University of California, Davis, NBER, and IZA

[c]Northwestern University, NBER, and IZA

[d]University of Hawaii at Manoa and IZA

May 4th, 2021

Abstract

We use unique data from journal submissions to identify and unpack publication bias and p-hacking. We find that initial submissions display significant bunching, suggesting the abnormal distribution among published statistics cannot be fully attributed to a publication bias in peer review. Desk rejected manuscripts display greater heaping than those sent for review, suggesting editors "sniff out" marginally significant results. Reviewer recommendations, on the other hand, are swayed significantly by statistical thresholds. The overall effect of peer review slightly smooths the distribution of test statistics. Lastly, tracking rejected papers, we find that never-published manuscripts display greater heaping than their eventually-published counterparts.

Keywords: Publication bias - p-hacking - selective reporting

JEL codes: A11, C13, C40

*We thank Adrian Amaya, Mohammad Elfeitori, Saori Fuji, Connor McKenney, Clemence Mugabo, Markus Tran, Shannon Tran, and Qian Yang for providing excellent research assistance. We also thank seminar and conference participants at BITSS, PUC-Chile, Shidler College of Business, UH Manoa and UC Santa Cruz for very useful remarks and encouragements. All errors are our own.

†Abel Brodeur, Email: abrodeur@uottawa.ca; Scott Carrell, Email: secarrell@ucdavis.edu; David Figlio, Email: figlio@northwestern.edu; Lester Lusher, Email: lrlusher@hawaii.edu

1 Introduction

P-hacking and publication biases are generally perceived to be pervasive issues in academia. Publication bias stems from a potential preference among editors and reviewers for results that display statistical significance. Furthermore, a belief about the existence of a publication bias may encourage authors to shelve a study if initial results are undesired or unpromising. P-hacking refers to actions that authors engage in, knowingly or otherwise, in order to produce "more favorable" p-values. Such actions include continuing to collect data, tinkering with econometric specifications, and imposing sample restrictions until certain thresholds of statistical significance are met. These behaviors may have large consequences as studies reporting significant effects of a particular program or policy may be more likely to end up published than studies with null results. This selectivity would then lead to biased estimates and misleading confidence sets in published research.

A large and growing literature has discussed potential publication biases and specification searching in economics and other disciplines (Andrews and Kasy, 2019; Ashenfelter et al., 1999; Bruns et al., 2019; De Long and Lang, 1992; Doucouliagos and Stanley, 2013; Ferraro and Shukla, 2020; Furukawa, 2020; Havránek, 2015; Ioannidis, 2005; Ioannidis et al., 2017; Leamer, 1983; McCloskey, 1985; Miguel et al., 2014; Stanley, 2005, 2008).[1] To document these phenomena, studies have plotted the distribution of test statistics from published manuscripts in a given literature or in top journals, finding significant bunching at well-known thresholds of statistical significance (e.g. Brodeur et al., 2016, 2020; Gerber and Malhotra, 2008a,b; Vivalt, 2019).

Since prior studies have strictly relied on *published* papers, one cannot convincingly identify the direct impact of the peer review process on the distribution of test statistics. Unpacking the role of authors, editors and reviewers is key for better understanding the extent and sources of p-hacking and publication bias. For instance, it may be that authors do not engage in p-hacking and the distribution of test statistics among submitted papers is normal, but then a publication bias distorts the distribution toward heaping at significance thresholds. Conversely, p-hacking may be so prevalent that the distribution of test statistics is even more skewed among journal submissions versus publications, suggesting that the peer review process mitigates the consequences of p-hacking. Potential interventions to combat these channels differ as well: For author behavior, academia has promoted pre-registrations of experiments and pre-analyses plans for empirical work, while other interventions such as pre-results review[2] and bias-corrected estimators and confidence

[1] See Christensen and Miguel (2018) for a recent relevant literature review, Stanley (2008) and Doucouliagos and Stanley (2013) for surveys of meta-regression methods and Havránek et al. (2020) for recent guidelines for meta-analysis. A growing literature also discusses which findings should be published (see, for example, Abadie (2020) and Frankel and Kasy (2021)).

[2] Pre-results review involves the reviewing and acceptance of detailed proposals for research studies prior to results being col-

intervals (e.g. Andrews and Kasy, 2019) target potential publication biases.

This study is the first, to our knowledge, to collect test statistics from manuscripts across the spectrum of the peer review process, from initial submission to desk rejection to reviewer reports to (potential) publication, in order to unpack the extent of p-hacking and publications bias on published statistics. Our data include over 11,000 test statistics across a random sample of nearly 400 manuscripts submitted for review to a prominent applied microeconomics journal (*Journal of Human Resources*) from the years 2013 to 2018.

We first find that the distribution of test statistics among submitted articles displays a hump around 10 and 5% significance thresholds, providing direct evidence that the abnormal distribution among published statistics cannot be fully attributed to a publication bias in peer review. Furthermore, solo-authored submissions display greater heaping than their multi-authored counterparts, arguably implying a strong p-hacking channel.

We then find that the distribution of desk rejections display greater bunching than those sent for review, suggesting that editors, on average, "filter out" false positives. We also find that this result is partially explained by author characteristics (e.g. solo-authorship) which correlate with both desk rejection outcomes and propensity to produce marginally significant estimates. Anonymous reviewers, on the other hand, appear to be swayed by statistical significance: As we move from rejection recommendations to strong positive recommendations, the distributions of test statistics sharply increase around significance thresholds. Finally, we find that the distribution of statistics from the final draft of accepted manuscripts is moderately smoother than their initial draft counterparts, once again suggesting that even authors of eventually-published manuscripts engaged in p-hacking prior to their initial submission.[3]

In total, by comparing the final draft of accepted manuscripts against all rejected submissions, we find that the peer review process slightly smooths the distribution of test statistics.[4] That is, the distribution of initial submissions displays greater heaping than does the distribution from published manuscripts. We further track papers after rejection to find that approximately 60% eventually publish elsewhere. To allay concerns that our results are anomalous or unique to our journal, we compare the distribution of tests for manuscripts that fail to publish elsewhere to their eventually published counterparts, finding that manuscripts that fail to publish elsewhere display more statistical bunching. This evidence suggests the presumed prevalence of publication biases is perhaps not as prominent as feared, though concern remains about the impartiality of

lected. Consequently, the journal commits to publishing the subsequent paper regardless of the study's results. Several journals adopting pre-results review in economics include the *Journal of Development Economics* and *Experimental Economics*.

[3] Still, given we focus strictly on "main" estimates in papers (and not robustness checks or heterogeneity analyses), the difference in this latter result is fairly small (i.e. main estimates seldom change from initial to final draft).

[4] Brodeur et al. (2020) compare the distribution of tests in published journal articles to that in the working papers. They find that the distributions are strikingly similar.

anonymous reviewers. At a minimum, our results suggest that the findings from prior studies can likely be more attributed to author behavior as opposed to the peer review process.

The main results from our paper suggest that the statistical bunching observed among published manuscripts cannot be (entirely) explained by the peer review process. Instead, our results suggest that authors engage in actions which cause skewed distributions. These actions are, however, unobserved in our data - for example, authors may refrain from submitting null result papers entirely, or they may tinker with specifications until desired thresholds are met. In an effort to document the types of behaviors authors engage in prior to submission, we conducted an anonymous survey across a broad sample of applied microeconomists. We find that roughly 30% of authors have stopped a research study or refrained from submitting a paper after finding null results within the past five years. We provide evidence that this behavior is in response to beliefs about the importance of statistical significance in influencing the editor's/reviewer's decision. We also asked applied microeconomists about other behaviors and find that around 50% of authors have (at least once) reported only a subset of the dependent variables and/or analyses conducted in the final draft of their paper. Less common behaviors include modifying original hypotheses to better match empirical results (26%), excluding or recategorizing data after seeing the effects of doing so (18%), and selecting regressors after looking at the results (26%). Finally, nearly 30% of authors have (at least once) decided to further expand their analytic sample or conduct more experiments after analyzing data. We also find that these behaviors are broadly consistent across authors who had previously submitted to the *Journal of Human Resources* versus other journals, suggesting our prior peer review results likely apply to journals outside our setting as well.

The findings in our study contribute to the literature in several important ways. Namely, they suggest that p-hacking among initial submissions is a strong driver of validity concerns, and interventions that target curbing author behavior away from p-hacking should be particularly impactful. These include growing practices of preregistering studies and developing pre-analyses plans.[5] Given the observed biases among reviewer recommendations, our findings also reinforce those from Blanco-Perez and Brodeur (2020), who suggest that interventions where editors instruct reviewers to evaluate studies on potential merit regardless of statistical significance can be particularly effective.

Our results also contribute to a large literature on replications and meta-analyses by better informing the form of selectivity in the publication process, which may help researchers to more appropriately correct the bias from selective publication (e.g. Andrews and Kasy, 2019; Havránek and Sokolova, 2020). The two most relevant studies are possibly DellaVigna and Linos (2020) and Franco et al. (2014). DellaVigna

[5]See Casey et al. (2012) for an in-depth example and analyses of a pre-analysis plan, and Ofosu and Posner (2020) for an analysis of the impact of pre-analysis plans on publication rates.

and Linos (2020) find that results from RCTs published in papers have significantly larger treatment effects (8.7pp) compared to RCTs conducted at a larger-scale via Nudge Units (1.4pp). Franco et al. (2014) follow 221 research proposals that won a competitive award to conduct survey-based experiments. They provide evidence that strong results are 40 (60) percentage points more likely to be published (written up) than are null results.

Last, our findings relate to a growing literature documenting editor and reviewer behavior. Card and DellaVigna (2020) find that 1) editor decisions closely follow referee recommendations, 2) papers by highly published authors receive more subsequent citations conditioning on referee recommendations and publication status, and 3) there are no differences in the predictive power of referee publication rate on paper citations, yet editors give significantly more weight to highly published referees. Card et al. (2020) document how the peer review process differentially treats male- and female-authored papers. Carrell et al. (2020) document signaling and network effects in how reviewers evaluate papers written by authors of matching characteristics: For example, the authors find evidence that reviewers positively evaluate research by authors who went to their same PhD program.

2 Data Sources and Background

Our data consist of two parts. The first are collected from the *Journal of Human Resources* (JHR). The JHR is often regarded as a highly selective applied microeconomics field journal. For the full population of submitted papers at this journal, roughly a third are desk rejected by the editor. Reviewers give a rejection recommendation over 50% of the time, while less than 10% of reviewer recommendations are strong positives. The overall acceptance rate at the journal is 6%.

Our sample of data from the JHR contains all manuscripts submitted for review from 2013 to 2018. This sample was then reduced through the following process: First we selected a random sample of 360 manuscripts stratified by year and outcome (desk rejected vs. rejected after reviewer comments vs. accepted) such that each year-outcome combination contained 20 manuscripts. Then, we included initial drafts of accepted papers into our sample.[6] Lastly, upon reading the paper, we removed manuscripts which did not contain a clear identification strategy (difference in differences, instrumental variables, regression discontinuity, and/or randomized control trials) for causal inference; this process closely followed that of Brodeur et al. (2020).[7] Our final analytic sample contains 396 manuscripts: 98 desk rejections, 110 rejections after

[6]For our random sample, there are no instances of a paper receiving a R&R but subsequently being rejected. Thus only accepted manuscripts may contain multiple drafts in our sample. Moreover, there was one paper in our sample that was accepted without reviewer reports, which we omit from our sample.

[7]Examples of omitted papers include literature reviews, methodology papers, descriptive exercises, and other identification

receiving reviews, 94 drafts of eventually-accepted manuscripts, and 94 published drafts.

We then coded coefficients and their standard errors from each paper. Following the previous literature (Brodeur et al., 2016; Blanco-Perez and Brodeur, 2020; Brodeur et al., 2020), we only collect estimates from main results tables. Estimates from summary statistics, appendices, robustness checks, and placebo tests were not collected, nor were results from figures. Within main tables, we only collected coefficients from the variable(s) of interest in the paper; thus we omit obvious regression controls and constant terms. Otherwise, within a main table, all coefficients on the covariate(s) of interest were collected. Any cases of ambiguity were marked accordingly; for our primary estimates, we exclude ambiguous estimates, but robustness analyses check for the sensitivity to the inclusion of ambiguous cases. Ultimately, we collected 11,165 test statistics.

Coefficients and standard errors are reported for the vast majority of tests, while p-values and t-statistics are reported for 3% and 2.5% of tests, respectively. We transform p-values into the equivalent z-statistics. For coefficients and standard errors, we construct the ratio of the two. We thus treat these ratios as if they were following an asymptotically standard normal distribution under the null hypothesis. One issue discussed in the literature is the overrepresentation of round values (e.g. coefficient of 0.02 and standard error of 0.01). As a robustness check, we follow Brodeur et al. (2016) and randomly redraw a number in the interval of potentially true numbers around each collected value using a uniform distribution. This derounding method has little impact on our conclusions. See Appendix 1 for more details on this methodology and results.

The second part of our data consists of manually-collected information on authors and reviewers. The following information was collected by visiting each individual's website(s), Google Scholar webpage and `ideas.repec.org` webpage: gender, institution of PhD, PhD graduation year, tenure status, and prior publication history. Rankings for the prestige of the author's PhD program were also collected from the department productivity rankings on `ideas.repec.org`.[8]

2.1 Summary Statistics

Table 1 presents summary statistics for our sample at the paper level, split by the four categories for paper outcomes: desk rejected, rejected after receiving reviewer comments, first drafts of accepted manuscripts, and final drafts of accepted manuscripts. Desk rejected papers tend to have fewer main estimates (25) compared to those sent our for review. Additionally, accepted submissions tend to contain slightly more

strategies such as propensity score matching.

[8] IDEAS rankings retrieved May 2019 from `https://ideas.repec.org/top/top.econdept.html`.

estimates (30) than submissions rejected at the reviewer stage (28). We deal with these differences in the number of tests reported in each category in two ways in our analysis. First, we use the inverse of the number of tests presented in the same article to weight observations. Second, we present a set of robustness estimates in which we focus on the first table (with main results) for each manuscript.

Next, we find several large discrepancies in author characteristics associated with the paper's outcome. For instance, papers with multiple authors tend to experience better outcomes: desk rejected papers are solo authored at a 41% rate, 29% of those rejected after review are solo authored, and 24% of accepted manuscripts are solo authored. Those with prior publication history at the journal experience more positive outcomes. More experienced authors (measured as years since PhD) and those who came from better ranked PhD programs tend to experience more positive paper outcomes. These correlations are unsurprising since these characteristics are generally associated with higher quality papers. Lastly, turning to identification strategy, randomized control trials appear to have relatively higher likelihood of getting past the desk and subsequently publish at the expense of instrumental variables strategies.

2.2 Where papers go before and after the *Journal of Human Resources*

In this section, we describe the list of journals that authors typically submit to prior to their JHR submission, and which journals authors publish in after rejection at the JHR. To do the former, we conducted a survey (described in greater detail in Section 5) across 143 applied microeconomists who listed which journals they had submitted to in the previous five years. Authors were then asked (for a random subset of journal submissions) which journals they had submitted to prior to a specific journal submission. In Figure 1, we plot the distributions of prior submissions, sorted by journal rank (according to `ideas.repec.org`), for each of several journals of interest including the JHR. In general we first see that (unsurprisingly) authors tend to submit to higher ranked journals first. The most common journal authors submit to prior to a JHR submission is the *American Economic Journal: Applied Economics* (AEJ:AE). The most common prior journals for AEJ:AE submissions are the *American Economic Review* (AER) and the *Quarterly Journal of Economics* (QJE). The distribution of prior submissions to the JHR closely resembles the distribution for the *Journal of Public Economics* (JPubE), perhaps confirming their reputation as top field journals.

To track whether and where papers published after rejection at the JHR, we collected additional data on publishing outcomes using searches on Google Scholar and `ideas.repec.org` for our sample of rejected manuscripts. We note that eventual publication rates are stable for papers submitted from 2013 to 2016, then drop steeply for the years 2017 and 2018; this likely reflects the long publication timeline in economics (median time of 2.25 years from submission to publication at a single journal according to Hadavand

et al. (2020)), thus creating a right-censoring issue of tracking long run publication. Consequently, for all analyses using eventual publication outcomes, we choose to focus on rejected manuscripts submitted prior to 2017. For this sample, the eventual publication rate was roughly 60%: 56% for desk rejected manuscripts and 62% for manuscripts rejected after receiving reviewer recommendations. The most common eventual publication outlets include *Economics of Education Review* (10% of eventual publications), *Journal of Health Economics* (9%), *Health Economics* (6%), *Journal of Economic Behavior and Organization* (6%), *Economic Development and Cultural Change* (5%), and *Labour Economics* (5%).

3 Main Results - Plotting Test Statistics

3.1 Initial submissions

We start with Figure 2 which plots the raw distribution of z-statistics for our full sample of initial submissions. Bins were constructed with a width of 0.10 along the interval $[0, 10]$ for a total of 100 bins. Three vertical lines are drawn for reference to z statistics associated with 10, 5, and 1% significance (1.65, 1.96, and 2.58, respectively). The distribution displays a two-humped shape, with one hump for test statistics below 1, and another around the 5% statistical threshold. Approximately 59, 49 and 35% of test statistics are significant at the 10, 5 and 1 percent levels, respectively. This distribution largely reflects the distribution from published manuscripts in Brodeur et al. (2016) and Brodeur et al. (2020).

These results provide evidence of p-hacking to the extent that the distribution of test statistics faced by editors is already significantly skewed toward statistical thresholds. In other words, we can rule out the case that the distribution of test statistics initially faced by editors is normal, and then a process of publication bias skews the distribution toward statistical significance. Thus, the observed distributions from prior studies cannot be strictly attributed to the peer review process. Also note that this distribution of initial statistics may be driven by a *belief* in a publication bias. That is, if authors' final results are statistically insignificant, and they believe this diminishes their odds of publication, then they may choose to not write up or submit their results.

In Figure 3 we investigate heterogeneity by solo-authorship by splitting the distribution of initial submissions by whether the paper was solo-authored or not. Papers with multiple authors are arguably less prone to p-hacking since multiple authors may be engaging in the paper's analyses and cross-checking sensitive results. Indeed we find that the second hump of test statistics is far steeper for single-authored papers than for multi-authored papers. Still, the heaping for all initial submissions cannot be strictly explained by solo-authorship since a second hump still emerges for multi-authored papers. Later in our supplemen-

tal econometric analyses section 4.1.1, we show that the difference in distributions by solo-authorship is statistically significant, while also investigating other potential heterogeneities.

3.2 Results by desk rejection

In Figure 4, we split the distribution of initial submissions by whether they were desk rejected by the editor, or sent out for review. While both distributions still display heaping at significance thresholds, the peak for desk rejections is much more pronounced. Thus, editors on average seem to "filter out" what are likely to be false positives and attenuate the consequences of p-hacked submissions.

This filtering out by editors could be driven by both conscious decision making (e.g. the editor is skeptical of the paper's statistical inference) and by correlates of paper quality and propensity for marginal significance. In the latter case, it may be that papers of lower quality (and thus higher likelihood for desk rejection) are also more likely to report marginal significance. Econometric results presented later in 4.1.2 find that author characteristics can partially explain the differential bunching in desk rejection versus non-desk rejected statistics, suggesting authors of certain characteristics (e.g. solo-authorship) who write lower quality papers tend to p-hack more.

3.3 Results by reviewer recommendations

Next we turn to all manuscripts sent out for review, splitting by the reviewer's specific recommendation on the paper. Thus, we utilize a dataset at the test statistic-paper-reviewer level, where each paper appears in the data as many times as reviewers it was assigned. Estimates were then split by the reviewer's recommendation on the paper. At this journal, a reviewer can give an overall ranking from 1 to 5, where 1 reflects "Reject" and 5 reflects "Accept as is." Figure 5 presents the distribution of test statistics split by rejection recommendations, non-rejection recommendations (ranking of 2+), and strong positive recommendations (ranking of 4 or 5). The second hump of test statistics around significance thresholds becomes more pronounced as we move from the first figure (desk rejections) to the third figure (strong positive recommendations). This suggests that reviewers have a positive bias toward statistical significance. Moreover, shown later in 4.1.3, the bunching remains even after accounting for author and reviewer characteristics, suggesting marginally significant results are not differentially assigned to reviewers who may happen to have differing propensity for positive or negative reviews.

3.4 Comparing initial vs. final drafts of accepted papers

In Figure 6 we juxtapose the distribution of test statistics from the final draft of accepted manuscripts against their initial submission counterparts. After an initial submission receives a positive response from an editor, authors may be asked to edit their main tables to address editor and reviewer comments. Here we find a larger hump among initial submissions, suggesting the peer review process smooths the distribution of test statistics, and further suggests evidence of p-hacked results among first drafts of eventually-accepted papers. The distributions, however, are somewhat noisy, and shown later do not display statistically significant differences from each other.

3.5 Overall impact of peer review - accepted versus rejected manuscripts

Lastly, in Figure 7 we compare the same distribution of test statistics from the final draft of accepted manuscripts against all rejections. This comparison allows us to evaluate the overall impact of the peer review process by effectively seeing the net effect of the prior three sections: first, editors on average filter out false positives (Figure 4), but then among non-desk rejections, reviewer recommendations favor statistical significance (Figure 5). Then, editors take reviewer recommendations to decide which of these papers deserve a chance for revision, and authors of these papers tend to produce a more smoothed distribution in their final revision (Figure 6). The net effect of these processes in Figure 7 shows more bunching among rejected papers than accepted papers just past the 10% significance threshold, suggesting that the "filtering out" of marginally significant results trumps the reviewer bias effect. This is perhaps unsurprising given that editors ultimately decide the final outcome of the paper, both along the desk rejection margin and the final acceptance margin. Note, however, that further analyses suggest that the differences between these distributions are not significant, particularly after controlling for covariates. Still, these results suggest that the overall peer review process does *not* favor statistical significance, and if anything, it filters out marginally significant results.

3.6 After journal rejection - eventually published versus never published manuscripts

The prior sections highlighted what happens to the distribution of test statistics at each stage of the peer review process. In this section, we investigate what happened to the papers that were rejected in our sample, with a particular focus on whether a rejected manuscript published elsewhere and whether publication is associated with greater p-hacking. If the distribution of eventually published manuscripts displays greater heaping, then this suggests there may still be a significant publication bias in the profession overall, and

that the impacts from our single journal on the distribution of published test statistics may be negligible or anomalous. To do so, we turn to our data that matched rejected papers to their (potential) eventual publication outlet. Recall that this sample focuses strictly on submissions to the JHR from 2013 to 2016 in order to allow submissions adequate time to publish after their JHR rejection.

Figure 8 compares the distribution of test statistics for manuscripts that published elsewhere after rejection versus those that failed to publish elsewhere. Though both distributions display significant heaping, never published manuscripts experience a sharper jump at the 5% threshold; shown later in our econometric analyses, the difference in heaping at the 5% level is statistically significant, and cannot be explained by observed differences in author or paper characteristics. This result rules out the idea that there is a "graveyard" of null result working papers that fail to publish, and suggests that the peer review phenomena identified in the prior sections are likely applicable to the broader economics profession. That is, the peer review process overall tends to slightly filter out false positives, and/or (unobservedly) "bad" papers tend to p-hack more.[9]

To sum up, we find that: initial submissions display significant bunching; co-editors on average attenuate the consequences of p-hacked submissions; reviewers in contrast have a positive bias toward marginally significant results; and that papers never published possess more marginally significant results. These results suggest that researchers engage in p-hacking prior to submitting to academic journals, possibly in response to beliefs about preferences of editors and reviewers for significant results. They also suggest that many papers with non-significant results are never submitted. We come back to these issues later when discussing our survey results, but first formally confirm these findings using two econometric approaches.

4 Additional Econometric Analyses

In this section, borrowing from several studies including Andrews and Kasy (2019) and Brodeur et al. (2020) among others, we investigate our results in the context of two separate econometric tests. These tests provide two distinct advantages beyond the graphical analysis. First, they allow us to quantify and conduct inference on the extent of the statistical bunching. Second, they allow us to identify other factors/heterogeneities that potentially mediate the observed bunching, such as differential assignment to co-editor or author characteristics. We first present results from Caliper tests, then follow with results from the excess test statistics method. See Appendix 2 for a thorough discussion of the econometric methods and

[9] Since this exercise only tracks papers submitted to the JHR and their subsequent long run outcomes, these findings may not be applicable to the peer review process at journals ranked higher than the JHR. We note, however, that if there is also a strong belief that higher ranked journals require greater statistical significance, then it would be the case that the distribution of initial submissions at top journals is even more heaped at thresholds than our observed distributions. Though we cannot provide direct evidence of this, our belief is that our identified phenomena are likely to be heightened at higher ranked journals where there is likely a stronger belief of required statistical significance.

specifications.[10]

4.1 Caliper Test

4.1.1 Does marginal significance differ by author and paper characteristics?

We start with the Caliper test to investigate selective reporting by author and paper characteristics near statistical significance thresholds among initially submitted papers. Table 2 tests whether our vector of covariates are significantly associated with marginal significance at the 10, 5, and 1 percent levels in the first, second, and third columns, respectively. Each column presents results from a single regression. We report standard errors adjusted for clustering by article in parentheses. We use the inverse of the number of tests presented in the same article to weight observations. We restrict the samples to $z \in [1.15, 2.15]$, $z \in [1.46, 2.45]$ and $z \in [2.08, 3.08]$ for 10, 5, and 1 percent levels, respectively. Positive coefficients suggest an increase in the likelihood that the reported test statistic is marginally significant.

The most notable result confirms the finding suggested in Figure 3, where papers that are solo-authored contain significantly more marginally significant test statistics than multi-authored papers. In particular, solo-authored papers are 17.5 percentage points more likely to report a marginally significant result at the 5 percent level compared to multi-authored papers. Another heterogeneity worth noting comes from the paper's identification strategy, where difference-in-differences and instrumental variables papers tend to contain more marginally significant estimates compared to regression discontinuities. This result is comparable to that of Brodeur et al. (2020), who look at differences in statistical bunching by identification strategy among *published* manuscripts. Given the similarity in our estimates, this suggests that the results in Brodeur et al. (2020) cannot be driven by the peer review process being biased simultaneously toward a) marginal significance and b) particular identification strategies. Finally, other considered heterogeneities do not appear to be significantly associated with marginal significance, including author tenure, years since PhD, gender, prior publication at the journal, and author PhD prestige.

4.1.2 Editors desk reject marginally significant results

In this section, we test whether submissions that are desk-rejected are more or less likely to report marginally (in)significant estimates. The dependent variable indicates whether a test statistic is statistically significant at the 10 and 5 percent levels in Panels A and B, respectively, of Table 3, respectively (see

[10] Another similar method used in Brodeur et al. (2020) is randomization tests. This method has the advantage of relying on small windows to effectively test whether the mass of tests just above versus just below a threshold differ significantly by status. See Appendix 2 for a description of this method and Appendix Tables A19 and A20 for our analysis. Our conclusions are robust to the use of randomization tests.

Appendix Table A1 for the 1% statistical significance threshold).[11] Coefficients for the variable "Desk Rejected" reflect increases in the probability of marginal statistical significance relative to the baseline category (not desk-rejected). In columns 1–4, we restrict the sample to $z \in [1.15, 2.15]$ and $z \in [1.46, 2.46]$ for 10 and 5 percent levels, respectively. Our sample size consists of about 2,000 test statistic observations.

In the most parsimonious specification, we find that desk-rejected estimates are about nine percentage points more likely to be statistically significant at the 10% level than estimates in manuscripts that are not desk-rejected. The estimate is statistically significant at the 5 percent level. This provides some evidence that editors, on average, filter out false positives (desk rejected papers display significantly more bunching at the 10% level). In contrast, desk-rejected estimates are *not* statistically more likely than non-desk rejected estimates to be marginally statistically significant at the 5% level.

At this journal, each manuscript is assigned one handling co-editor, each of whom have complete autonomy over rejection, revision, and publication decisions.[12] One plausible explanation for our findings is that co-editors may have been differentially assigned papers with marginally (in)significant estimates, and that co-editors may have different propensities to desk reject papers. More specifically, it may be that co-editors with a high propensity to reject papers tended to receive submissions with marginally significant results. We provide evidence that this is not the case by enriching our specification with co-editor fixed effects (column 2). The point estimate in Panel A increases in magnitude (to 13 percentage points) and is now statistically significant at the 1% level. The point estimate for the 5% significance level also increases but remains statistically insignificant.

Moving to column (3), we similarly test whether differences in the paper's identification strategy can explain the results. In our setting, it may be that certain identification strategies are both more likely to be desk rejected and to be p-hacked. Similar to the results with co-editors, the statistical bunching cannot be explained by the paper's identification strategy. Results for 10% significance remain statistically significant at the 5% level, while the results for 5% significance remain positive but insignificant.

Finally, in column (4) we include the full vector of author characteristics. This model presents several interesting findings. Our results for 10% significance shrink from 12 percentage points to 9, and lose their statistical significance. This suggests that certain types of authors who are more prone to finding marginally significant results are also writing papers which tend to be desk rejected. As identified in the previous section, one of these correlates is solo-authorship: From Table 3 we see a large increase in the likelihood of

[11] We find no evidence that marginally rejecting the null hypothesis at the 1% level is related to desk-rejection rates.

[12] Our sample spans 24 individual co-editors. In total, these 24 co-editors have served across 40 journal editorial boards in economics.

desk rejection for solo authored papers (statistically significant at the 10% level).[13]

4.1.3 Reviewer bias toward statistical significance

We now turn to the question of whether manuscripts that were positively reviewed by external reviewers were more or less likely to contain marginally significant estimates compared to negatively reviewed manuscripts. In other words, we test whether reviewers have a preference toward statistical significance among papers that were not desk rejected. We present these results in Table 4 (see Appendix Table A2 for the 1% statistical significance threshold). The first four columns of this table reflect the same structure as Table 3. The variables of interest are $WeakR\&R_{sr}$ and $StrongR\&R_{sr}$ which equal one if the reviewer's recommendation was weakly positive or strongly positive, respectively. In this analysis, we only focus on the first round of reviews (i.e. we drop any additional rounds of review conducted after the first).

Overall we find evidence of a "reviewer bias" toward statistical significance. From column (1) in Table 4, we see that papers that received either a weakly positive or strongly positive review were more likely to be marginally significant at the 10% level than negative reviews (though these estimates are not statistically significant). Then, sequentially including covariates through the next three columns, our estimates become more precise and slightly increase, culminating in a large and statistically significant bunching effect for strong positive reviews (relative to negative reviews).[14] This suggests that at the reviewer stage, differences in reviewer recommendations by marginal significance cannot be explained the paper's co-editor, identification strategy, or author characteristics. Note that this result differs from the desk rejection results, where the co-editor was "filtering out" marginally significant results via desk rejection, and this filtering out was partially explained by author characteristics (authors who submit marginally significant estimates also tend to write desk-rejected papers).

Much like in our previous analysis where we included co-editor fixed effects to account for potential correlations in an editor's set of manuscripts and the editor's propensity for rejection, in column (5) we include a vector of reviewer-level covariates to account for potential correlation in (a) the assignment of manuscripts with marginally significant results to (b) reviewers with a higher propensity to review manuscripts positively. We find virtually no difference in our estimates between columns (4) and (5), suggesting editors do not choose reviewers based on both the paper's marginal significance and the reviewers propensity to review papers positively or negatively, and further supporting the notion that reviewers are biased toward statistical

[13] As a final robustness check, in column (5) and (6), we replicate columns (2) and (4) respectively but for a narrower bandwidth of test statistics, and the results remain largely unchanged.

[14] Interpreting our results from column (4), estimates from strong positive reviews are nearly 12 percentage points more likely to be marginally significant at the 10% level compared to desk rejections, while weak positive reviews are over 3 percentage points more likely to be marginally significant.

significance. Finally, in Panel B, we turn to potential reviewer preference for statistical significance at the 5% level, where we overall find an effect on weakly positive reviews but not strong positive reviews.

4.1.4 Comparing accepted manuscripts against their first draft

This section analyzes how peer review influences the distribution of test statistics among papers that were ex-post accepted for publication. We present estimates from the Caliper tests in Table 5 in the same manner as in Table 3 (see Appendix Table A3 for the 1% statistical significance threshold). This analysis focuses strictly on the first and final drafts of published manuscripts (i.e. middle drafts are ignored). Similar to the graphical evidence, with positive coefficients, we see that initial submissions were more likely to display marginally significant results. Note, however, that not only are these estimates imprecisely estimated, but the magnitude of the effects are rather small, seldom exceeding two percentage points. Recall that our data collection process only involved "main" tables, and not robustness checks or secondary heterogeneity analyses. Given responses to reviewer and editor comments likely manifest through robustness checks and supplementary analyses, we find it unsurprising that there is little change in the probability of reporting a marginally significant estimate between a paper's first and final main results. Still, the lack of a negative effect reveals that the peer review process among accepted papers does *not* push papers toward marginally significant estimates.

4.1.5 Comparing accepted manuscripts against all rejections

Finally, in Table 6 (see Appendix Table A4 for the 1% statistical significance threshold), we test for the overall impact of peer review on initial submissions, comparing the final drafts of accepted manuscripts against all rejected manuscripts (desk rejects or after review). Note that these models are effectively aggregating all the observed effects from editor desk rejections and reviewer recommendations while controlling for editor, author, and paper characteristics. In total we find little difference in the propensity for marginal significance in accepted manuscripts versus rejections. Thus, it seems that after accounting for various covariates, the peer review process does not exacerbate nor attenuate issues of p-hacking.

4.1.6 Comparing published elsewhere manuscripts against never published

Table 7 (see Appendix Table A5 for the 1% statistical significance threshold) tests whether there are statistically significant differences in distributions for manuscripts that published elsewhere after rejection against those that failed to publish anywhere. As described in section 3.6, this analysis focuses strictly on rejected manuscripts that were submitted before 2017. We first note in Panel A that never published

manuscripts display *less* bunching at the 10% threshold, but the point estimates are mostly statistically insignificant. In Panel B, we see that never published manuscripts possess significantly more marginally significant estimates at the 5% level. The difference cannot be explained by other covariates that may be associated with both probability of rejection and probability of p-hacking (e.g. solo authorship). This finding provides evidence that the observed affects from the prior sections are likely broadly applicable to the economics profession, and cannot be strictly attributed to our journal being an outlier. At a minimum, this result rejects the idea that there may be a "graveyard" of null result working papers that fail to publish due to a publication bias.

4.1.7 Further robustness checks

For further robustness, we conduct additional Caliper tests in the Appendix. Appendix Tables A6-A10 mirror our primary tables except they additionally include test statistics that we coded as "ambiguous" during the data collection phase. These results are very similar to our main results. Then, to account for the possibility that papers across different phases of the peer review process have differing quantities of main results tables, Appendix Tables A11-A15 conduct Caliper tests for our primary bandwidths while restricting our sample to the first main results table for each manuscript. Again, these results are our largely similar to the main estimates, with estimated magnitudes generally increasing along with their standard errors.

4.2 Excess analysis

For our second method, we evaluate the extent of selective reporting by comparing the observed distribution of test statistics for each group of manuscripts to counterfactual distributions.[15] We follow Brodeur et al. (2020) and flexibly calibrate a different counterfactual t distribution to each group of manuscripts. This method allows to endogenize the potential differences in the distribution of tests across groups of manuscripts. This is an important robustness check if desk-rejected manuscripts are, for instance, less powered or use different methods than non-desk-rejected manuscripts.

We calibrate a non-central input distribution by group of manuscripts. We assume that there are no selective reporting –publication bias and p-hacking– above $z > 5$. This assumption is based on the idea that there should be no incentives to engage in specification searching past the 0.1% threshold. We proceed as follows. For 0 to 10 degrees of freedom, we calculate the non- centrality parameter that minimizes the

[15] As a robustness check, we show in the Appendix that our conclusions remain unchanged if we hypothesize that the underlying distribution of tests follows either a t distribution with 1 degree of freedom or a Cauchy(0,0.5) distribution. The choice of these two counterfactual distributions is based on the fact that the observed distribution of tests for $z > 5$ for our different groups of manuscripts behaves similarly to these two input distributions in this region of statistical significance.

difference in the range $[5, \infty)$ between the observed distribution and the expected distribution. Importantly, we compute this difference for each group of manuscripts individually. We then choose the 'best' of the 10 optimized t distributions by degree of freedom.[16]

Appendix Figures A11, A12 and A13 compare the distribution of tests for each group of manuscripts to counterfactual distributions. Two striking facts are worth noting. First, the optimization procedure provides very precise fitting curves. Second, the difference between the observed and input distributions is positive between 1.65 and 2.58 and negative from 0 to 1.65 for each group of manuscripts suggesting selective reporting for all our publication outcomes.

Table 8 presents our results for the following intervals: : $[0 - 1.65]$, $[1.65 - 1.96]$, $[1.96 - 2.58]$, $[2.58 - 5]$ and $[5, \infty)$. To calculate the excess statistics in these intervals, we use the CDF of observed t statistics $\widehat{F}(upper) - \widehat{F}(lower)$ and subtract $F_{t(df,np)}(upper) - F_{t(df,np)}(lower)$. (Recall that all our groups have 2 degrees of freedom but different optimal non-centrality parameters.) This subtraction allows us to calculate misallocation of tests as the difference between the observed and expected distributions for each group given our calibration in the interval $z > 5$.

We first look at desk-rejected manuscripts. The mass difference between expected and observed for the non-significance region is negative and quite large with a dearth of 19.5% in comparison to only 8.6% for non-desk-rejected manuscripts. These 'missing' tests can be found, in part, in the regions $[1.65 - 1.96]$ and $[1.96 - 2.58]$, with an excess of 4.3% and 4.1%, respectively. For non-desk-rejected tests, the 'missing' tests can be found mostly in the $[2.58 - 5]$ and $[5, \infty)$ regions, with a surplus of 4.3% in both regions. These results are consistent with our previous findings of misallocated tests at the 10% level for desk-rejected manuscripts.

At the reviewer stage, we find confirm our previous findings that there is relatively more misallocation of tests in the 10% and 5% significance regions for Strong R&R (and to some extent (Weak R&R) reviews than negative reviews. More precisely, we find an excess of tests for Strong R&R reviews of about 5% for the $[1.96 - 2.58]$ region in comparison to 1.8% for the negative reviews.

We also investigate the extent of misallocation for accepted manuscripts. We find that the extent of misallocation is quite similar between initial and final versions for the 10% and 5% significance regions, but that tests in the final version are more likely to be misallocated for the region past the 1% significance threshold than for accepted manuscripts' initial version.

Last, we do not find much evidence for misallocation differences between accepted and all rejected

[16] We explore the entire region of $0 < df < 10$ and $0 < np < 4$. Of note, we optimize at 2 degrees of freedom for all groups of manuscripts, but the non-centrality parameter differs across groups. This is akin to Brodeur et al. (2020) who optimize at 2 degrees of freedom for each method.

manuscripts. Rejected tests are slightly more likely to have misallocation in the $[1.65 - 1.96]$ region, while accepted manuscripts are more likely to be misallocated past the 5% significance threshold.

5 Results from Anonymous Survey

The main results from our paper suggest that the peer review process has an overall negligible effect on the distribution of published test statistics. Consequently, the observed statistical bunching along popular thresholds must be driven by decisions authors made in the process of conducting research and writing up their results. These actions are, of course, unobserved in our data, yet they are important to identify and understand in order to best implement potential policies to combat selective reporting.

In an effort to document the types of behaviors authors engage in prior to submission, in early 2021 we conducted an anonymous survey across a broad sample of applied microeconomists. In particular, we collected emails for all authors who had published a paper using one of the four identification strategies in our sample (IV, DID, RD, RCT) in a top 25 journal in the year 2018. The journals selected mirror the sample selection from Brodeur et al. (2020). We then dropped authors whose email we could not find or with invalid emails. Ultimately, we sent an invitation email to 561 authors, 143 of whom fully completed our survey. The survey asked questions about the author's publication history, submission history (in the past five years), and their behavior in conducting research.

The survey results regarding the author's behavior are presented in Table 9. Of particular interest, we first find that approximately 30% of authors have stopped a research study or refrained from submitting a paper after finding null results. This result directly speaks to the distribution of test statistics for initial submissions to the *Journal of Human Resources*, and confirms our intuition that many null results are never submitted to academic journals.

We also investigate beliefs about the importance of statistical significance in influencing editor and reviewer decisions. We find that on a scale from 1 to 10, with 10 being "very important," authors on average reported an 8 in response to the following question: "For studies that are claiming to identify an effect of x on y, how important do you think statistical significance is in influencing the editor's/reviewer's decision, ceteris paribus?" We also asked six additional questions about the respondent's behavior over the previous five years. Roughly half of authors have (at least once) reported only a subset of the dependent variables and/or analyses conducted in the final draft of their paper. Less common behaviors include modifying original hypotheses to better match empirical results (24%), excluding or recategorizing data after seeing the effects of doing so (17%), and selecting regressors after looking at the results (24%). Finally, around 26%

of authors have (at least once) decided to further expand their analytic sample or conduct more experiments after analyzing data.

In a separate exercise, we predict whether authors stopped a research study or refrained from submitting a paper after finding null results (i.e. an indicator equal to one if the survey-taker responded positively to either of the first two categories in Table 9) as a function of their 10 point scale beliefs (i.e. the final question in Table 9) and their publication history. We first find that authors who believe that statistical significance is important for publication are significantly more likely to stop their research study or refrain from submitting their paper after finding null results (coefficient of 0.08, p-value of 0.02).[17] Thus, the "missing" null results among initial submissions are driven by authors who possess stronger beliefs in a publication bias.

We also investigate whether authors with a wider spread of publications, and authors who had previously published in a top five journal, were more likely to stop or withhold their study after finding null results. Assuming well-published authors write stronger papers, this would suggest that higher quality null result papers would be less likely to be submitted for review compared to their lower quality counterparts. This is important given that we see greater p-hacking among desk rejected papers - if never-submitted null articles are of higher quality, then the effect we observe in this paper would in fact be understating the true "filtering out" effect done by editors, since these higher quality papers would be more likely to get past the desk, thus further smoothing the distribution of non-desk rejected papers. We confirm that authors with a greater spread of publications and with a top five journal publication were more likely to stop or withhold their study after finding null results, though both of these estimates are statistically insignificant at conventional levels.

Another benefit from conducting the survey is that we can compare differences in author behavior for those who submitted to the *Journal of Human Resources* relative to other journals and other authors in their field. To start, we report in the second column of Table 9 the means for respondents whose research specialty was at least one of public, labor, education, or health economics. Then, in the next column, we report the means for authors who had submitted to the *Journal of Human Resources* at least once in the prior five years. We then repeat the same exercise in the remaining columns for authors who had submitted to the *American Economic Review* (AER), *American Economic Journal: Applied Economics* (AEJ:AE), *Journal of Labor Economics* (JoLE), *Journal of Public Economics* (JPubE), and *Labour Economics* (Labour). Overall we find that behaviors are consistent across authors who had previously submitted to the *Journal of Human Resources* versus other journals. This suggests that our prior peer review results likely cannot be explained

[17] Interpreting the coefficient, increasing a respondent's belief in publication bias along the 10 point scale by one unit increases the likelihood they stopped or withheld their study after finding null results by eight percentage points.

by a unique set of authors who engage in differential behavior at the *Journal of Human Resources* relative to authors submitting to other journals.

6 Conclusion

A large and growing literature has documented abnormal distributions in test statistics among published manuscripts. This study is the first, to our knowledge, to collect test statistics across the full spectrum of the peer review process, from initial submissions to publication, in order to directly identify the effect of peer review on the distribution of test statistics. Our data come from the *Journal of Human Resources*, a journal largely regarded as a top applied microeconomics journal. Test statistics were collected from a random sample of nearly 400 manuscripts submitted from the years 2013 to 2018.

We first find that initial submissions display significant heaping at common thresholds of statistical significance (e.g. 5 percent), suggesting that findings from earlier studies cannot be strictly attributed to the peer review process. Solo-authored initial submissions also display significantly more bunching than multi-authored papers, perhaps suggesting prominent p-hacking among authors. Then, we find that editors on average "filter out" papers with false positives, where papers sent for review display significantly less bunching than desk rejected papers. Anonymous reviewers, on the other hand, appear to be influenced by statistical significance: Papers with (strong) positive scores are more likely to possess marginally significant results. In total, the peer review process modestly improves the statistical bunching problem, where the final draft of accepted manuscripts display a slightly smoother distribution than rejected manuscripts. Thus, our results suggest that author behavior (as opposed to peer review) is the primary culprit for issues of marginal significance.

We conduct two additional exercises to further unpack the role of authors. We first conduct an anonymous survey across a broad sample of applied microeconomists and find that approximately 30% of authors have stopped a research study or refrained from submitting a paper after finding null results. This result is possibly driven by authors' beliefs that a publication bias exists as we find that most economists report that statistical significance is important in influencing the editor's/reviewer's decision.

We then investigate whether the manuscripts rejected in the *Journal of Human Resources* end up published elsewhere, and compare the distribution of test statistics for manuscripts that eventually published elsewhere to those that remain unpublished. We first find that the eventual publication rate was roughly 60%. Interestingly, we also find that the extent of p-hacking is larger for manuscripts that are never published, suggesting that the peer review phenomenon that we identify is likely broadly applicable to peer

review in economics overall.

In total, our results suggest that economists (falsely) believe that editors and reviewers have strong preferences for significant results, leading them to engage in selective reporting prior to submitting to academic journals and withholding their non-significant results from journal submission (Franco et al., 2014).

References

ABADIE, A. (2020): "Statistical Nonsignificance in Empirical Economics," *American Economic Review: Insights*, 2, 193–208.

ANDREWS, I. AND M. KASY (2019): "Identification of and Correction for Publication Bias," *American Economic Review*, 109, 2766–94.

ASHENFELTER, O., C. HARMON, AND H. OOSTERBEEK (1999): "A Review of Estimates of the Schooling/Earnings Relationship, with Tests for Publication Bias," *Labour Economics*, 6, 453–470.

BLANCO-PEREZ, C. AND A. BRODEUR (2020): "Publication Bias and Editorial Statement on Negative Findings," *Economic Journal*, 130, 1226–1247.

BRODEUR, A., N. COOK, AND A. HEYES (2020): "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics," *American Economic Review*, 110.

BRODEUR, A., M. LÉ, M. SANGNIER, AND Y. ZYLBERBERG (2016): "Star Wars: The Empirics Strike Back," *American Economic Journal: Applied Economics*, 8, 1–32.

BRUNS, S. B., I. ASANOV, R. BODE, M. DUNGER, ET AL. (2019): "Reporting Errors and Biases in Published Empirical Findings: Evidence from Innovation Research," *Research Policy*, 48, 103796.

BUGNI, F. A. AND I. A. CANAY (Forthcoming): "Testing Continuity of a Density Via g-Order Statistics in the Regression Discontinuity Design," *Journal of Econometrics*.

CARD, D. AND S. DELLAVIGNA (2020): "What do Editors Maximize? Evidence from Four Economics Journals," *Review of Economics and Statistics*, 102, 195–217.

CARD, D., S. DELLAVIGNA, P. FUNK, AND N. IRIBERRI (2020): "Are Referees and Editors in Economics Gender Neutral?" *Quarterly Journal of Economics*, 135, 269–327.

CARRELL, S., D. FIGLIO, AND L. LUSHER (2020): "Clubs and Networks in Economics Reviewing," Tech. rep., working paper.

CASEY, K., R. GLENNERSTER, AND E. MIGUEL (2012): "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan," *Quarterly Journal of Economics*, 127, 1755–1812.

CHRISTENSEN, G. AND E. MIGUEL (2018): "Transparency, Reproducibility, and the Credibility of Economics Research," *Journal of Economic Literature*, 56, 920–80.

DE LONG, J. B. AND K. LANG (1992): "Are all Economic Hypotheses False?" *Journal of Political Economy*, 100, 1257–1257.

DELLAVIGNA, S. AND E. LINOS (2020): "RCTs to Scale: Comprehensive Evidence from Two Nudge Units," Tech. rep., UC Berkeley mimeo.

DOUCOULIAGOS, C. AND T. D. STANLEY (2013): "Are All Economic Facts Greatly Exaggerated? Theory Competition and Selectivity," *Journal of Economic Surveys*, 27, 316–339.

FERRARO, P. J. AND P. SHUKLA (2020): "Is a Replicability Crisis on the Horizon for Environmental and Resource Economics?" *Review of Environmental Economics and Policy*, 14, 339–351.

FRANCO, A., N. MALHOTRA, AND G. SIMONOVITS (2014): "Publication Bias in the Social Sciences: Unlocking the File Drawer," *Science*, 345, 1502–1505.

FRANKEL, A. AND M. KASY (2021): "Which Findings Should Be Published?" *American Economic Journal: Microeconomics*.

FURUKAWA, C. (2020): "Publication Bias under Aggregation Frictions: From Communication Model to New Correction Method," Tech. rep., MIT mimeo.

GERBER, A. AND N. MALHOTRA (2008a): "Do Statistical Reporting Standards Affect what is Published? Publication Bias in Two Leading Political Science Journals," *Quarterly Journal of Political Science*, 3, 313–326.

GERBER, A. S. AND N. MALHOTRA (2008b): "Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?" *Sociological Methods & Research*, 37, 3–30.

HADAVAND, A., D. S. HAMERMESH, AND W. W. WILSON (2020): "Is Scholarly Refereeing Productive (at the Margin)?" Tech. rep., National Bureau of Economic Research.

HAVRÁNEK, T. (2015): "Measuring Intertemporal Substitution: The Importance of Method Choices and Selective Reporting," *Journal of the European Economic Association*, 13, 1180–1204.

HAVRÁNEK, T. AND A. SOKOLOVA (2020): "Do Consumers Really Follow a Rule of Thumb? Three Thousand Estimates from 144 Studies Say "probably not"," *Review of Economic Dynamics*, 35, 97–122.

HAVRÁNEK, T., T. STANLEY, H. DOUCOULIAGOS, P. BOM, J. GEYER-KLINGEBERG, I. IWASAKI, W. R. REED, K. ROST, AND R. VAN AERT (2020): "Reporting Guidelines for Meta-Analysis in Economics," *Journal of Economic Surveys*.

IOANNIDIS, J. P. (2005): "Why Most Published Research Findings Are False," *PLoS medecine*, 2, e124.

IOANNIDIS, J. P., T. D. STANLEY, AND H. DOUCOULIAGOS (2017): "The Power of Bias in Economics Research," *Economic Journal*, 127, F236–F265.

LEAMER, E. E. (1983): "Let's Take the Con Out of Econometrics," *American Economic Review*, 73, pp. 31–43.

MCCLOSKEY, D. N. (1985): "The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests," *American Economic Review: Papers and Proceedings*, 75, 201–205.

MIGUEL, E., C. CAMERER, K. CASEY, J. COHEN, K. M. ESTERLING, A. GERBER, R. GLENNERSTER, D. GREEN, M. HUMPHREYS, G. IMBENS, ET AL. (2014): "Promoting Transparency in Social Science Research," *Science*, 343, 30–31.

OFOSU, G. K. AND D. N. POSNER (2020): "Do Pre-Analysis Plans Hamper Publication?" *AEA Papers and Proceedings*, 110, 70–74.

STANLEY, T. D. (2005): "Beyond Publication Bias," *Journal of Economic Surveys*, 19, 309–345.

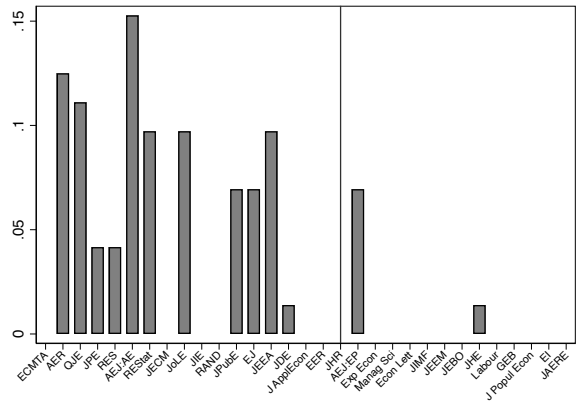
——— (2008): "Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection," *Oxford Bulletin of Economics and Statistics*, 70, 103–127.

VIVALT, E. (2019): "Specification Searching and Significance Inflation Across Time, Methods and Disciplines," *Oxford Bulletin of Economics and Statistics*, 81, 797–816.

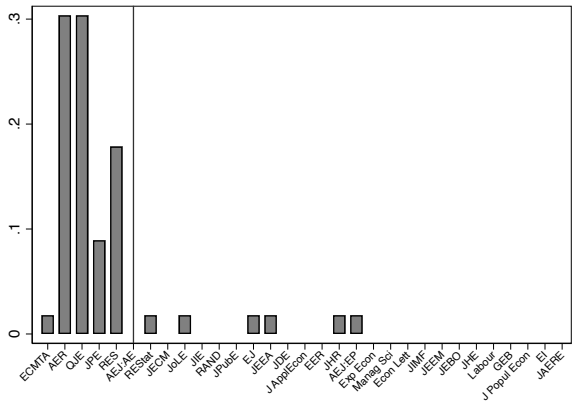
7 Figures and Tables

Figure 1: Distribution of journal submissions made prior to a submission at a particular journal

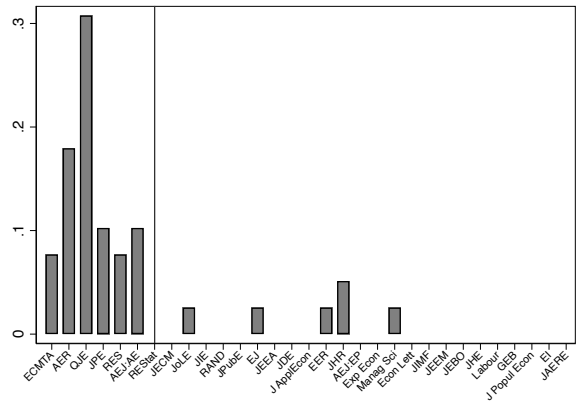
(a) *Journal of Human Resources*



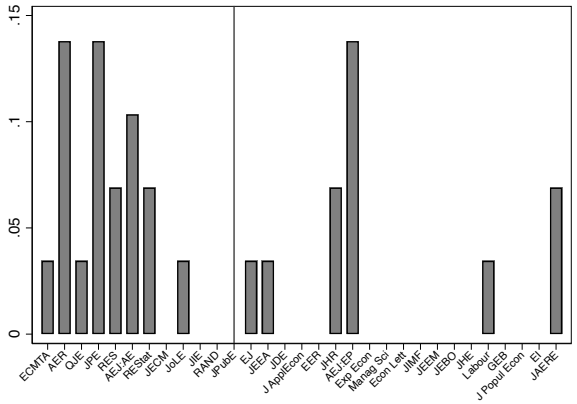
(b) *American Economic Journal: Applied Economics*



(c) *Review of Economics and Statistics*

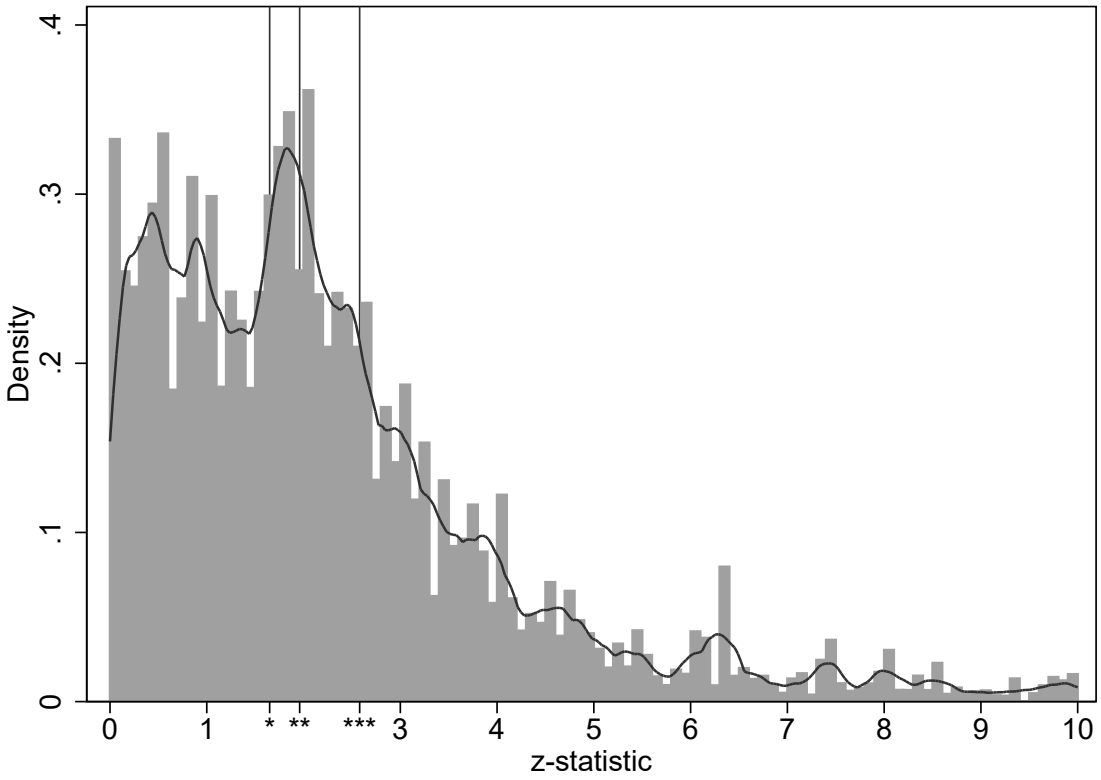


(d) *Journal of Public Economics*



Notes: Results based on survey data described in section 5. Survey participants were first asked to report which journals they had submitted to in the prior five years. Then for a random subset of those journals, participants were asked which journals they had submitted to prior to the relevant journal submission. For example, figure (a) reports the distribution of journals authors had submitted to prior to their most recent journal submission to the *Journal of Human Resources*. Journals along the x-axis are sorted by journal rank retrieved from ideas.repec.org.

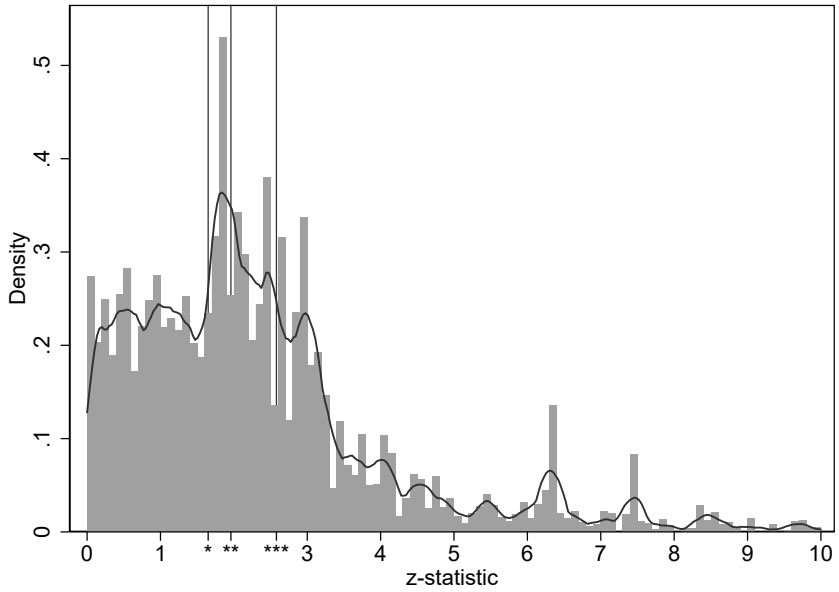
Figure 2: Distribution of z-statistics for initial submissions



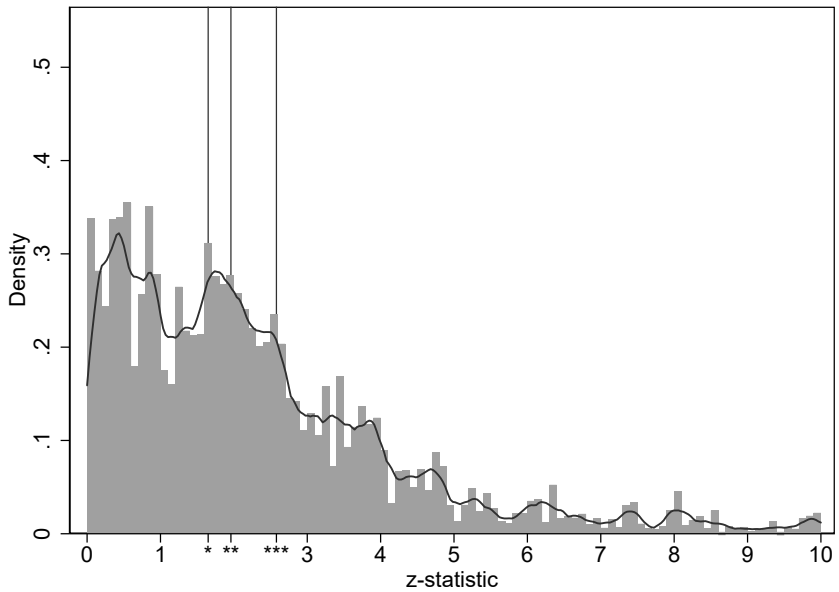
Notes: This figure displays an histogram of test statistics for $z \in [0, 10]$ for initial submission. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure 3: Distribution of z-statistics from initial submissions split by number of coauthors

(a) Solo authored



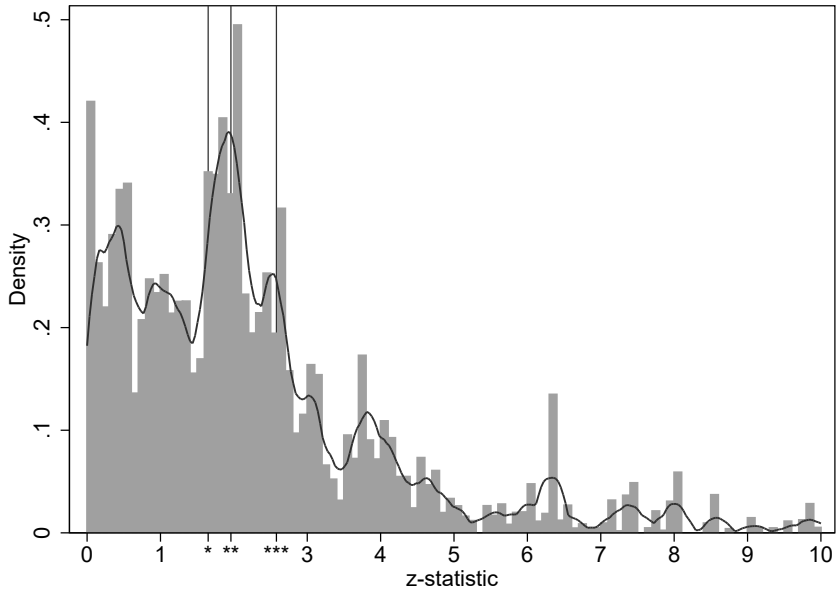
(b) Multiple authors



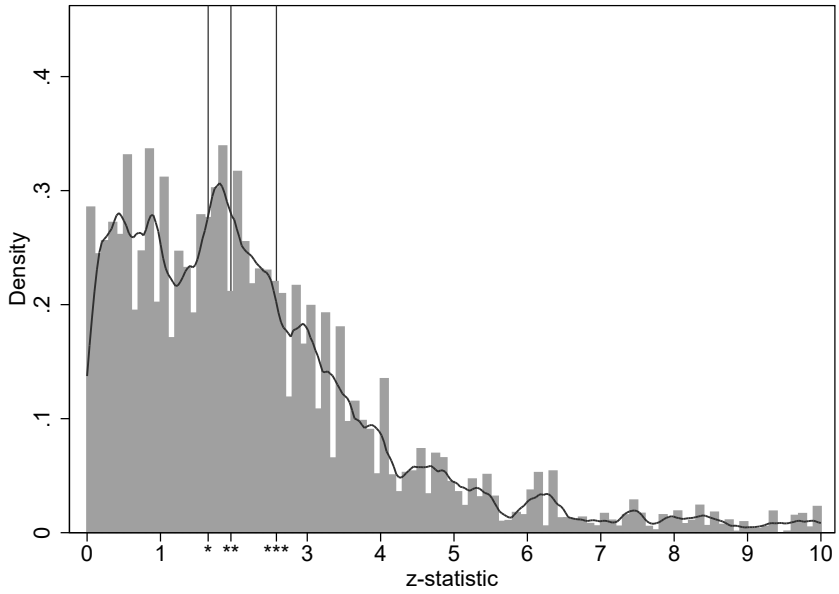
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ for initial submission split by number of coauthors. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure 4: Editor's first decision - Distributions of z-statistics by desk rejection

(a) Desk rejections



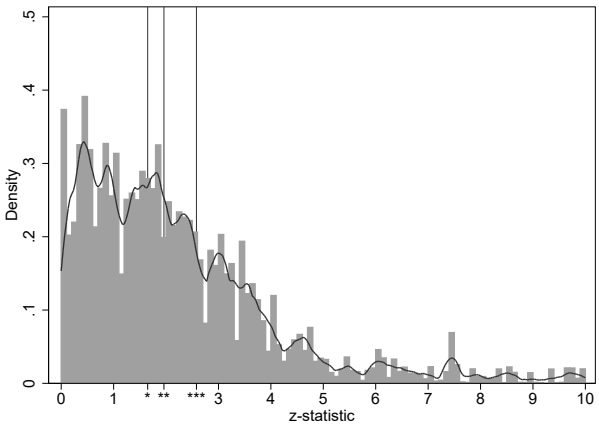
(b) Not desk rejected (received reviewer reports)



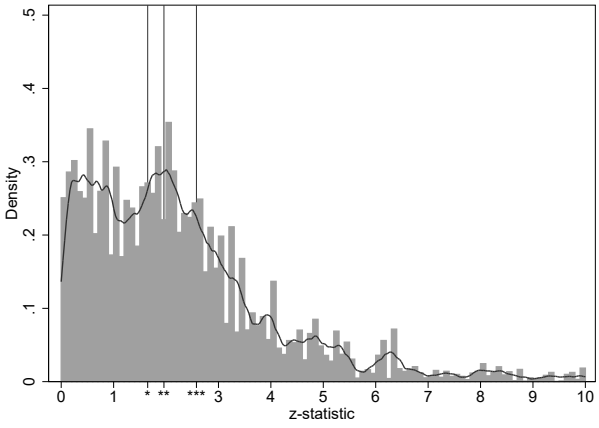
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ by editor's first decision. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure 5: Reviewer stage - Distributions of z-statistics by reviewer recommendation

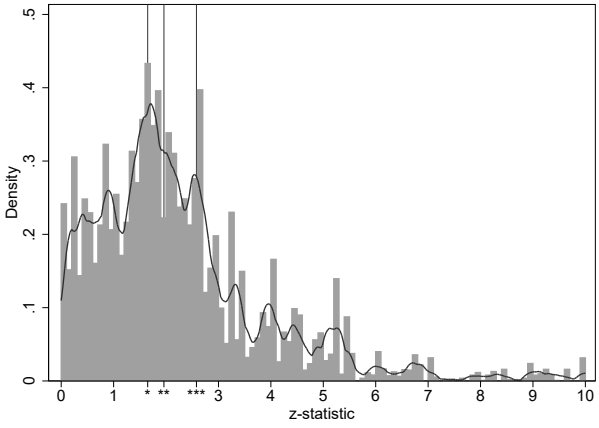
(a) Recommended rejection



(b) Recommend against outright rejection



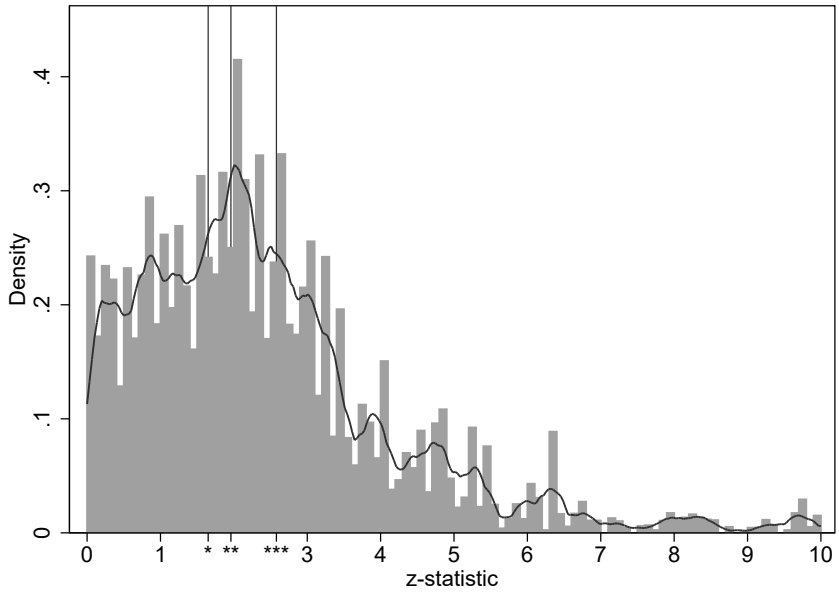
(c) Recommend accept as is or with only minor edits



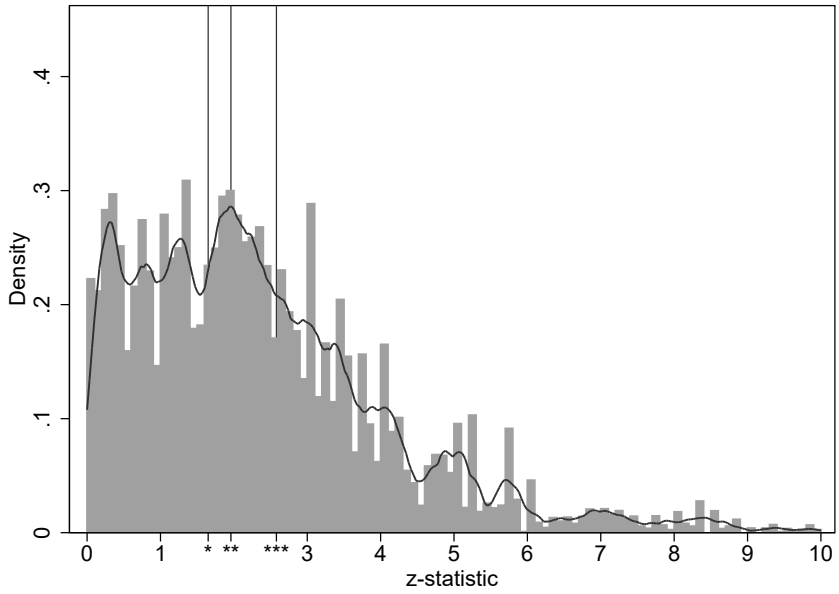
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ for the reviewer stage. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure 6: Distributions of z-statistics by draft versions of accepted manuscripts

(a) First draft (initial submission)



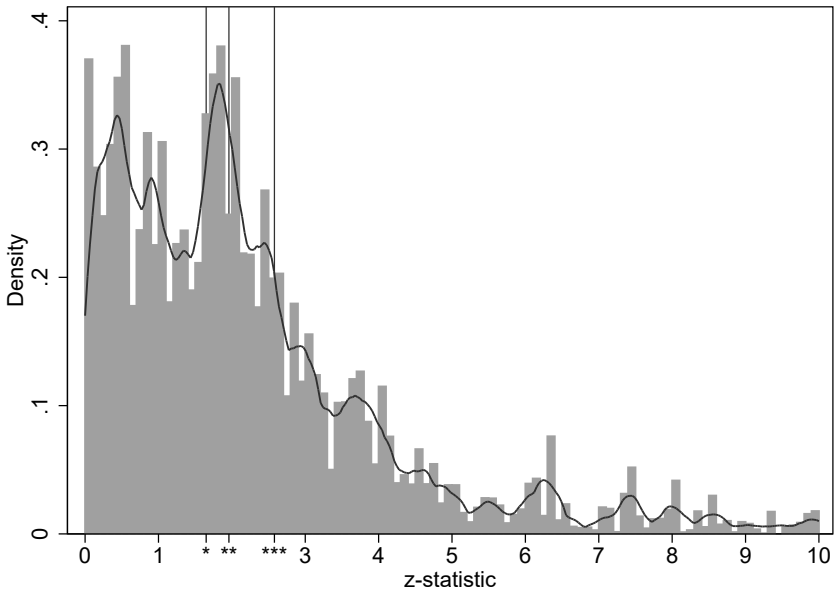
(b) Final draft (published version)



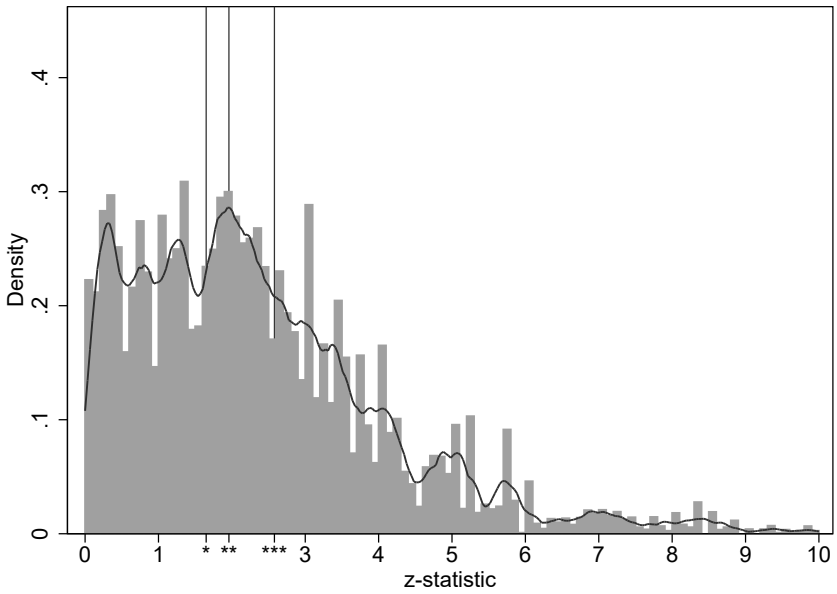
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ for the reviewer stage. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure 7: Peer review - Distributions of z-statistics by rejected and final draft of accepted manuscripts

(a) All rejections



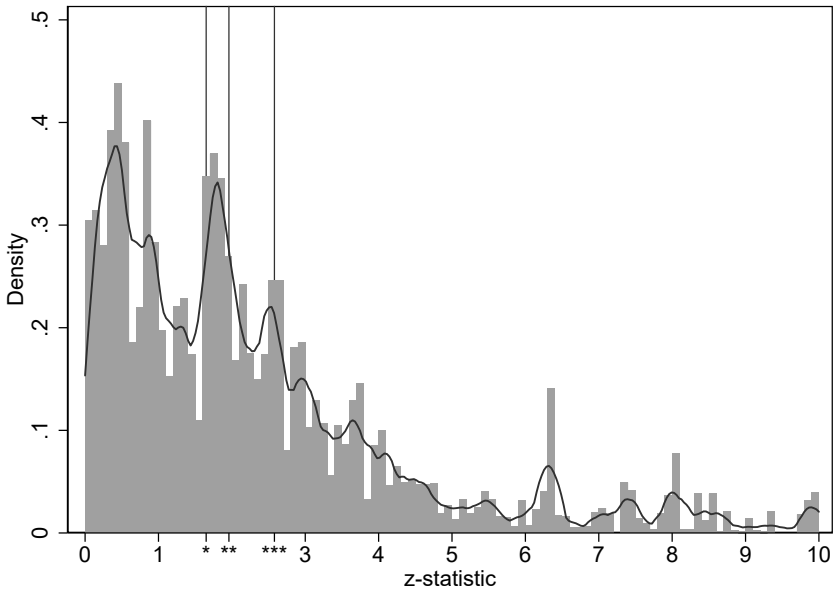
(b) Accepted manuscripts



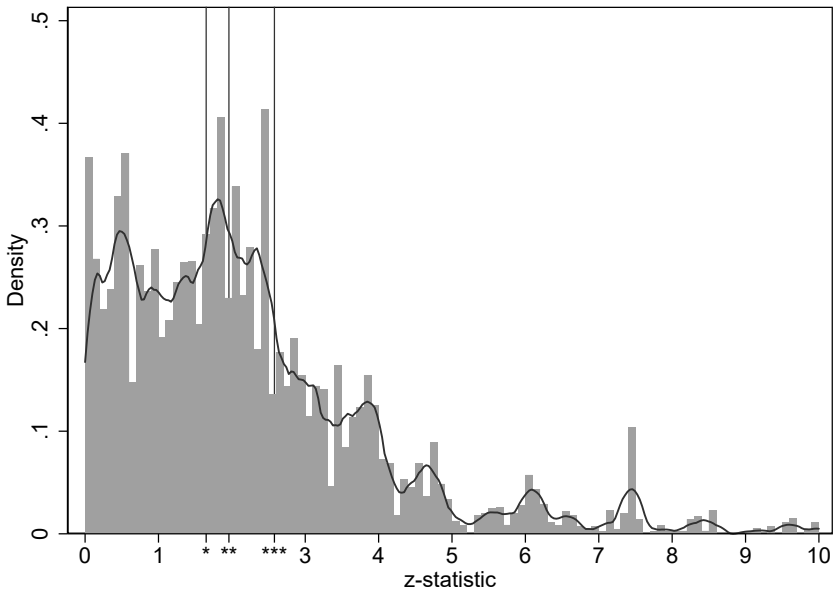
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ for all rejected manuscripts vs the published version. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure 8: After rejection - Distributions of z-statistics by whether the paper eventually published elsewhere

(a) Published elsewhere after rejection



(b) Failed to publish after rejection



Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ for rejected manuscripts that eventually published elsewhere vs. rejected manuscripts that failed to publish anywhere else. Sample restricted to manuscripts that were submitted prior to 2017 to mitigate right-censoring concerns. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Table 1: Summary Statistics at Paper Level

	Desk rejected		Rejected after review		Accepted initial		Accepted final	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Number of test statistics	24.80	31.55	27.94	28.48	30.33	40.30	28.37	34.24
Solo authored	0.41	0.49	0.29	0.46	0.24	0.43	0.24	0.43
Share of authors tenured	0.32	0.36	0.29	0.34	0.38	0.37	0.38	0.37
Share of authors female	0.33	0.39	0.33	0.39	0.37	0.37	0.37	0.37
Share of authors previously publish	0.09	0.23	0.14	0.28	0.45	0.40	0.45	0.40
Author avg. years since PhD	6.73	7.39	7.13	6.95	8.37	6.93	8.37	6.93
Oldest author (years since PhD)	10.04	10.56	10.73	10.78	13.82	12.55	13.82	12.55
Author avg. PhD rank	137.09	100.18	102.58	79.65	77.44	78.01	77.44	78.01
Authors highest PhD rank	98.84	106.83	65.17	79.85	45.06	68.41	45.06	68.41
Identification strategy:								
-Difference-in-differences	0.22	0.42	0.25	0.44	0.28	0.45	0.29	0.45
-Instrumental variables	0.47	0.50	0.29	0.46	0.32	0.47	0.31	0.46
-Regression discontinuity	0.20	0.41	0.29	0.46	0.17	0.38	0.17	0.38
-Randomize control trial	0.10	0.30	0.16	0.37	0.23	0.43	0.23	0.43
Observations	98		110		94		94	

Notes: This table presents summary statistics for our sample at the paper level, split by the four categories for paper outcomes: desk rejected, rejected after receiving reviewer comments, first drafts of accepted manuscripts, and final drafts of accepted manuscripts.

Table 2: Caliper Test, Author Heterogeneity in Initial Submissions

	10% significant	5% significant	1% significant
Solo authored	0.027	0.118*	0.012
	(0.055)	(0.064)	(0.068)
Share tenured	0.136*	0.016	-0.080
	(0.070)	(0.075)	(0.074)
Author avg. years since PhD	-0.014	-0.001	0.014
	(0.010)	(0.011)	(0.013)
(Author avg. years since PhD)^2	0.033	-0.015	-0.070*
	(0.027)	(0.035)	(0.042)
max(Author years since PhD)	0.002	0.007	-0.001
	(0.004)	(0.004)	(0.005)
Share female	0.015	0.001	-0.024
	(0.051)	(0.065)	(0.071)
Share published in JHR	-0.067	0.009	-0.070
	(0.061)	(0.064)	(0.077)
Author avg. PhD rank	-0.000	0.000	-0.000
	(0.000)	(0.001)	(0.000)
Authors highest PhD rank	0.001*	-0.000	0.000
	(0.000)	(0.001)	(0.001)
Identification strategy:			
-Diff-in-diff	0.045	0.112*	0.027
	(0.054)	(0.064)	(0.059)
-IV	0.090*	0.061	0.002
	(0.050)	(0.055)	(0.063)
-RCT	0.055	0.010	-0.050
	(0.062)	(0.062)	(0.091)
Observations	2001	1912	1345
z Sample Bounds	$[1.15, 2.15]$	$[1.46, 2.46]$	$[2.08, 3.08]$

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variables are dummies for whether the test statistics are significant at the 10, 5, and 1 percent levels in columns (1), (2), and (3), respectively. The sample is restricted to initial submissions. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table 3: Desk Rejection: Caliper Test, Significant at the 10% and 5% Levels

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 10% Significant						
Desk Rejected	0.085**	0.102**	0.092*	0.088*	0.083	0.051
	(0.042)	(0.051)	(0.051)	(0.049)	(0.070)	(0.066)
Observations	1952	1952	1952	1952	985	985
z Sample Bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.40, 1.90]$	$[1.40, 1.90]$
Co-editor FE		Y	Y	Y	Y	Y
Identification Strategy			Y	Y		Y
Paper-author Controls				Y		Y
Panel B: 5% Significant						
Desk Rejected	0.032	0.118*	0.117*	0.124*	0.104	0.142*
	(0.051)	(0.061)	(0.063)	(0.066)	(0.071)	(0.079)
Observations	1857	1857	1857	1857	952	952
z Sample Bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.71, 2.21]$	$[1.71, 2.21]$
Co-editor FE		Y	Y	Y	Y	Y
Identification Strategy			Y	Y		Y
Paper-author Controls				Y		Y

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable in Panel A (B) is a dummy for whether the test statistic is significant at the 10 (5) percent level. "Paper-author Controls" include indicators for whether the paper is solo authored, the share of the paper's authors who: are female, are tenured, and published previously in the *Journal of Human Resources*, the authors' average years since receiving their PhD (and its square), the number of years since receiving their PhD for the oldest author, the average of the authors' PhD rank, the highest PhD rank among all authors, and indicators for the primary identification strategy used in the paper. The sample is restricted to initial submissions. The variable of interest "Desk Rejected" equals one if the submission was desk rejected. In columns 1–4, we restrict the sample to $z \pm 0.50$. Columns 5 and 6 restrict the sample to $z \pm 0.25$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table 4: Reviewer Rejection: Caliper Test, Significant at the 10% and 5% Levels

	(1)	(2)	(3)	(4)	(5)
Panel A: 10% Significant					
-Weakly Positive Recommendation	0.033	0.044	0.036	0.039	0.039
	(0.039)	(0.040)	(0.039)	(0.037)	(0.032)
-Minor Edits or Accept As Is	0.020	0.044	0.043	0.063	0.077*
	(0.048)	(0.048)	(0.046)	(0.045)	(0.044)
Observations	2748	2748	2748	2748	2748
z Sample Bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$
Co-editor FE		Y	Y	Y	Y
Reviewer Controls			Y	Y	Y
Identification Strategy				Y	Y
Paper-author Controls					Y
Panel B: 5% Significant					
-Weakly Positive Recommendation	0.037	0.057	0.069*	0.065*	0.042
	(0.040)	(0.038)	(0.037)	(0.036)	(0.028)
-Minor Edits or Accept As Is	-0.069	-0.042	-0.033	-0.046	-0.026
	(0.055)	(0.047)	(0.046)	(0.048)	(0.042)
Observations	2658	2658	2658	2658	2658
z Sample Bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$
Co-editor FE		Y	Y	Y	Y
Reviewer Controls			Y	Y	Y
Identification Strategy				Y	Y
Paper-author Controls					Y

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable in Panel A (B) is a dummy for whether the test statistic is significant at the 10 (5) percent level. "Reviewer Controls" include number of years since PhD (and its square), their PhD rank, and indicators for whether the reviewer is female, an NBER affiliate, and whether they previously published in a "top five" economics journal. "Paper-author Controls" include indicators for whether the paper is solo authored, the share of the paper's authors who: are female, are tenured, and published previously in the *Journal of Human Resources*, the authors' average years since receiving their PhD (and its square), the number of years since receiving their PhD for the oldest author, the average of the authors' PhD rank, the highest PhD rank among all authors, and indicators for the primary identification strategy used in the paper. The sample is restricted to manuscripts that received recommendations from reviewers. The variable of interests "Weakly Positive" and "Minor Edits or Accept As Is" equal one if the manuscript was given a weakly positive or strong positive review, respectively. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table 5: Initial vs Final (Accepted) Submissions: Caliper Test, Significant at the 10% and 5% Level

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 10% Significant						
Initial Draft	0.019	0.017	0.011	0.017	-0.032	-0.017
	(0.046)	(0.043)	(0.043)	(0.038)	(0.052)	(0.047)
Observations	1445	1445	1445	1445	714	714
z Sample Bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.40, 1.90]$	$[1.40, 1.90]$
Co-editor FE		Y	Y	Y	Y	Y
Identification Strategy			Y	Y		Y
Paper-author Controls				Y		Y
Panel B: 5% Significant						
Initial Draft	0.003	-0.002	-0.003	-0.015	0.032	0.033
	(0.051)	(0.041)	(0.041)	(0.036)	(0.057)	(0.047)
Observations	1397	1397	1397	1397	730	730
z Sample Bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.71, 2.21]$	$[1.71, 2.21]$
Co-editor FE		Y	Y	Y	Y	Y
Identification Strategy			Y	Y		Y
Paper-author Controls				Y		Y

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable in Panel A (B) is a dummy for whether the test statistic is significant at the 10 (5) percent level. "Paper-author Controls" include indicators for whether the paper is solo authored, the share of the paper's authors who: are female, are tenured, and published previously in the *Journal of Human Resources*, the authors' average years since receiving their PhD (and its square), the number of years since receiving their PhD for the oldest author, the average of the authors' PhD rank, the highest PhD rank among all authors, and indicators for the primary identification strategy used in the paper. The sample is restricted to initial and final submissions of accepted manuscripts. The variable of interest "Initial Draft" equals one if the initial submission and zero for the final submission. In columns 1–4, we restrict the sample to $z \pm 0.50$. Columns 5 and 6 restrict the sample to $z \pm 0.25$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table 6: Accepted vs. Rejected Manuscripts: Caliper Test, Significant at the 10% and 5% Level

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 10% Significant						
Accepted Manuscripts	-0.030	-0.015	-0.023	0.026	-0.013	0.039
	(0.043)	(0.045)	(0.044)	(0.049)	(0.054)	(0.056)
Observations	1917	1917	1917	1917	957	957
z Sample Bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.40, 1.90]$	$[1.40, 1.90]$
Co-editor FE		Y	Y	Y	Y	Y
Identification Strategy			Y	Y		Y
Paper-author Controls				Y		Y
Panel B: 5% Significant						
Accepted Manuscripts	0.029	0.001	0.003	0.019	-0.021	-0.031
	(0.050)	(0.045)	(0.045)	(0.047)	(0.057)	(0.058)
Observations	1793	1793	1793	1793	905	905
z Sample Bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.71, 2.21]$	$[1.71, 2.21]$
Co-editor FE		Y	Y	Y	Y	Y
Identification Strategy			Y	Y		Y
Paper-author Controls				Y		Y

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. An observation is a test statistic. The dependent variable in Panel A (B) is a dummy for whether the test statistic is significant at the 10 (5) percent level. "Paper-author Controls" include indicators for whether the paper is solo authored, the share of the paper's authors who: are female, are tenured, and published previously in the *Journal of Human Resources*, the authors' average years since receiving their PhD (and its square), the number of years since receiving their PhD for the oldest author, the average of the authors' PhD rank, the highest PhD rank among all authors, and indicators for the primary identification strategy used in the paper. The sample includes all submissions. The variable of interest "Accepted Manuscripts" equals one if the submission was accepted. In columns 1–4, we restrict the sample to $z \pm 0.50$. Columns 5 and 6 restrict the sample to $z \pm 0.25$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table 7: Never Published vs. Published Elsewhere: Caliper Test, Significant at the 10% and 5% Levels

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 10% Significant						
Never Published	-0.083	-0.084	-0.104*	-0.080	-0.118	-0.147**
	(0.059)	(0.060)	(0.060)	(0.072)	(0.072)	(0.072)
Observations	862	862	862	862	434	434
z Sample Bounds	[1.15, 2.15]	[1.15, 2.15]	[1.15, 2.15]	[1.15, 2.15]	[1.40, 1.90]	[1.40, 1.90]
Co-editor FE		Y	Y	Y	Y	Y
Identification Strategy			Y	Y		Y
Paper-author Controls				Y		Y
Panel B: 5% Significant						
Never Published	0.130*	0.140**	0.152***	0.174***	0.125	0.120
	(0.069)	(0.055)	(0.057)	(0.059)	(0.082)	(0.078)
Observations	761	761	761	761	379	379
z Sample Bounds	[1.46, 2.46]	[1.46, 2.46]	[1.46, 2.46]	[1.46, 2.46]	[1.71, 2.21]	[1.71, 2.21]
Co-editor FE		Y	Y	Y	Y	Y
Identification Strategy			Y	Y		Y
Paper-author Controls				Y		Y

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. An observation is a test statistic. The dependent variable in Panel A (B) is a dummy for whether the test statistic is significant at the 10 (5) percent level. "Paper-author Controls" include indicators for whether the paper is solo authored, the share of the paper's authors who: are female, are tenured, and published previously in the *Journal of Human Resources*, the authors' average years since receiving their PhD (and its square), the number of years since receiving their PhD for the oldest author, the average of the authors' PhD rank, the highest PhD rank among all authors, and indicators for the primary identification strategy used in the paper. The sample includes all rejected manuscripts submitted from 2016 to 2016. The variable of interest "Never Published" equals one if the rejected manuscript failed to publish elsewhere. In columns 1–4, we restrict the sample to $z \pm 0.50$. Columns 5 and 6 restrict the sample to $z \pm 0.25$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table 8: Excess Coefficients by Significance Region

	Desk rejected?		Reviewer Report			Accepted		All
	Yes	No	Weak	Strong	Reject	First	Final	Rejects
[0,1.65)								
Observed	0.391	0.404	0.395	0.362	0.420	0.356	0.385	0.419
Expected	0.586	0.490	0.482	0.574	0.525	0.555	0.628	0.467
Difference	-0.195	-0.086	-0.087	-0.212	-0.106	-0.199	-0.241	-0.048
Ratio of Excess to Expected	-0.332	-0.176	-0.180	-0.369	-0.200	-0.359	-0.385	-0.103
[1.65,1.96)								
Observed	0.117	0.089	0.082	0.111	0.094	0.080	0.084	0.107
Expected	0.074	0.082	0.082	0.075	0.079	0.077	0.070	0.083
Difference	0.043	0.008	-0.000	0.037	0.015	0.003	0.014	0.024
Ratio of Excess to Expected	0.573	0.094	-0.005	0.487	0.189	0.037	0.205	0.283
[1.96,2.58)								
Observed	0.146	0.144	0.148	0.160	0.135	0.165	0.150	0.135
Expected	0.105	0.123	0.124	0.108	0.116	0.111	0.097	0.127
Difference	0.041	0.021	0.024	0.052	0.018	0.054	0.053	0.008
Ratio of Excess to Expected	0.389	0.170	0.189	0.486	0.153	0.481	0.549	0.062
[2.58,5)								
Observed	0.216	0.242	0.250	0.231	0.240	0.284	0.282	0.210
Expected	0.157	0.199	0.203	0.162	0.183	0.170	0.140	0.210
Difference	0.059	0.043	0.048	0.069	0.057	0.113	0.143	0.000
Ratio of Excess to Expected	0.376	0.214	0.235	0.423	0.310	0.664	1.019	0.000
[5,∞)								
Observed	0.129	0.121	0.125	0.135	0.111	0.116	0.099	0.130
Expected	0.078	0.106	0.108	0.081	0.094	0.086	0.068	0.113
Difference	0.059	0.043	0.017	0.016	0.016	0.029	0.031	0.016
Ratio of Excess to Expected	0.668	0.143	0.158	0.673	0.173	0.341	0.460	0.145
Degrees of Freedom	2	2	2	2	2	2	2	2
Non-centrality Parameter	0.92	1.31	1.34	0.97	1.17	1.05	0.74	1.4

Notes: Each panel of the table is a separate significance region. In each panel and for each method, we report four statistics: 1) The observed mass of test statistics 2) The expected mass informed by a calibrated t distribution 3) The difference and 4) The ratio of the observed to expected. For the difference and ratio, a negative value implies 'missing' test statistics in the region whereas a positive number implies an excess of test statistics. The degrees of freedom and the noncentrality parameter for the t distributions that fit the observed data best are presented at the bottom.

Table 9: Results From Survey with Applied Microeconomists

	Full Sample	Pub/Labor/ Ed/Health	Submitted to: JHR	AER	AEJ:AE	JoLE	JPubE	Labour
Within the past five years, have you ever stopped a research study and/or refrained from submitting a paper to a journal after finding null results?								
- Stopped	0.30	0.39	0.36	0.34	0.35	0.39	0.31	0.41
- Refrained	0.27	0.33	0.34	0.30	0.30	0.25	0.34	0.41
Within the past five years, for a given submission have you ever:								
- Reported only a subset of the dependent variables explored during analysis	0.46	0.53	0.62	0.52	0.56	0.61	0.55	0.71
- Reported only a subset of the analyses or experiments that were conducted	0.45	0.54	0.60	0.51	0.50	0.50	0.55	0.65
- Modified your original hypothesis to better match the empirical results	0.24	0.24	0.36	0.32	0.33	0.43	0.25	0.41
- Excluded or recategorized data after looking at the effect of doing so	0.17	0.20	0.28	0.19	0.20	0.21	0.22	0.24
- Selected regressors after looking at the results	0.24	0.32	0.34	0.26	0.31	0.29	0.28	0.35
- Analyzed data, then decided to expand your sample or conduct more experiments	0.26	0.28	0.34	0.25	0.31	0.25	0.31	0.35
On a 10 point scale (10=Very Important), for studies that are claiming to identify an effect of x on y, how important do you think statistical significance is in influencing the editor's/reviewer's decision, ceteris paribus?	8.06	7.97	7.93	8.04	8.12	7.56	8.19	8.35
	(1.46)	(1.55)	(1.60)	(1.44)	(1.30)	(1.89)	(1.40)	(1.11)
Observations	143	85	47	96	84	28	64	17

Notes: Survey sent to 561 economists who had published a paper with an identification strategy in a top 25 journal in the year 2018. See Table 1 in Brodeur et al. (2020) for the full list of 25 journals. "Submitted to JHR" is the sample of respondents who had submitted to the *Journal of Human Resources* at least once in the prior five years.

Appendix 1: Rounding Issue

Rounding Issue and Uniform Distribution

As mentioned in Section 2, we compute z-statistics by taking the ratios of coefficients and standard errors. This lead to an overrepresentation of small integers because of the low precision used in submitted manuscripts. For example, if the coefficient is reported to be 0.020 and the standard error is 0.010, then our reconstructed z-statistic is two, but the true coefficient lies in the interval [0.0195,0.0205] and the true standard error lies in the interval [0.0095,0.0105]. As a robustness check, we follow Brodeur et al. (2016) and Bruns et al. (2019) and independently redraw an estimate and a standard error in these intervals using a uniform distribution. Using these two random numbers, we reconstruct new *derounded* z-statistics.

Derounded Distributions

We show the derounded distribution of tests for our main subsamples in Appendix Figures A14-A18. Overall, using derounded z-statistics smooth potential discontinuities in histograms, but does not change the shape of the distributions. Bunching just passed the 10% level slightly increases, while the extent of bunching around 1.96 slightly decreases.

Appendix 2: Econometric Specifications

Caliper Test

The Caliper test compares the number of test statistics in a narrow range above and below a statistical significance threshold. For instance, for the 5% threshold:

$$R_{-,h} = [1.96 - h, 1.96], R_{+,h} = [1.96, 1.96 + h] \tag{1}$$

for a bandwidth parameter h.

The main advantage of this methodology over other methods is that it allows us to control for the co-editor handling the submission and for articles' and authors' characteristics. These in turn control for potential (1) differences in co-editors' rejection and acceptance rates, and (2) differences in manuscript quality correlated with paper and author characteristics.

We start with the following equation, focusing strictly on initial submissions:

$$Pr(Significant_{ise} = 1) = \Phi(\alpha + \beta_e + X'_{is}\delta + \gamma DeskRejected_{se}) \tag{2}$$

where $Significant_{ise}$ is an indicator variable for whether test *i* in submission *s* reviewed by co-editor *e* is statistically significant at the 10, 5 or 1% level. We rely on logit models throughout and present standard errors clustered at the submission level. We restrict the sample to $z \in [1.46, 2.46]$ for the 5% statistical significance and to $z \in [1.15, 2.15]$ for the 10% threshold. We also check the robustness of our results to a smaller bandwidth. The variable of interest is $DeskRejected_{se}$, which represents the decision made by the co-editor on the manuscript to either desk reject the manuscript or send it out for further review.

We include the term X_{is} in our model. This vector includes dummy variables for how results are reported (i.e., whether a submission reports p-values, standard errors or t statistics), whether the submission is solo-authored, the identification strategy implemented[18] and the following author-level characteristics aggregated to the paper-level: average years since PhD, maximum years since PhD (i.e. experience of oldest co-author), average PhD institutional rank, minimum PhD institution rank (i.e. rank of university for highest ranked author), share of female authors, share of tenured authors, and share of authors who had published in the journal prior to submission. We also include 24 co-editor fixed effects in most models.

Moving to reviewer recommendations, we estimate the following equation:

$$Pr(Significant_{isr} = 1) = \Phi(\alpha + X'_{is}\delta + \gamma_1 WeakR\&R_{sr} + \gamma_2 StrongR\&R_{sr}) \tag{3}$$

where $Significant_{isr}$ and X_{is} behave as previous described. At this journal, reviewers are given five different options for recommendations ranging from outright rejection to publish as is. $WeakR\&R_{sr}$ and $StrongR\&R_{sr}$ are indicators for whether the manuscript s was weakly or strongly positively reviewed by the reviewer r, respectively. More precisely, $StrongR\&R_{sr}$ indicates a review of accepting the manuscript as is or only requesting minor revisions, while $WeakR\&R_{sr}$ indicates both a non-rejection recommendation, but also not a strongly supportive review. Note that we only estimate this equation for papers that received reviews, and we only focus on the first round of review.

Lastly, to estimate the effect of the peer review process on eventually-accepted manuscripts, we estimate

[18] We classify manuscripts based on the method used by the authors. More precisely, we coded manuscripts as using difference-in-differences, instrumental variables, randomized control trials, or regression discontinuity design.

the following equations:

$$Pr(Significant_{ise} = 1) = \Phi(\alpha + \beta_e + X'_{is}\delta + \gamma_1 Initial_{se}) \tag{4}$$

$$Pr(Significant_{ise} = 1) = \Phi(\alpha + \beta_e + X'_{is}\delta + \gamma_2 Accepted_{se}) \tag{5}$$

where in equation (4) we first restrict our sample to accepted manuscripts and their first drafts. $Initial_{se}$ is an indicator for the initial draft of the eventually-accepted manuscript, and γ_1 reflects the increased bunching in marginally significant tests in first drafts relative to final drafts. Then, in equation (5), $Accepted_{se}$ compares accepted manuscripts against all rejected manuscripts (desk rejected or rejected after review) in order to evaluate the overall impact of peer review on the distribution of test statistics (from initial submissions to final publications).

Excess Test Statistics

Our second method determines the amount of excess tests for different regions by comparing the observed distribution of tests to distributions absent of p-hacking or publication bias. The main advantage of this method is that it allows us to evaluate the absolute level of selective reporting for each group of manuscripts whereas the Caliper test method compares bunching across covariates.

As a robustness check, we follow Brodeur et al. (2016) and hypothesize that the underlying distribution of tests follows either a t distribution with 1 degree of freedom or a Cauchy(0,0.5) distribution. The choice of these two counterfactual distributions is based on the fact that the observed distribution of tests for $z > 5$ for our different groups of manuscripts behaves similarly to these two input distributions in this region of statistical significance.

We complement our excess statistics analysis here by comparing the observed distribution of tests for each group of manuscripts to counterfactual distributions. Appendix Figures A19 and A20 illustrate the observed distributions for each possible decision and the two input distributions, i.e., student's t-distribution with one degree of freedom and Cauchy(0,0.5). We first note that both input distributions behave like the observed distributions for $z > 5$. Moreover, the difference between the observed and input distributions is positive between 1.65 and 2.58 and negative from 0 to 1.65 for each group of manuscripts suggesting selective reporting for all our publication outcomes.

Appendix Tables A17 and A18 present our results using the student's t-distribution with one degree of freedom and Cauchy(0,0.5), respectively. We first normalize the area below each observed distribution (and the input distributions) to one in the interval $z = [0, 10]$. We then compute the difference between the observed and input distributions for each group of manuscripts. We partition the interval $z = [0, 10]$ in four intervals: $[0 - 1.65]$, $[1.65 - 1.96]$, $[1.96 - 2.58]$ and $[2.58 - 10]$. Overall, we find evidence of selective reporting for all our groups of manuscripts with negative values for the interval $[0-1.65]$ and positive values for the other intervals.

We focus on the student's t-distribution with one degree of freedom for the interpretation of the results as the estimates for misallocation are smaller in magnitude. We first look at desk-rejected and non-desk-rejected manuscripts. We find that about 28% of desk-rejected tests and 27% of non-desk-rejected tests are 'missing' from the interval $z = [0, 1.65)$. Approximately 6.6% of the surplus of tests for desk rejected manuscripts can be found in the 1.65 in comparison to 3.9% for non-rejected manuscripts. Similarly, we find a greater extent of misallocation in the window $[1.96 - 2.58]$ for desk-rejected than non-desk-rejected manuscripts with a surplus of 9.3% of tests versus 8.2%.

Last, we investigate the extent of misallocation for accepted manuscripts. We find that approximately 30% of tests are 'missing' from the non-significance region. The difference between initial and final submission is very small for the window $[1.65-1.96]$ and slightly larger for initial submission than final submission for the range $[1.96 - 2.58]$.

Randomization Tests

We now turn to a third method used in other studies to document the extent of p-hacking. Randomization tests consider small windows around significance thresholds. For the 5% threshold:

$$R_{-,h} = [1.96 - h, 1.96], R_{+,h} = [1.96, 1.96 + h] \tag{6}$$

for h a bandwidth parameter, and define ratios

$$r_{M,h} = \frac{Pr(Z \in R_{+,h}|M)}{Pr(Z \in R_{-,h}|M)}. \tag{7}$$

As $h \to 0$, this ratio recovers the (proportional) discontinuity in f_Z^M at the threshold,

$$\lim_{h \to 0} r_{M,h} = \lim_{h \to 0} \frac{f_Z(1.96 + h)}{f_Z(1.96 - h)}. \tag{8}$$

These values are comparable across reviewing stages, since they are unaffected by the distribution of true effects Θ.

In the analysis, we consider small bandwidths h and test if the observed test statistics are binomial-distributed around a threshold with equal probability (Andrews and Kasy (2019)). We conduct this analysis for different sets of manuscripts such as desk-rejected manuscripts. This methodology is akin to Bugni and Canay (forthcoming) who applies a similar methodology to check for jumps in the density in regression discontinuity settings.

Results from Randomization Test

We first compare manuscripts that were desk-rejected to those that were not desk-rejected. We keep only the initial submissions for manuscripts not desk-rejected. The results are reported in Appendix Tables A19 and A20 for the 10 and 5% levels.[19] We report whether desk-rejected and non-desk-rejected manuscripts are statistically differently distributed around significance thresholds for many windows. In Panel A, we rely on a window of half-width $h = 0.5$, while the other panels have the following windows of half-width: $h = 0.25$ and $h = 0.1$. We report one sided p-values.

In Panel A, for the 10% threshold, we find that 65% of tests are statistically significant. In comparison, 56% of test statistics can be found in the same region for non-desk-rejected tests. We test whether each subsample is equally likely to be significant and nonsignificant, i.e. is the random variable $z_{method} \sim Binomial(p = 0.5)$. The probability of observing 65% or greater statistically significant desk-rejected tests is 0.000. Reducing the width of the windows has no effect on this result. Similarly, the probability of observing 56% or greater statistically significant non-desk-rejected tests is 0.000, but reducing the width of the windows increases the p-value with values greater than 0.10 for two windows. These results provide a first piece of evidence that desk-rejected manuscripts and to some extent non-desk-rejected manuscripts have a statistically significant discontinuity in the distribution around the 10% threshold.

For the 5% significance threshold, we find no significant discontinuity for non-desk-rejected manuscripts. For desk-rejected manuscripts, the discontinuity in the distribution is significant only for small windows (i.e.

[19] We find no evidence throughout for selective reporting for the 1% significance threshold. This is consistent with the literature (e.g. Brodeur et al. (2020)) and points to a lack of incentives to engage in specification searching for this threshold.

$h < 0.25$). In the half-width $h = 0.1$ window, about 57% of desk-rejected tests are statistically significant, with a one sided p-value of 0.000.

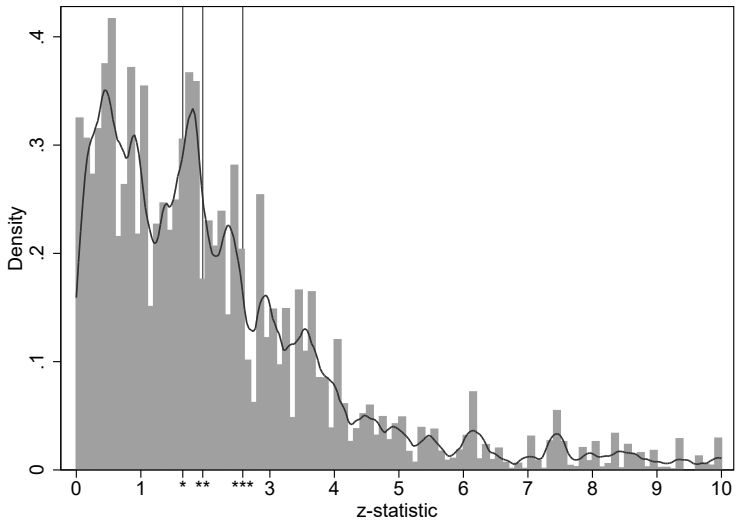
The third to fifth columns in each table shows the proportion for manuscripts that were reviewed by reviewers. For the 10% threshold, we find that 57%, 61% and 55% of tests are statistically significant for the status 'weakly positive', 'strongly positive' and 'reject', respectively. The probability of observing these proportions or greater statistically significant tests is 0.000 for each status. For the 5% significance threshold, we only find evidence of bunching for the status 'weak positive' with 51% of tests being statistically significant (one sided p-value of 0.000).

We now turn our attention to accepted manuscripts. For this analysis, we directly compare the initial and final submission of manuscripts that are ultimately accepted for publication. In Panel A, for the 10% threshold, approximately 56% of tests are statistically significant for both the initial and final versions. There is a small difference for the 5% threshold in Panel A, with 52% and 50% of tests being significant for the initial and final versions, respectively. But this difference is not robust to using smaller windows around this significance threshold.

Overall, these results confirm our conclusions using our other methods.

Appendix 3: Additional Figures and Tables

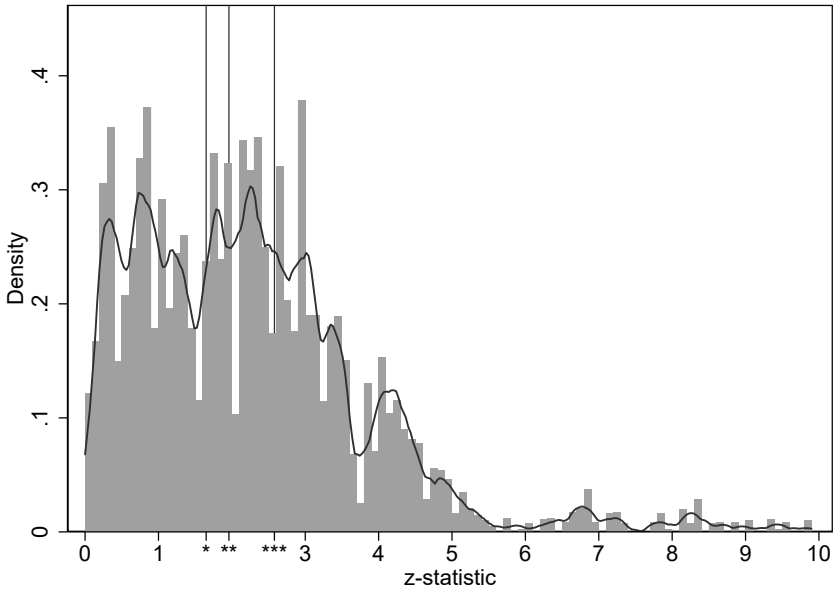
Figure A1: Distribution of z-statistics for submissions rejected but received reviewer reports



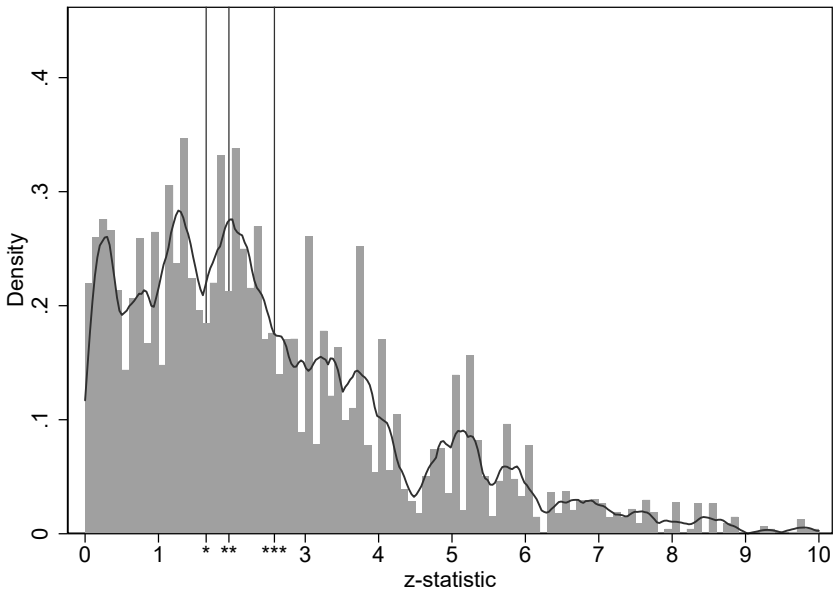
Notes: This figure displays an histogram of test statistics for $z \in [0, 10]$ for manuscripts rejected but that received reviewer reports. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure A2: Distribution of z-statistics from published manuscripts split by number of tenured authors

(a) No tenured authors



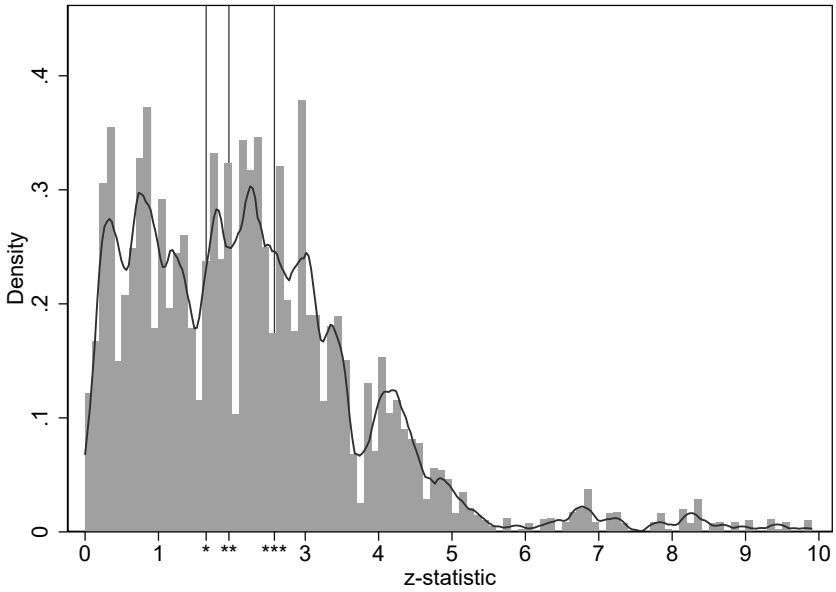
(b) At least one tenured author



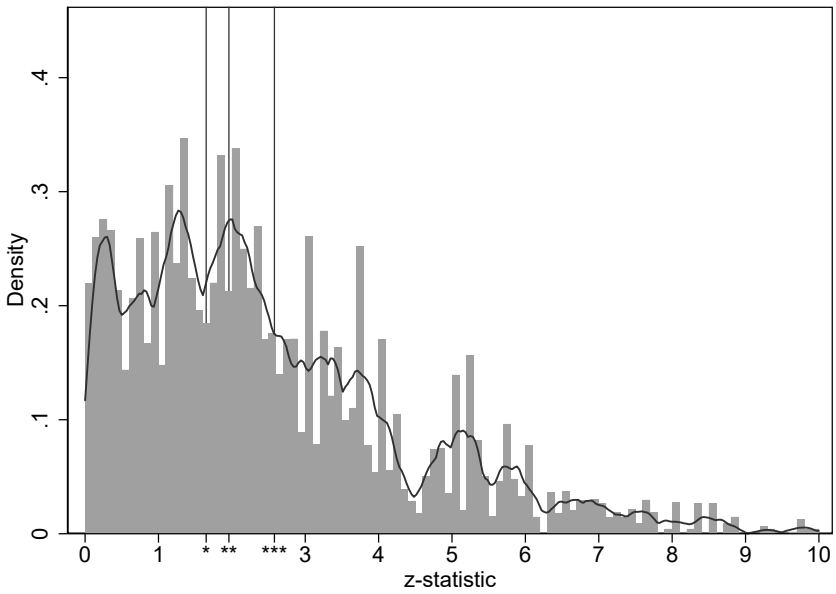
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ for published manuscripts split by the number of tenured authors. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure A3: Distribution of z-statistics from published manuscripts split by authors' years since PhD

(a) All authors within 7 years since PhD



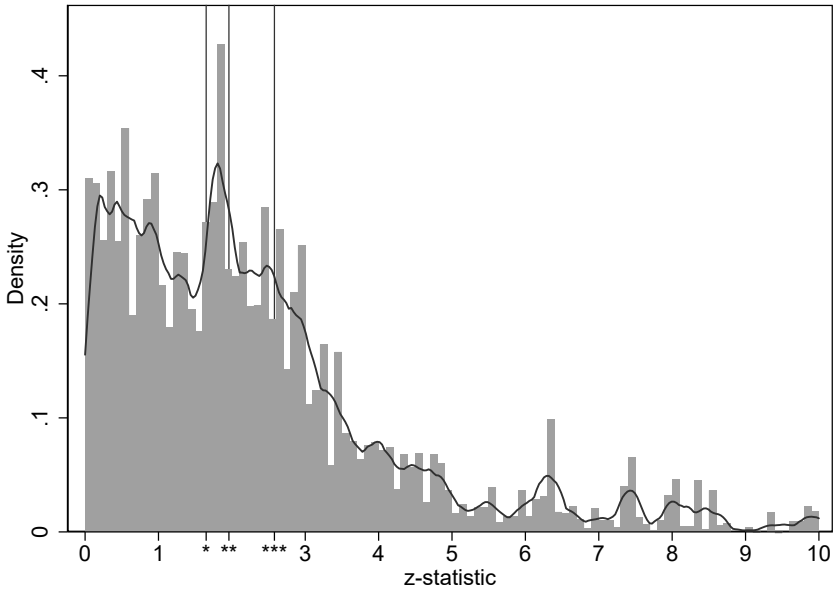
(b) At least one author >7 years since PhD



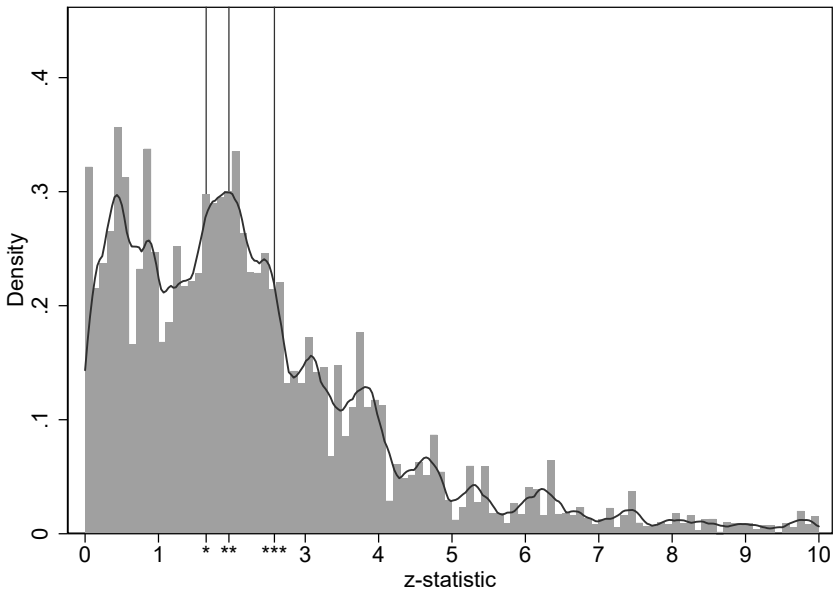
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ for published manuscripts split by authors' years since PhD. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure A4: Distribution of z-statistics from initial submissions split by number of tenured coauthors

(a) No tenured authors



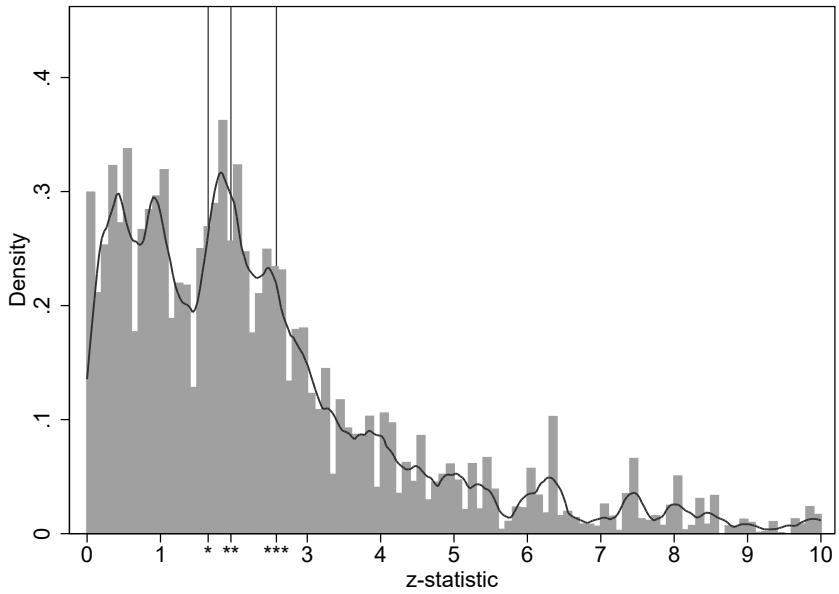
(b) At least one tenured author



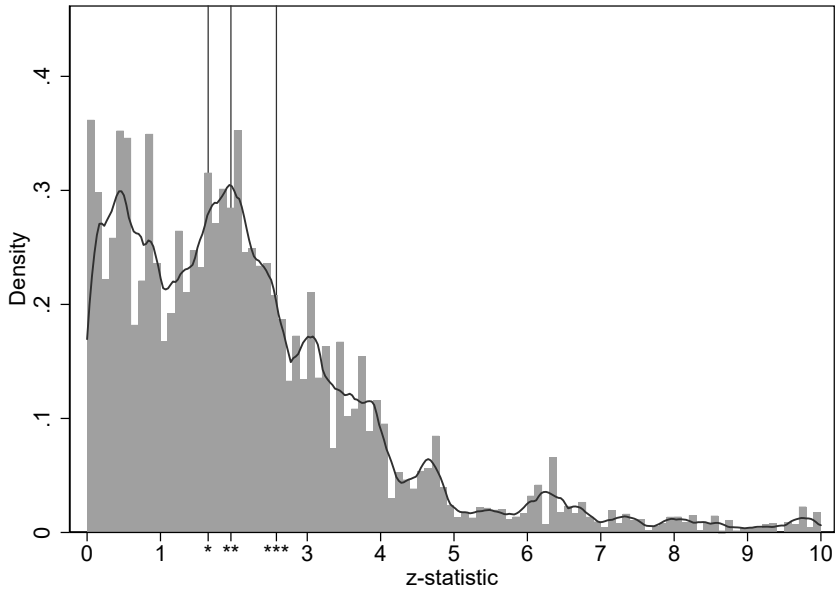
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ for initial submissions split by number of tenured coauthors. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure A5: Distribution of z-statistics from initial submissions split by authors' years since PhD

(a) All authors within seven years of PhD

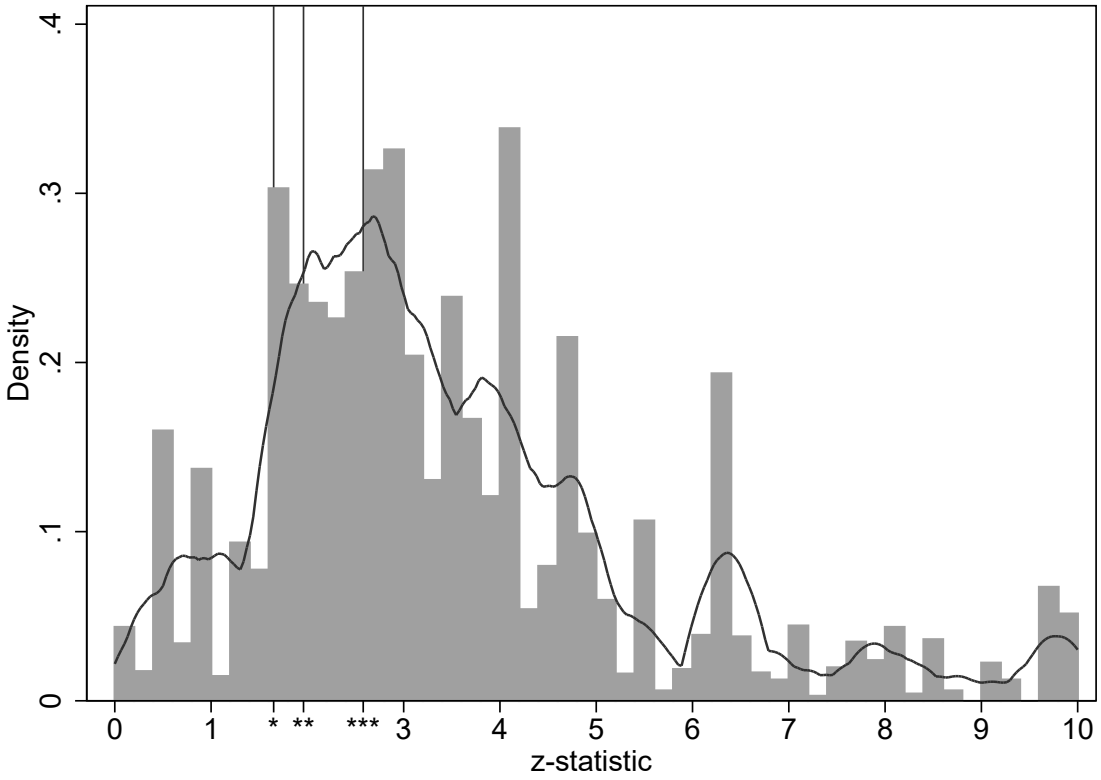


(b) At least one author >7 years since PhD



Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ for initial submissions split by authors' years since PhD. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

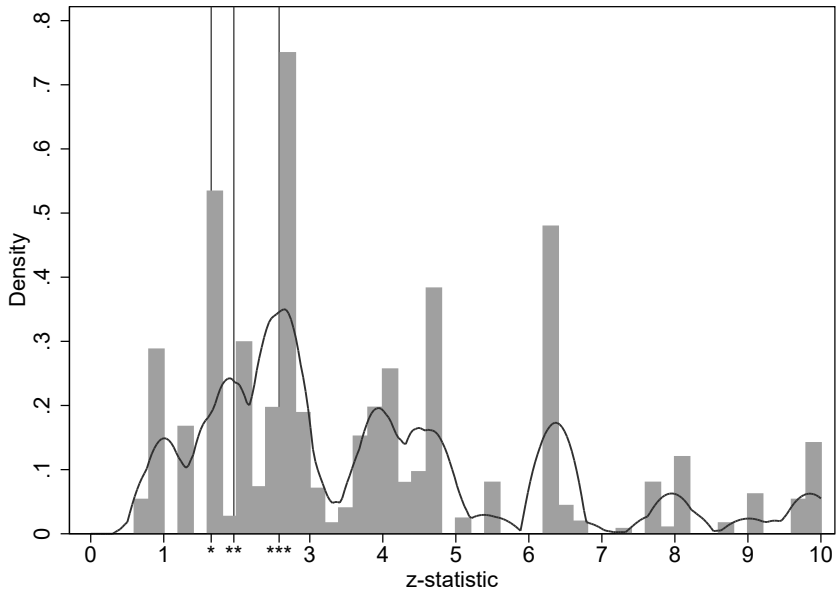
Figure A6: Distribution of z-statistics for initial submissions - largest z-stat from main table



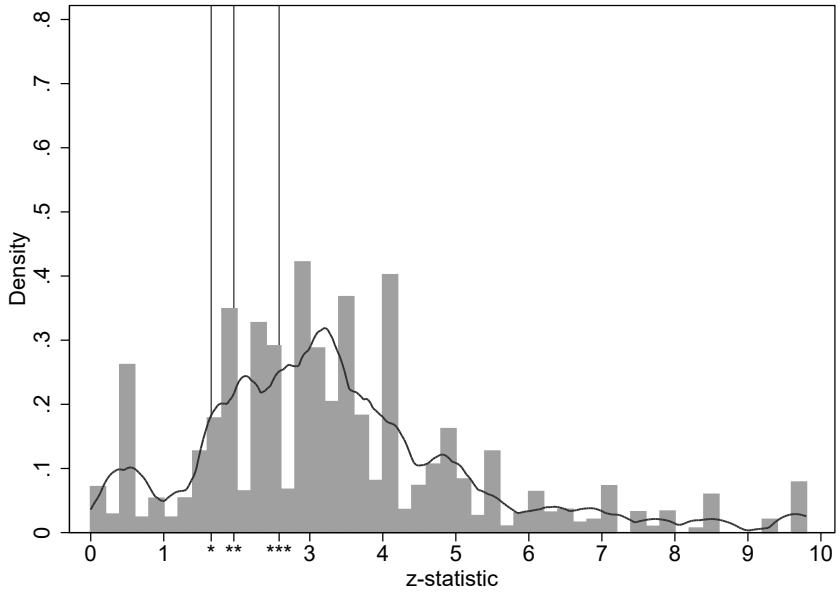
Notes: This figure displays an histogram of test statistics for $z \in [0, 10]$ for initial submission. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure A7: Editor's first decision - Distributions of z-statistics by desk rejection - largest z-stat from main table

(a) Desk rejections



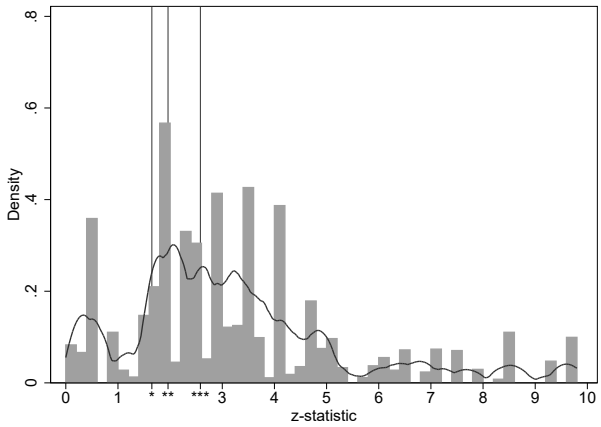
(b) Not desk rejected (received reviewer reports)



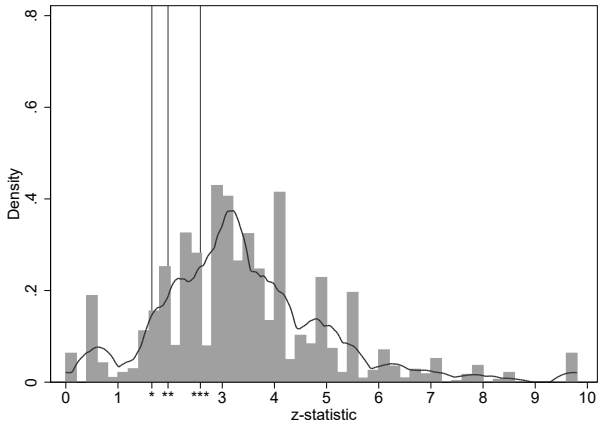
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ by editor's first decision. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure A8: Reviewer stage - Distributions of z-statistics by reviewer recommendation - largest z-stat from main table

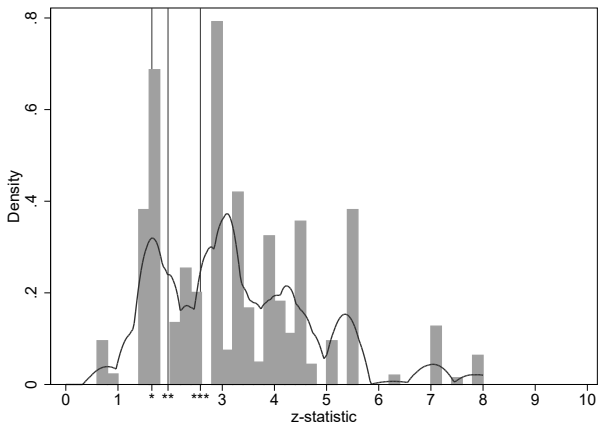
(a) Recommended rejection



(b) Recommend against outright rejection



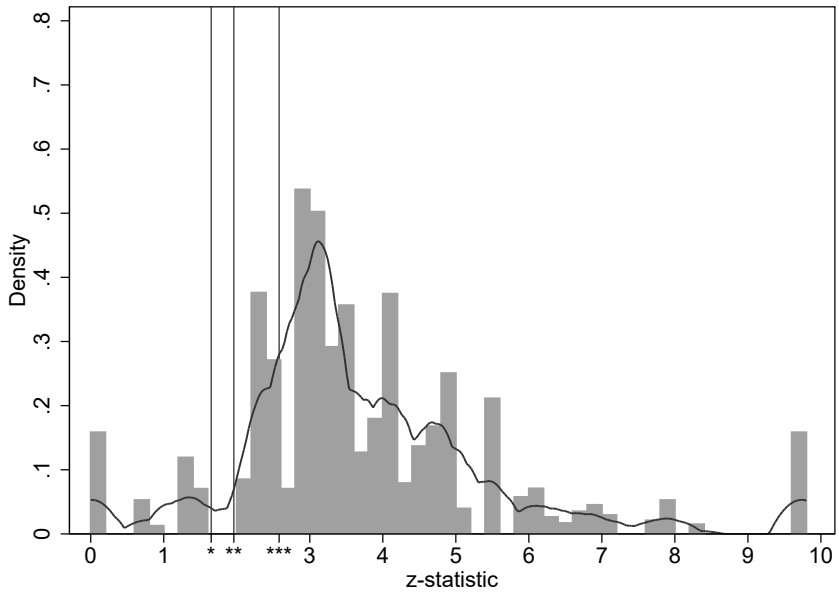
(c) Recommend accept as is or with only minor edits



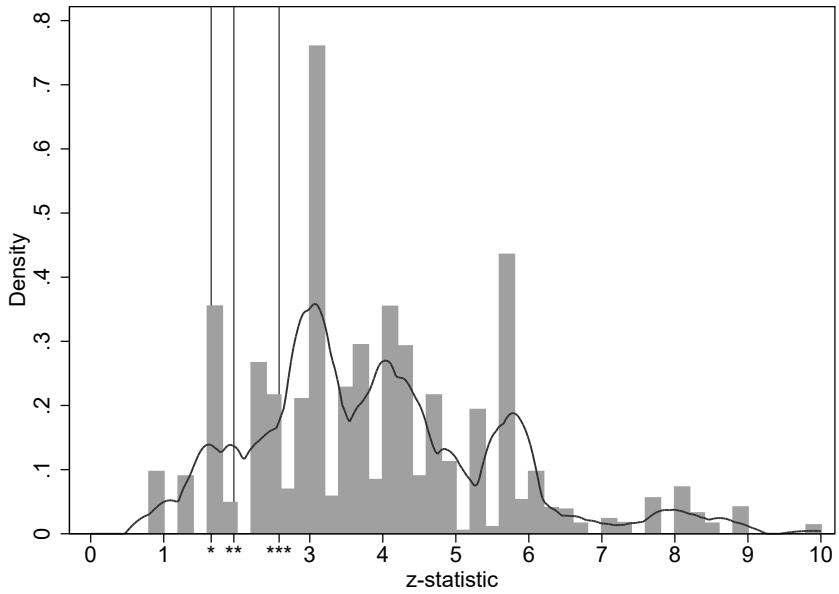
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ for the reviewer stage. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure A9: Distributions of z-statistics by draft versions of accepted manuscripts - largest z-stat from main table

(a) First draft (initial submission)



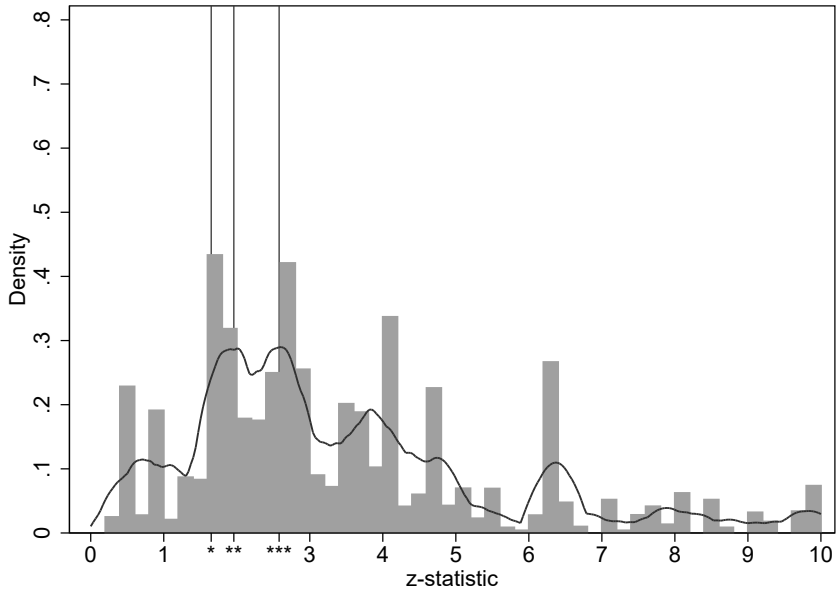
(b) Final draft (published version)



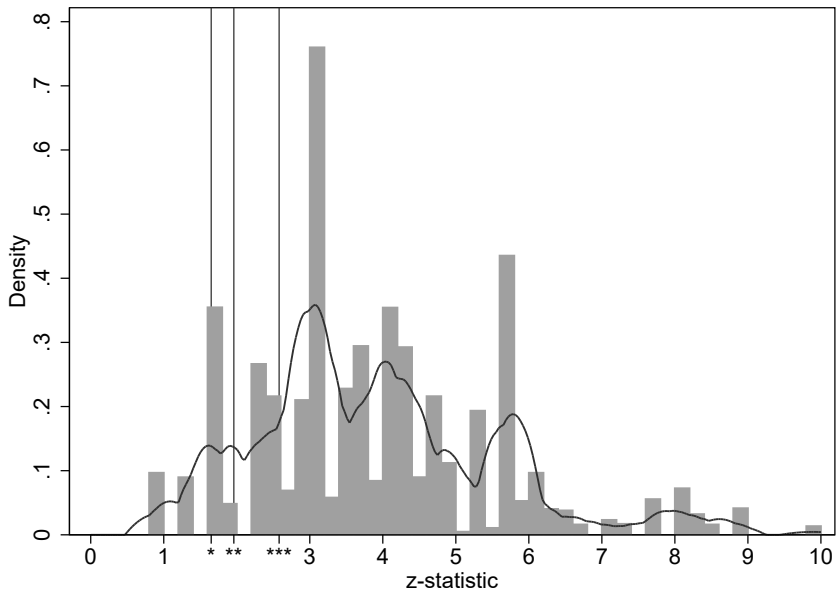
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ for the reviewer stage. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure A10: Peer review - Distributions of z-statistics by rejected and final draft of accepted manuscripts - largest z-stat from main table

(a) All rejections

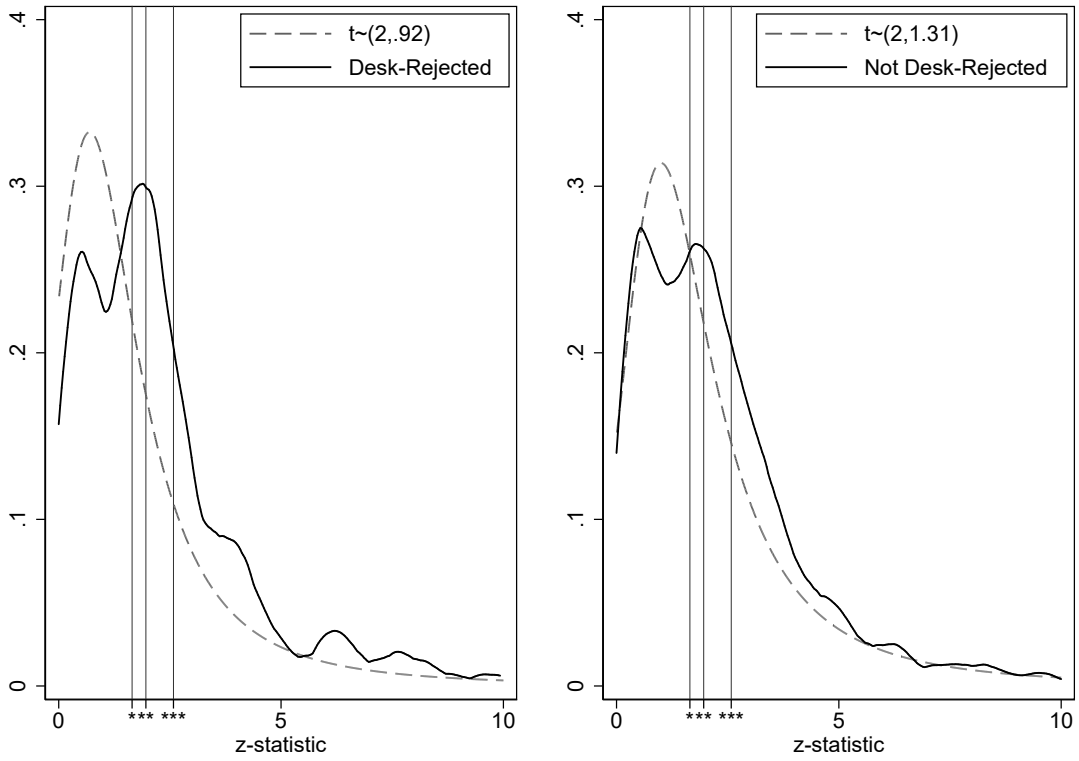


(b) Accepted manuscripts



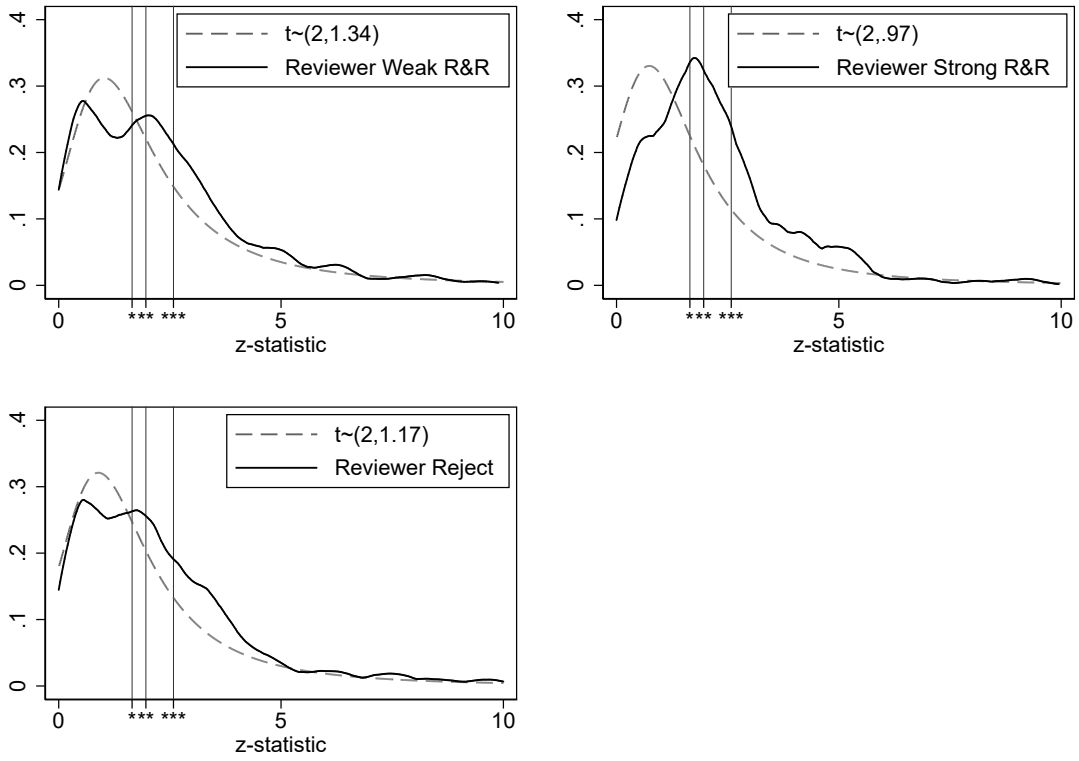
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ for all rejected manuscripts vs the published version. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure A11: Excess Test Statistics by Manuscript Decision: Desk-Rejection Stage



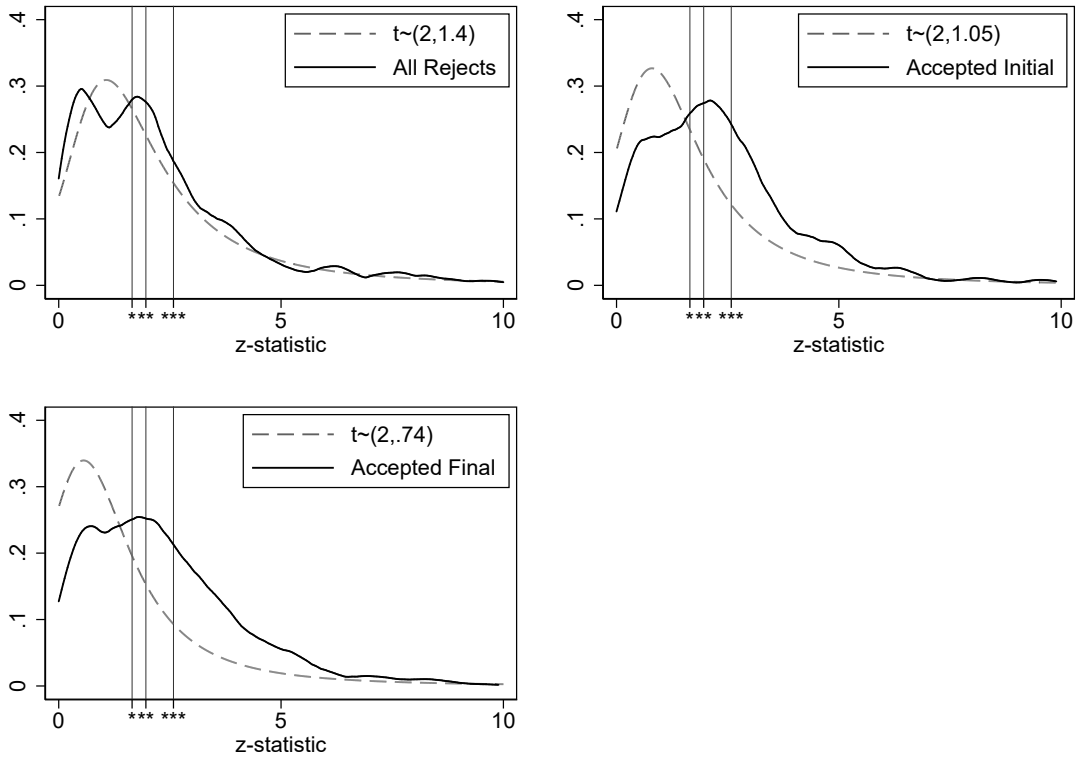
Notes: This figure presents the calibrated input distributions with the observed distributions (weighted using the inverse of the number of tests presented in the same article). We optimize for each category of manuscript at student t distribution with 2 degrees of freedom. The optimal non-centrality parameter is presented for each group of manuscripts.

Figure A12: Excess Test Statistics by Manuscript Decision: Reviewer Stage



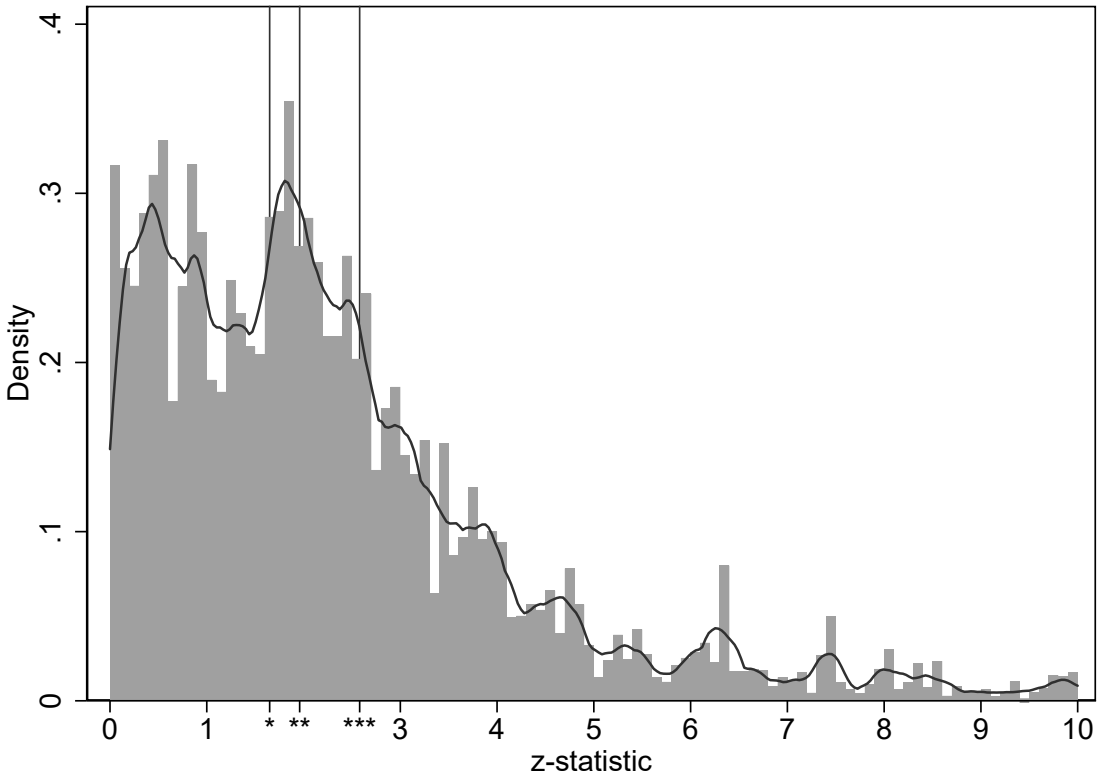
Notes: This figure presents the calibrated input distributions with the observed distributions (weighted using the inverse of the number of tests presented in the same article). We optimize for each category of manuscript at student t distribution with 2 degrees of freedom. The optimal non-centrality parameter is presented for each group of manuscripts.

Figure A13: Excess Test Statistics by Manuscript Decision: Rejection vs Acceptance



Notes: This figure presents the calibrated input distributions with the observed distributions (weighted using the inverse of the number of tests presented in the same article). We optimize for each category of manuscript at student t distribution with 2 degrees of freedom. The optimal non-centrality parameter is presented for each group of manuscripts.

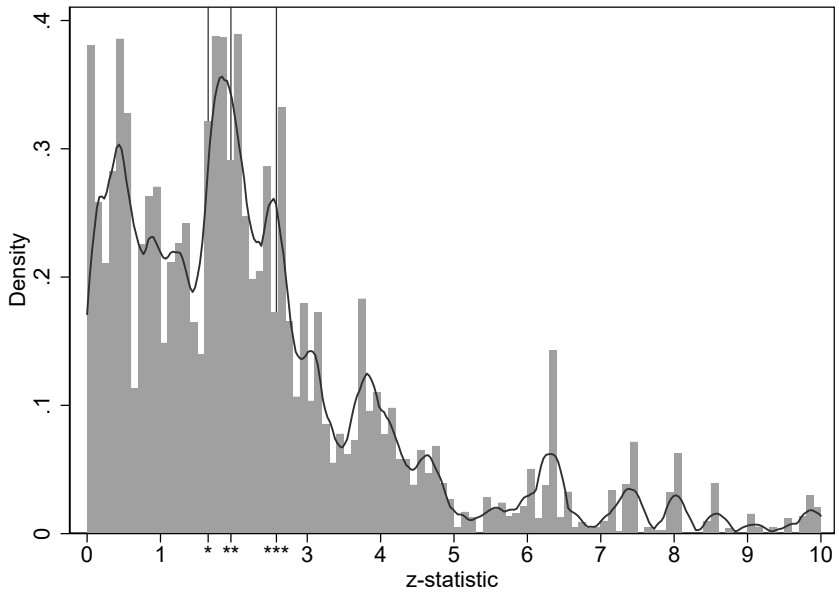
Figure A14: De-rounded distribution of z-statistics for initial submissions



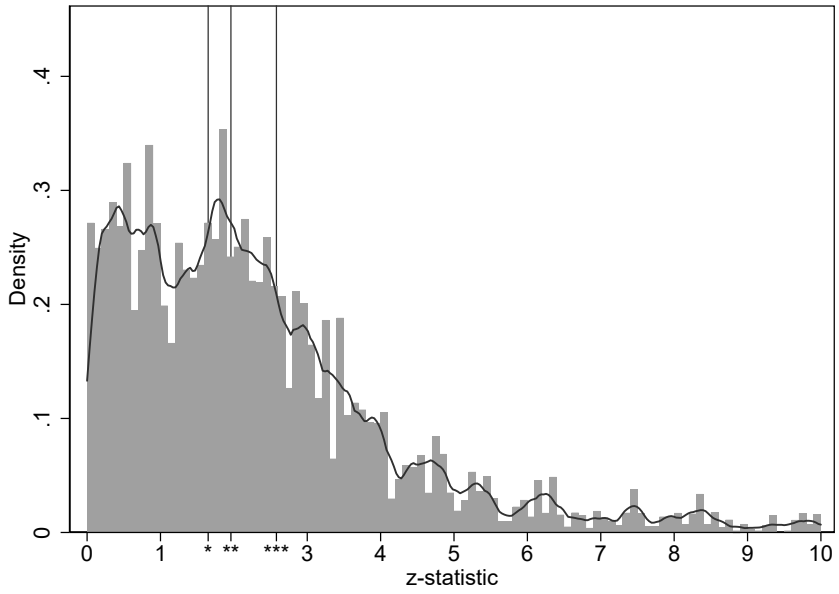
Notes: See Appendix 1 for the de-rounding method. This figure displays an histogram of test statistics for $z \in [0, 10]$ for initial submission. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure A15: Editor's first decision - De-rounded distributions of z-statistics by desk rejection

(a) Desk rejections



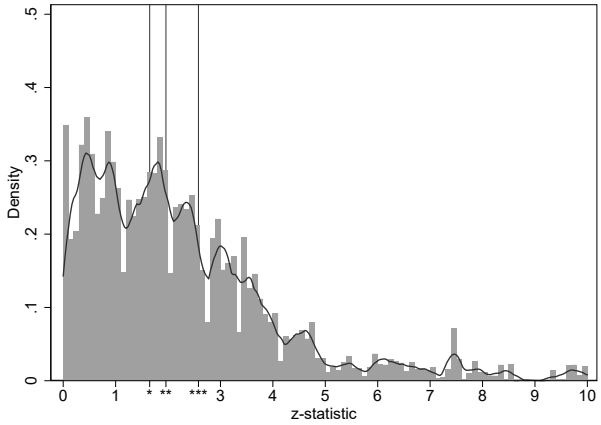
(b) Not desk rejected (received reviewer reports)



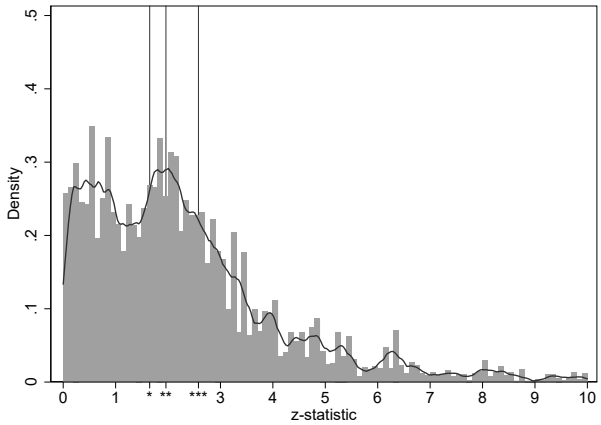
Notes: See Appendix 1 for the de-rounding method. This figure displays histograms of test statistics for $z \in [0, 10]$ by editor's first decision. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure A16: Reviewer stage - De-rounded distributions of z-statistics by reviewer recommendation

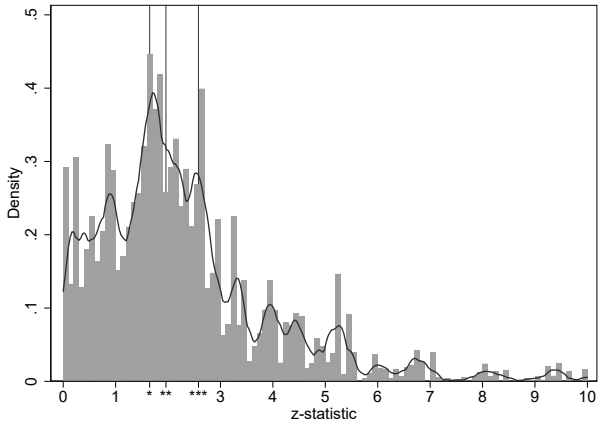
(a) Recommended rejection



(b) Recommend against outright rejection



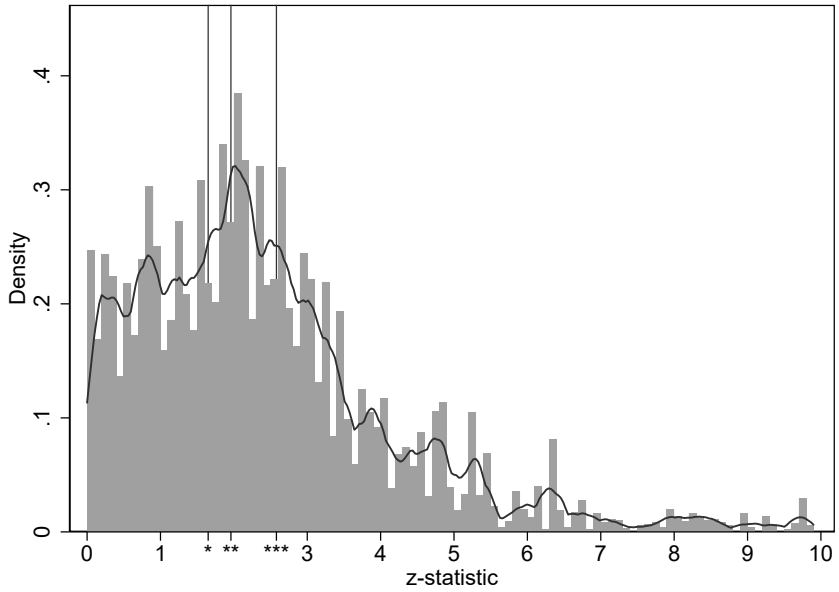
(c) Recommend accept as is or with only minor edits



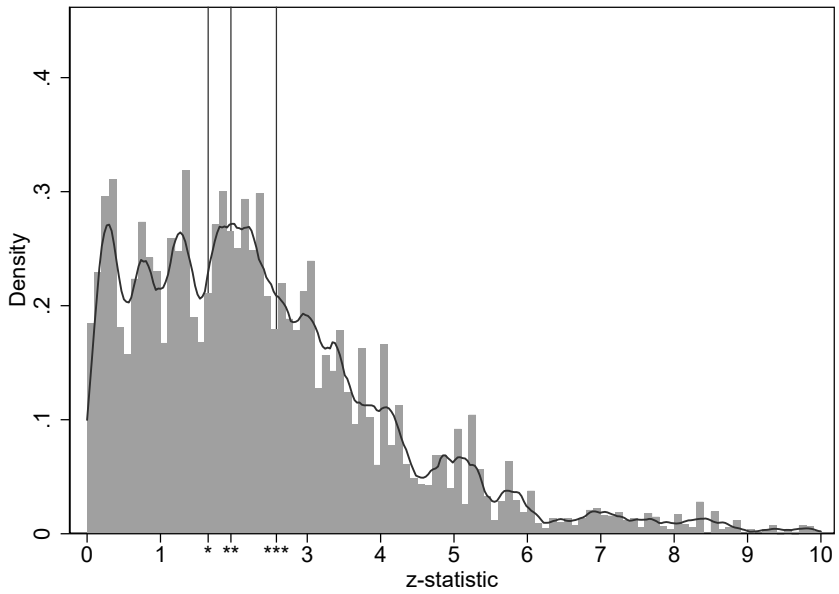
Notes: See Appendix 1 for the de-rounding method. This figure displays histograms of test statistics for $z \in [0, 10]$ for the reviewer stage. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure A17: De-rounded distributions of z-statistics by draft versions of accepted manuscripts

(a) First draft (initial submission)



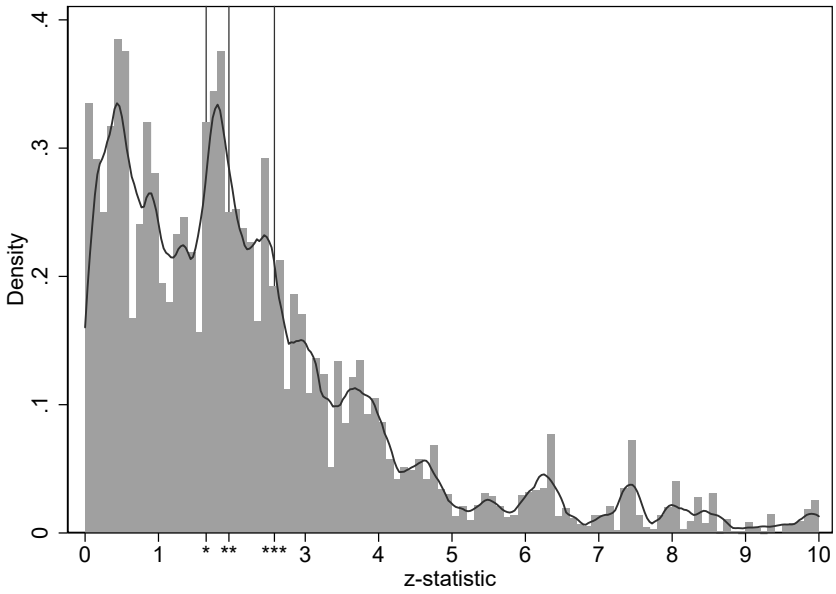
(b) Final draft (published version)



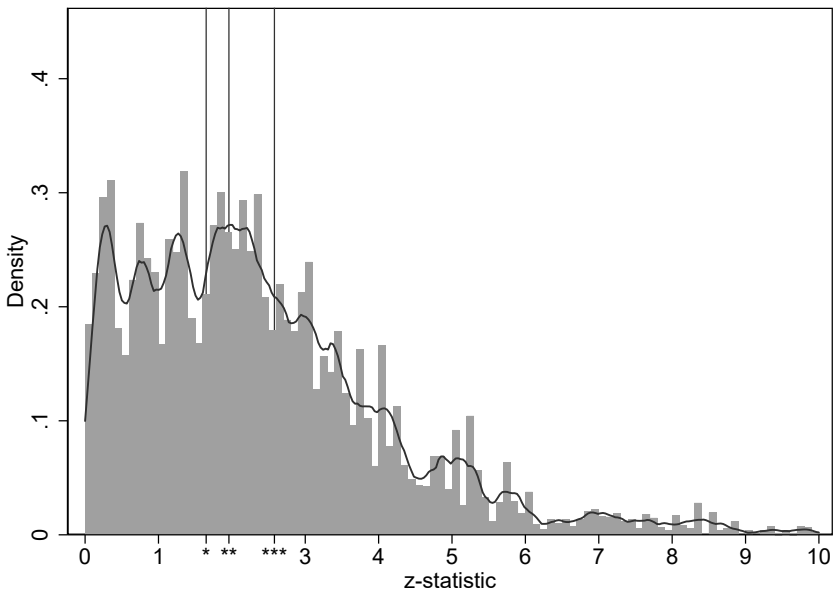
Notes: See Appendix 1 for the de-rounding method. This figure displays histograms of test statistics for $z \in [0, 10]$ for the reviewer stage. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure A18: Peer review - De-rounded distributions of z-statistics by rejected and final draft of accepted manuscripts

(a) All rejections

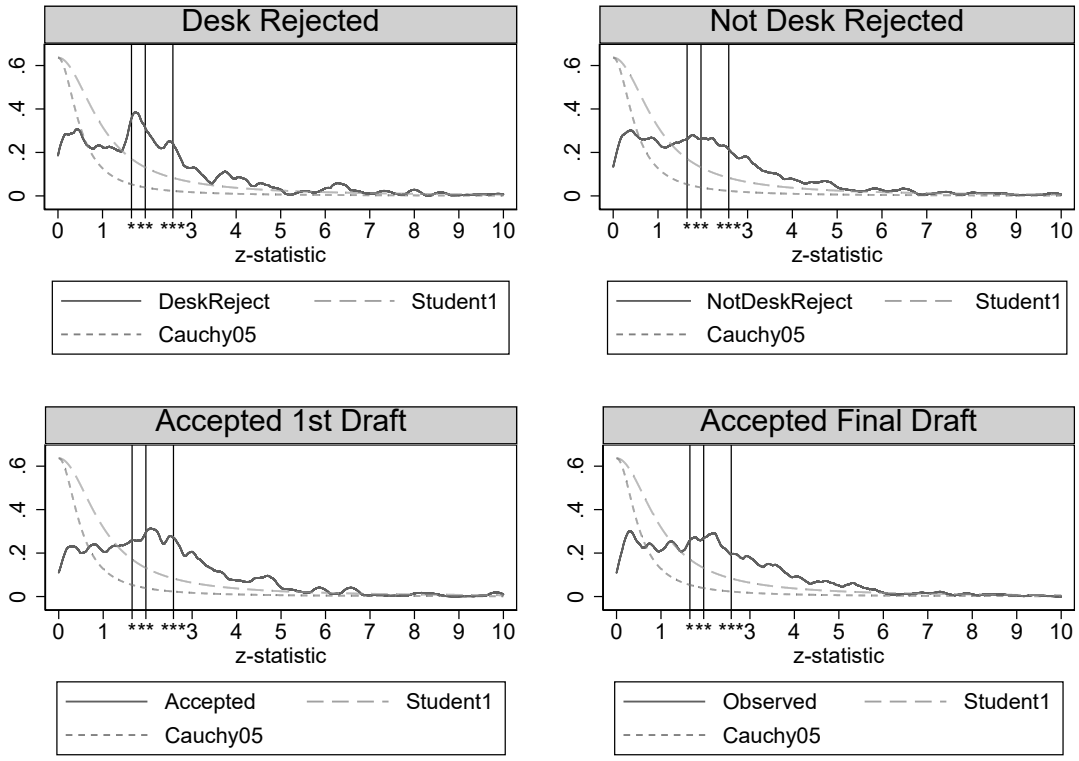


(b) Accepted manuscripts



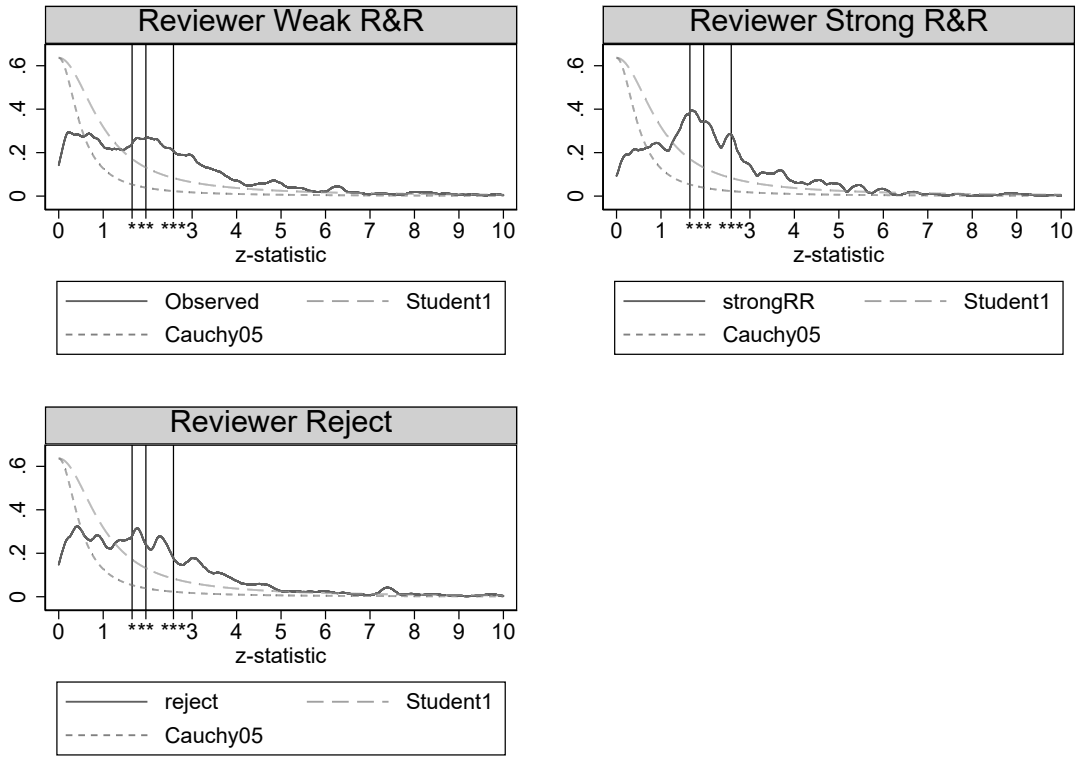
Notes: See Appendix 1 for the de-rounding method. This figure displays histograms of test statistics for $z \in [0, 10]$ for all rejected manuscripts vs the published version. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure A19: Observed Distributions, Student (1) and Cauchy (0,0.5)



Notes: This figure displays the smoothed densities from Figures xx. The Student's t-distribution with one degree of freedom and Cauchy(0,0.5) are used as a reference distribution to detect excess (or missing) tests. Reference lines are displayed at the conventional two-tailed significance levels.

Figure A20: Observed Distributions, Student (1) and Cauchy (0,0.5)



Notes: This figure displays the smoothed densities from Figures xx. The Student's t-distribution with one degree of freedom and Cauchy(0,0.5) are used as a reference distribution to detect excess (or missing) tests. Reference lines are displayed at the conventional two-tailed significance levels.

Table A1: Desk Rejection: Caliper Test, Significant at the 1% Level

	(1)	(2)	(3)	(4)	(5)	(6)
Desk Rejected	-0.028	-0.046	-0.071	-0.125**	-0.049	-0.075
	(0.056)	(0.059)	(0.060)	(0.063)	(0.078)	(0.076)
Observations	1304	1304	1304	1304	648	648
z Sample Bounds	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.33, 2.83]$	$[2.33, 2.83]$
Co-editor FE		Y	Y	Y	Y	Y
Identification Strategy			Y	Y		Y
Paper-author Controls				Y		Y

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable is a dummy for whether the test statistic is significant at the 1 percent level. "Paper-author Controls" include indicators for whether the paper is solo authored, the share of the paper's authors who: are female, are tenured, and published previously in the *Journal of Human Resources*, the authors' average years since receiving their PhD (and its square), the number of years since receiving their PhD for the oldest author, the average of the authors' PhD rank, the highest PhD rank among all authors, and indicators for the primary identification strategy used in the paper. The sample is restricted to initial submissions. The variable of interest "Desk Rejected" equals one if the submission was desk rejected. In columns 1–4, we restrict the sample to $z \pm 0.50$. Columns 5 and 6 restrict the sample to $z \pm 0.25$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table A2: Reviewer Rejection: Caliper Test, Significant at the 1% Level

	(1)	(2)	(3)	(4)	(5)
Panel A: 10% Significant					
-Weakly Positive Recommendation	0.059	0.059	0.050	0.060	0.071**
	(0.053)	(0.042)	(0.041)	(0.038)	(0.035)
-Minor Edits or Accept As Is	0.039	0.001	-0.023	-0.011	-0.007
	(0.074)	(0.057)	(0.056)	(0.051)	(0.050)
Observations	1906	1906	1906	1906	1906
z Sample Bounds	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.08, 3.08]$
Co-editor FE		Y	Y	Y	Y
Reviewer Controls			Y	Y	Y
Identification Strategy				Y	Y
Paper-author Controls					Y

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable is a dummy for whether the test statistic is significant at the 1 percent level. "Reviewer Controls" include number of years since PhD (and its square), their PhD rank, and indicators for whether the reviewer is female, an NBER affiliate, and whether they previously published in a "top five" economics journal. "Paper-author Controls" include indicators for whether the paper is solo authored, the share of the paper's authors who: are female, are tenured, and published previously in the *Journal of Human Resources*, the authors' average years since receiving their PhD (and its square), the number of years since receiving their PhD for the oldest author, the average of the authors' PhD rank, the highest PhD rank among all authors, and indicators for the primary identification strategy used in the paper. The sample is restricted to manuscripts that received recommendations from reviewers. The variable of interests "Weakly Positive" and "Minor Edits or Accept As Is" equal one if the manuscript was given a weakly positive or strong positive review, respectively. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table A3: Initial vs Final (Accepted) Submissions: Caliper Test, Significant at the 1% Level

	(1)	(2)	(3)	(4)	(5)	(6)
Initial Draft	0.014	0.015	0.019	0.029	0.021	0.027
	(0.072)	(0.050)	(0.049)	(0.045)	(0.057)	(0.052)
Observations	1040	1040	1040	1040	482	482
z Sample Bounds	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.08, 3.08]$	$[2.33, 2.83]$	$[2.33, 2.83]$
Co-editor FE		Y	Y	Y	Y	Y
Identification Strategy			Y	Y		Y
Paper-author Controls				Y		Y

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable is a dummy for whether the test statistic is significant at the 1 percent level. "Paper-author Controls" include indicators for whether the paper is solo authored, the share of the paper's authors who: are female, are tenured, and published previously in the *Journal of Human Resources*, the authors' average years since receiving their PhD (and its square), the number of years since receiving their PhD for the oldest author, the average of the authors' PhD rank, the highest PhD rank among all authors, and indicators for the primary identification strategy used in the paper. The sample is restricted to initial submissions. The variable of interest "Initial Draft" equals one if the initial submission and zero for the final submission. In columns 1–4, we restrict the sample to $z \pm 0.50$. Columns 5 and 6 restrict the sample to $z \pm 0.25$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table A4: Accepted vs. Rejected Manuscripts: Caliper Test, Significant at the 1% Level

	(1)	(2)	(3)	(4)	(5)	(6)
Accepted Manuscripts	0.043	0.060	0.067	0.059	0.115**	0.138**
	(0.059)	(0.052)	(0.050)	(0.055)	(0.058)	(0.061)
Observations	1272	1272	1272	1272	634	634
z Sample Bounds	[2.08, 3.08]	[2.08, 3.08]	[2.08, 3.08]	[2.08, 3.08]	[2.33, 2.83]	[2.33, 2.83]
Co-editor FE		Y	Y	Y	Y	Y
Identification Strategy			Y	Y		Y
Paper-author Controls				Y		Y

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. An observation is a test statistic. The dependent variable is a dummy for whether the test statistic is significant at the 1 percent level. "Paper-author Controls" include indicators for whether the paper is solo authored, the share of the paper's authors who: are female, are tenured, and published previously in the *Journal of Human Resources*, the authors' average years since receiving their PhD (and its square), the number of years since receiving their PhD for the oldest author, the average of the authors' PhD rank, the highest PhD rank among all authors, and indicators for the primary identification strategy used in the paper. The sample includes all submissions. The variable of interest "Accepted Manuscripts" equals one if the submission was accepted. In columns 1–4, we restrict the sample to $z \pm 0.50$. Columns 5 and 6 restrict the sample to $z \pm 0.25$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table A5: Never Published vs. Published Elsewhere: Caliper Test, Significant at the 1% Level

	(1)	(2)	(3)	(4)	(5)	(6)
Never Published	0.010	0.041	0.046	0.035	0.105	0.093
	(0.080)	(0.068)	(0.062)	(0.074)	(0.100)	(0.102)
Observations	526	526	526	526	260	260
z Sample Bounds	[2.08, 3.08]	[2.08, 3.08]	[2.08, 3.08]	[2.08, 3.08]	[2.33, 2.83]	[2.33, 2.83]
Co-editor FE		Y	Y	Y	Y	Y
Identification Strategy			Y	Y		Y
Paper-author Controls				Y		Y

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. An observation is a test statistic. The dependent variable is a dummy for whether the test statistic is significant at the 1 percent level. "Paper-author Controls" include indicators for whether the paper is solo authored, the share of the paper's authors who: are female, are tenured, and published previously in the *Journal of Human Resources*, the authors' average years since receiving their PhD (and its square), the number of years since receiving their PhD for the oldest author, the average of the authors' PhD rank, the highest PhD rank among all authors, and indicators for the primary identification strategy used in the paper. The sample includes all rejected manuscripts submitted from 2016 to 2016. The variable of interest "Never Published" equals one if the rejected manuscript failed to publish elsewhere. In columns 1–4, we restrict the sample to $z \pm 0.50$. Columns 5 and 6 restrict the sample to $z \pm 0.25$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table A6: Robustness to Including Ambiguous Estimates - Desk Rejection: Caliper Test, Significant at the 10% and 5% Levels

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 10% Significant						
Desk Rejected	0.082**	0.091*	0.082	0.083*	0.071	0.055
	(0.041)	(0.050)	(0.050)	(0.047)	(0.069)	(0.064)
Observations	2124	2124	2124	2124	1077	1077
z Sample Bounds	[1.15, 2.15]	[1.15, 2.15]	[1.15, 2.15]	[1.15, 2.15]	[1.40, 1.90]	[1.40, 1.90]
Co-editor FE		Y	Y	Y	Y	Y
Identification Strategy			Y	Y		Y
Paper-author Controls				Y		Y
Panel B: 5% Significant						
Desk Rejected	0.029	0.104*	0.103*	0.115*	0.098	0.130*
	(0.049)	(0.060)	(0.061)	(0.064)	(0.069)	(0.077)
Observations	2012	2012	2012	2012	1022	1022
z Sample Bounds	[1.46, 2.46]	[1.46, 2.46]	[1.46, 2.46]	[1.46, 2.46]	[1.71, 2.21]	[1.71, 2.21]
Co-editor FE		Y	Y	Y	Y	Y
Identification Strategy			Y	Y		Y
Paper-author Controls				Y		Y

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable in Panel A (B) is a dummy for whether the test statistic is significant at the 10 (5) percent level. "Paper-author Controls" include indicators for whether the paper is solo authored, the share of the paper's authors who: are female, are tenured, and published previously in the *Journal of Human Resources*, the authors' average years since receiving their PhD (and its square), the number of years since receiving their PhD for the oldest author, the average of the authors' PhD rank, the highest PhD rank among all authors, and indicators for the primary identification strategy used in the paper. The sample is restricted to initial submissions. The variable of interest "Desk Rejected" equals one if the submission was desk rejected. In columns 1–4, we restrict the sample to $z \pm 0.50$. Columns 5 and 6 restrict the sample to $z \pm 0.25$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table A7: Robustness to Including Ambiguous Estimates - Reviewer Rejection: Caliper Test, Significant at the 10% and 5% Levels

	(1)	(2)	(3)	(4)	(5)
Panel A: 10% Significant					
-Weakly Positive Recommendation	0.032	0.044	0.035	0.041	0.044
	(0.038)	(0.039)	(0.038)	(0.037)	(0.032)
-Minor Edits or Accept As Is	0.021	0.043	0.041	0.065	0.079*
	(0.047)	(0.049)	(0.046)	(0.046)	(0.044)
Observations	3108	3108	3108	3108	3108
z Sample Bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$
Co-editor FE		Y	Y	Y	Y
Reviewer Controls			Y	Y	Y
Identification Strategy				Y	Y
Paper-author Controls					Y
Panel B: 5% Significant					
-Weakly Positive Recommendation	0.038	0.055	0.066*	0.062*	0.038
	(0.039)	(0.037)	(0.037)	(0.036)	(0.029)
-Minor Edits or Accept As Is	-0.074	-0.050	-0.040	-0.051	-0.038
	(0.054)	(0.046)	(0.045)	(0.048)	(0.042)
Observations	2977	2977	2977	2977	2977
z Sample Bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$
Co-editor FE		Y	Y	Y	Y
Reviewer Controls			Y	Y	Y
Identification Strategy				Y	Y
Paper-author Controls					Y

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable in Panel A (B) is a dummy for whether the test statistic is significant at the 10 (5) percent level. "Reviewer Controls" include number of years since PhD (and its square), their PhD rank, and indicators for whether the reviewer is female, an NBER affiliate, and whether they previously published in a "top five" economics journal. "Paper-author Controls" include indicators for whether the paper is solo authored, the share of the paper's authors who: are female, are tenured, and published previously in the *Journal of Human Resources*, the authors' average years since receiving their PhD (and its square), the number of years since receiving their PhD for the oldest author, the average of the authors' PhD rank, the highest PhD rank among all authors, and indicators for the primary identification strategy used in the paper. The sample is restricted to manuscripts that received recommendations from reviewers. The variable of interests "Weakly Positive" and "Minor Edits or Accept As Is" equal one if the manuscript was given a weakly positive or strong positive review, respectively. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table A8: Robustness to Including Ambiguous Estimates - Initial vs Final (Accepted) Submissions: Caliper Test, Significant at the 10% and 5% Levels

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 10% Significant						
Initial Draft	0.025	0.023	0.017	0.023	-0.015	0.000
	(0.045)	(0.042)	(0.042)	(0.037)	(0.051)	(0.046)
Observations	1533	1533	1533	1533	771	771
z Sample Bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.40, 1.90]$	$[1.40, 1.90]$
Co-editor FE		Y	Y	Y	Y	Y
Identification Strategy			Y	Y		Y
Paper-author Controls				Y		Y
Panel B: 5% Significant						
Initial Draft	-0.005	-0.010	-0.011	-0.025	0.028	0.029
	(0.050)	(0.041)	(0.041)	(0.037)	(0.057)	(0.046)
Observations	1482	1482	1482	1482	769	769
z Sample Bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.71, 2.21]$	$[1.71, 2.21]$
Co-editor FE		Y	Y	Y	Y	Y
Identification Strategy			Y	Y		Y
Paper-author Controls				Y		Y

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable in Panel A (B) is a dummy for whether the test statistic is significant at the 10 (5) percent level. "Paper-author Controls" include indicators for whether the paper is solo authored, the share of the paper's authors who: are female, are tenured, and published previously in the *Journal of Human Resources*, the authors' average years since receiving their PhD (and its square), the number of years since receiving their PhD for the oldest author, the average of the authors' PhD rank, the highest PhD rank among all authors, and indicators for the primary identification strategy used in the paper. The sample is restricted to initial submissions. The variable of interest "Initial Draft" equals one if the initial submission and zero for the final submission. In columns 1–4, we restrict the sample to $z \pm 0.50$. Columns 5 and 6 restrict the sample to $z \pm 0.25$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table A9: Robustness to Including Ambiguous Estimates - Accepted vs. Rejected Manuscripts: Caliper Test, Significant at the 10% and 5% Levels

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 10% Significant						
Accepted Manuscripts	-0.030	-0.010	-0.016	0.032	-0.002	0.052
	(0.041)	(0.044)	(0.043)	(0.048)	(0.053)	(0.055)
Observations	2058	2058	2058	2058	1025	1025
z Sample Bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.40, 1.90]$	$[1.40, 1.90]$
Co-editor FE		Y	Y	Y	Y	Y
Identification Strategy			Y	Y		Y
Paper-author Controls				Y		Y
Panel B: 5% Significant						
Accepted Manuscripts	0.031	0.001	0.004	0.015	-0.033	-0.033
	(0.049)	(0.045)	(0.045)	(0.048)	(0.057)	(0.057)
Observations	1934	1934	1934	1934	965	965
z Sample Bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.71, 2.21]$	$[1.71, 2.21]$
Co-editor FE		Y	Y	Y	Y	Y
Identification Strategy			Y	Y		Y
Paper-author Controls				Y		Y

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. An observation is a test statistic. The dependent variable in Panel A (B) is a dummy for whether the test statistic is significant at the 10 (5) percent level. "Paper-author Controls" include indicators for whether the paper is solo authored, the share of the paper's authors who: are female, are tenured, and published previously in the *Journal of Human Resources*, the authors' average years since receiving their PhD (and its square), the number of years since receiving their PhD for the oldest author, the average of the authors' PhD rank, the highest PhD rank among all authors, and indicators for the primary identification strategy used in the paper. The sample includes all submissions. The variable of interest "Accepted Manuscripts" equals one if the submission was accepted. In columns 1–4, we restrict the sample to $z \pm 0.50$. Columns 5 and 6 restrict the sample to $z \pm 0.25$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table A10: Robustness to Including Ambiguous Estimates - Never Published vs. Published Elsewhere: Caliper Test, Significant at the 10% and 5% Levels

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 10% Significant						
Never Published	-0.087	-0.088	-0.103*	-0.088	-0.119*	-0.109*
	(0.054)	(0.057)	(0.057)	(0.065)	(0.068)	(0.065)
Observations	987	987	987	987	489	489
z Sample Bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.40, 1.90]$	$[1.40, 1.90]$
Co-editor FE		Y	Y	Y	Y	Y
Identification Strategy			Y	Y		Y
Paper-author Controls				Y		Y
Panel B: 5% Significant						
Never Published	0.132**	0.127**	0.132**	0.167***	0.123	0.128*
	(0.062)	(0.055)	(0.056)	(0.056)	(0.079)	(0.078)
Observations	881	881	881	881	431	431
z Sample Bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.71, 2.21]$	$[1.71, 2.21]$
Co-editor FE		Y	Y	Y	Y	Y
Identification Strategy			Y	Y		Y
Paper-author Controls				Y		Y

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. An observation is a test statistic. The dependent variable in Panel A (B) is a dummy for whether the test statistic is significant at the 10 (5) percent level. "Paper-author Controls" include indicators for whether the paper is solo authored, the share of the paper's authors who: are female, are tenured, and published previously in the *Journal of Human Resources*, the authors' average years since receiving their PhD (and its square), the number of years since receiving their PhD for the oldest author, the average of the authors' PhD rank, the highest PhD rank among all authors, and indicators for the primary identification strategy used in the paper. The sample includes all rejected manuscripts submitted from 2016 to 2016. The variable of interest "Never Published" equals one if the rejected manuscript failed to publish elsewhere. In columns 1–4, we restrict the sample to $z \pm 0.50$. Columns 5 and 6 restrict the sample to $z \pm 0.25$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table A11: Robustness to First Table Subsample - Desk Rejection: Caliper Test, Significant at the 10% and 5% Levels

	(1)	(2)	(3)	(4)
Panel A: 10% Significant				
Desk Rejected	0.066	0.111	0.103	0.079
	(0.057)	(0.068)	(0.067)	(0.067)
Observations	904	904	904	904
z Sample Bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$
Co-editor FE		Y	Y	Y
Identification Strategy			Y	Y
Paper-author Controls				Y
Panel B: 5% Significant				
Desk Rejected	0.016	0.122	0.122	0.136
	(0.075)	(0.081)	(0.083)	(0.084)
Observations	895	895	895	895
z Sample Bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$
Co-editor FE		Y	Y	Y
Identification Strategy			Y	Y
Paper-author Controls				Y

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable in Panel A (B) is a dummy for whether the test statistic is significant at the 10 (5) percent level. "Paper-author Controls" include indicators for whether the paper is solo authored, the share of the paper's authors who: are female, are tenured, and published previously in the *Journal of Human Resources*, the authors' average years since receiving their PhD (and its square), the number of years since receiving their PhD for the oldest author, the average of the authors' PhD rank, the highest PhD rank among all authors, and indicators for the primary identification strategy used in the paper. The sample is restricted to initial submissions. The variable of interest "Desk Rejected" equals one if the submission was desk rejected. In columns 1–4, we restrict the sample to $z \pm 0.50$. Columns 5 and 6 restrict the sample to $z \pm 0.25$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table A12: Robustness to First Table Subsample - Reviewer Rejection: Caliper Test, Significant at the 10% and 5% Levels

	(1)	(2)	(3)	(4)	(5)
Panel A: 10% Significant					
-Weakly Positive Recommendation	0.037	0.042	0.049	0.054	0.049
	(0.054)	(0.048)	(0.047)	(0.046)	(0.041)
-Minor Edits or Accept As Is	0.048	0.051	0.045	0.063	0.093
	(0.064)	(0.061)	(0.063)	(0.063)	(0.059)
Observations	1179	1179	1179	1179	1179
z Sample Bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$
Co-editor FE		Y	Y	Y	Y
Reviewer Controls			Y	Y	Y
Identification Strategy				Y	Y
Paper-author Controls					Y
Panel B: 5% Significant					
-Weakly Positive Recommendation	0.026	0.051	0.065	0.056	0.032
	(0.052)	(0.046)	(0.045)	(0.043)	(0.035)
-Minor Edits or Accept As Is	-0.127	-0.092	-0.071	-0.090	-0.031
	(0.080)	(0.065)	(0.063)	(0.063)	(0.060)
Observations	1229	1229	1229	1229	1229
z Sample Bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$
Co-editor FE		Y	Y	Y	Y
Reviewer Controls			Y	Y	Y
Identification Strategy				Y	Y
Paper-author Controls					Y

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable in Panel A (B) is a dummy for whether the test statistic is significant at the 10 (5) percent level. "Reviewer Controls" include number of years since PhD (and its square), their PhD rank, and indicators for whether the reviewer is female, an NBER affiliate, and whether they previously published in a "top five" economics journal. "Paper-author Controls" include indicators for whether the paper is solo authored, the share of the paper's authors who: are female, are tenured, and published previously in the *Journal of Human Resources*, the authors' average years since receiving their PhD (and its square), the number of years since receiving their PhD for the oldest author, the average of the authors' PhD rank, the highest PhD rank among all authors, and indicators for the primary identification strategy used in the paper. The sample is restricted to manuscripts that received recommendations from reviewers. The variable of interests "Weakly Positive" and "Minor Edits or Accept As Is" equal one if the manuscript was given a weakly positive or strong positive review, respectively. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table A13: Robustness to First Table Subsample - Initial vs Final (Accepted) Submissions: Caliper Test, Significant at the 10% and 5% Levels

	(1)	(2)	(3)	(4)
Panel A: 10% Significant				
Initial Draft	0.049	0.042	0.036	0.043
	(0.062)	(0.056)	(0.056)	(0.048)
Observations	618	618	618	618
z Sample Bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$
Co-editor FE		Y	Y	Y
Identification Strategy			Y	Y
Paper-author Controls				Y
Panel B: 5% Significant				
Initial Draft	-0.012	-0.011	-0.012	-0.024
	(0.077)	(0.060)	(0.058)	(0.051)
Observations	596	596	596	596
z Sample Bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$
Co-editor FE		Y	Y	Y
Identification Strategy			Y	Y
Paper-author Controls				Y

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable in Panel A (B) is a dummy for whether the test statistic is significant at the 10 (5) percent level. "Paper-author Controls" include indicators for whether the paper is solo authored, the share of the paper's authors who: are female, are tenured, and published previously in the *Journal of Human Resources*, the authors' average years since receiving their PhD (and its square), the number of years since receiving their PhD for the oldest author, the average of the authors' PhD rank, the highest PhD rank among all authors, and indicators for the primary identification strategy used in the paper. The sample is restricted to initial submissions. The variable of interest "Initial Draft" equals one if the initial submission and zero for the final submission. In columns 1–4, we restrict the sample to $z \pm 0.50$. Columns 5 and 6 restrict the sample to $z \pm 0.25$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table A14: Robustness to First Table Subsample - Accepted vs. Rejected Manuscripts: Caliper Test, Significant at the 10% and 5% Levels

	(1)	(2)	(3)	(4)
Panel A: 10% Significant				
Accepted Manuscripts	-0.053	-0.058	-0.072	-0.040
	(0.056)	(0.057)	(0.055)	(0.062)
Observations	929	929	929	929
z Sample Bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$
Co-editor FE		Y	Y	Y
Identification Strategy			Y	Y
Paper-author Controls				Y
Panel B: 5% Significant				
Accepted Manuscripts	0.033	0.010	0.012	0.037
	(0.071)	(0.061)	(0.061)	(0.070)
Observations	896	896	896	896
z Sample Bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$
Co-editor FE		Y	Y	Y
Identification Strategy			Y	Y
Paper-author Controls				Y

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. An observation is a test statistic. The dependent variable in Panel A (B) is a dummy for whether the test statistic is significant at the 10 (5) percent level. "Paper-author Controls" include indicators for whether the paper is solo authored, the share of the paper's authors who: are female, are tenured, and published previously in the *Journal of Human Resources*, the authors' average years since receiving their PhD (and its square), the number of years since receiving their PhD for the oldest author, the average of the authors' PhD rank, the highest PhD rank among all authors, and indicators for the primary identification strategy used in the paper. The sample includes all submissions. The variable of interest "Accepted Manuscripts" equals one if the submission was accepted. In columns 1–4, we restrict the sample to $z \pm 0.50$. Columns 5 and 6 restrict the sample to $z \pm 0.25$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table A15: Robustness to First Table Subsample - Never Published vs. Published Elsewhere: Caliper Test, Significant at the 10% and 5% Levels

	(1)	(2)	(3)	(4)
Panel A: 10% Significant				
Never Published	-0.130	-0.117	-0.162**	-0.122
	(0.079)	(0.083)	(0.079)	(0.091)
Observations	443	443	443	443
z Sample Bounds	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$	$[1.15, 2.15]$
Co-editor FE		Y	Y	Y
Identification Strategy			Y	Y
Paper-author Controls				Y
Panel B: 5% Significant				
Never Published	0.150	0.152*	0.161*	0.167**
	(0.099)	(0.082)	(0.089)	(0.085)
Observations	407	407	407	407
z Sample Bounds	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$	$[1.46, 2.46]$
Co-editor FE		Y	Y	Y
Identification Strategy			Y	Y
Paper-author Controls				Y

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. An observation is a test statistic. The dependent variable in Panel A (B) is a dummy for whether the test statistic is significant at the 10 (5) percent level. "Paper-author Controls" include indicators for whether the paper is solo authored, the share of the paper's authors who: are female, are tenured, and published previously in the *Journal of Human Resources*, the authors' average years since receiving their PhD (and its square), the number of years since receiving their PhD for the oldest author, the average of the authors' PhD rank, the highest PhD rank among all authors, and indicators for the primary identification strategy used in the paper. The sample includes all rejected manuscripts submitted from 2016 to 2016. The variable of interest "Never Published" equals one if the rejected manuscript failed to publish elsewhere. In columns 1–4, we restrict the sample to $z \pm 0.50$. Columns 5 and 6 restrict the sample to $z \pm 0.25$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table A16: Heterogeneity in Reviewer Rejection: Caliper Test

	NBER affiliate		Not NBER		Publish in top 5		Not top 5		Junior		Senior	
	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%
Panel B: 5% Significant												
-Weakly Positive Recommendation	0.033	0.018	0.063*	0.014	0.080	0.022	0.007	-0.004	-0.001	0.039	0.108**	0.008
	(0.066)	(0.050)	(0.036)	(0.031)	(0.090)	(0.065)	(0.041)	(0.040)	(0.054)	(0.050)	(0.051)	(0.043)
-Minor Edits or Accept As Is	-0.025	-0.074	0.147**	-0.012	0.426**	0.277**	0.060	-0.050	0.001	-0.020	0.133**	0.057
	(0.082)	(0.081)	(0.066)	(0.049)	(0.181)	(0.132)	(0.062)	(0.063)	(0.055)	(0.059)	(0.061)	(0.064)
Observations	1015	947	1733	1711	474	448	1754	1714	1649	1581	1098	1074
z Sample Bounds	[1.15, 2.15]	[1.46, 2.46]	[1.15, 2.15]	[1.46, 2.46]	[1.15, 2.15]	[1.46, 2.46]	[1.15, 2.15]	[1.46, 2.46]	[1.15, 2.15]	[1.46, 2.46]	[1.15, 2.15]	[1.46, 2.46]
Co-editor FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Reviewer Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Identification Strategy	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Paper-author Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

Notes: This table reports marginal effects from logit regressions. An observation is a test statistic. The dependent variable is either a dummy for whether the test statistic is significant at the 10 or 5 percent level. "Reviewer Controls" include number of years since PhD (and its square), their PhD rank, and indicators for whether the reviewer is female, an NBER affiliate, and whether they previously published in a "top five" economics journal. "Paper-author Controls" include indicators for whether the paper is solo authored, the share of the paper's authors who: are female, are tenured, and published previously in the *Journal of Human Resources*, the authors' average years since receiving their PhD (and its square), the number of years since receiving their PhD for the oldest author, the average of the authors' PhD rank, the highest PhD rank among all authors, and indicators for the primary identification strategy used in the paper. The sample is restricted to manuscripts that received recommendations from reviewers. The variable of interests "Weakly Positive" and "Minor Edits or Accept As Is" equal one if the manuscript was given a weakly positive or strong positive review, respectively. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table A17: Excess Test Statistics - Student(1) Distribution

z-Statistic	Desk rejected?		Reviewer Report			Accepted		All
	Yes	No	Weak	Strong	Reject	First	Final	Rejects
(0-1.65)	-0.297	-0.282	-0.286	-0.322	-0.264	-0.333	-0.307	-0.265
(1.65-1.96)	0.062	0.046	0.035	0.081	0.045	0.040	0.032	0.056
(1.96-2.58)	0.109	0.080	0.081	0.099	0.076	0.097	0.084	0.086
(2.58-10)	0.126	0.156	0.170	0.143	0.143	0.196	0.191	0.123

Notes: This table displays the percentage of misallocated tests in each confidence interval. This table uses a Student(1) distribution and weighted observed distributions.

Table A18: Excess Test Statistics - Cauchy Distribution

z-Statistic	Desk rejected?		Reviewer Report			Accepted		All
	Yes	No	Weak	Strong	Reject	First	Final	Rejects
(0-1.65)	-0.439	-0.424	-0.428	-0.464	-0.406	-0.475	-0.449	-0.407
(1.65-1.96)	0.082	0.066	0.055	0.101	0.066	0.060	0.053	0.076
(1.96-2.58)	0.140	0.111	0.112	0.130	0.107	0.128	0.115	0.117
(2.58-10)	0.217	0.246	0.261	0.233	0.233	0.287	0.281	0.214

Notes: This table displays the percentage of misallocated tests in each confidence interval. This table uses a Cauchy(0.5) distribution and weighted observed distributions.

Table A19: Randomization Tests, 10% Significance Threshold

	Desk rejected?		Reviewer Report			Accepted		All
	Yes	No	Weak	Strong	Reject	First	Final	Rejects
Proportion Significant in 1.65±0.5	0.653	0.564	0.574	0.608	0.550	0.563	0.561	0.610
One Sided p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Proportion Significant in 1.65±0.25	0.678	0.591	0.587	0.615	0.603	0.548	0.596	0.654
One Sided p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Proportion Significant in 1.65±0.1	0.623	0.605	0.559	0.667	0.651	0.469	0.627	0.685
One Sided p-value	0.000	0.495	0.000	0.000	0.000	1.000	0.000	0.000

Notes: In this table we present the results of binomial proportion tests where a success is defined as a statistically significant observation at the threshold level. In the first panel we use observations where $(1.15 < z < 2.15)$. The other panels use observations for smaller windows. We test whether the proportion of tests statistically significant is statistically greater than 0.5. The associated p-values are then reported.

Table A20: Randomization Tests, 5% Significance Threshold

	Desk rejected?		Reviewer Report			Accepted		All
	Yes	No	Weak	Strong	Reject	First	Final	Rejects
Proportion Significant in 1.96±0.5	0.487	0.468	0.507	0.417	0.437	0.522	0.500	0.452
One Sided p-value	1.000	0.505	0.000	1.000	1.000	0.000	0.784	1.000
Proportion Significant in 1.96±0.25	0.500	0.454	0.507	0.434	0.396	0.546	0.462	0.438
One Sided p-value	0.791	0.504	0.000	1.000	1.000	0.000	1.000	1.000
Proportion Significant in 1.96±0.1	0.574	0.410	0.433	0.420	0.368	0.443	0.425	0.489
One Sided p-value	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Notes: In this table we present the results of binomial proportion tests where a success is defined as a statistically significant observation at the threshold level. In the first panel we use observations where $(1.46 < z < 2.46)$. The other panels use observations for smaller windows. We test whether the proportion of tests statistically significant is statistically greater than 0.5. The associated p-values are then reported.