American Economic Review 2023, 113(11): 2974–3002
https://doi.org/10.1257/aer.20210795

Unpacking p-Hacking and Publication Bias[†]

By Abel Brodeur, Scott Carrell, David Figlio, and Lester Lusher*

We use unique data from journal submissions to identify and unpack publication bias and p-hacking. We find initial submissions display significant bunching, suggesting the 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; i.e., marginally significant results are more likely to be desk rejected. Reviewer recommendations, in contrast, are positively associated with statistical significance. Overall, the peer review process has little effect on the distribution of test statistics. Lastly, we track rejected papers and present evidence that the prevalence of publication biases is perhaps not as prominent as feared. (*JEL* A11, A14, C13, L82)

Publication biases and p-hacking are generally perceived to be pervasive issues in academia. Publication bias reflects a potential preference among editors and reviewers for results that display statistical significance. P-hacking generally refers to undesirable actions that authors engage in, knowingly or otherwise, in order to produce "more favorable" *p*-values.[1] Such actions include continuing to collect data, tinkering with econometric specifications, and imposing sample restrictions until certain thresholds of statistical significance are met. The motivations for p-hacking could be driven by the presence of a publication bias. Furthermore, a belief about the existence of a publication bias may encourage authors to shelve a study if initial results are undesired or unpromising. These behaviors may have large consequences, as studies reporting significant effects of a particular program or policy may be more

* Brodeur: Department of Economics, University of Ottawa and Institute for Replication (email: abrodeur@uottawa.ca); Carrell: Department of Economics, University of Texas at Austin, NBER, and IZA (email: scott.carrell@austin.utexas.edu); Figlio: Department of Economics and Warner School of Education, University of Rochester, NBER, and IZA (email: david.figlio@rochester.edu); Lusher: Department of Economics, University of Pittsburgh and IZA (email: lesterlusher@pitt.edu). Isaiah Andrews was the coeditor for this article. 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, Keio University, PUC-Chile, SEA, Shidler College of Business, UC Santa Cruz, and UH Manoa for very useful remarks and encouragements. All errors are our own.

[†] Go to https://doi.org/10.1257/aer.20210795 to visit the article page for additional materials and author disclosure statements.

[1] The exact definition of publication bias and p-hacking has slightly differed across the literature. In this study, for simplicity, we refer to "publication bias" as behaviors that reviewers and editors (i.e., the peer review process) engage in that are skewed in favor of statistical significance, while "p-hacking" refers to behaviors engaged in by authors.

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 discusses potential publication biases and specification searching in economics and other disciplines (Andrews and Kasy 2019; Ashenfelter, Harmon, and Oosterbeek 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, Stanley, and Doucouliagos 2017; Leamer 1983; Lybbert and Buccola 2021; McCloskey 1985; Miguel et al. 2014; Stanley 2005, 2008).[2] To document these phenomena, researchers 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; Gerber and Malhotra 2008a,b; Vivalt 2019). Brodeur, Cook, and Heyes (2020) also collect test statistics from working papers and compare this distribution against their published counterparts, finding no evidence that the journal "revise and resubmit" process mitigates bunching of test statistics.

Since prior studies have almost exclusively 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 smooth, 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[3] and bias-corrected estimators and confidence intervals (e.g., Andrews and Kasy 2019) correct reported results for publication bias and some forms of p-hacking.

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 20,000 test statistics across a random sample of over 700 manuscripts submitted for review to a prominent applied microeconomics journal (*Journal of Human Resources*) from the year 2013 to 2018. The *Journal of Human Resources* (JHR) is largely regarded as a "top field" journal and has been shown to be important for tenure and promotion decisions, even among the top economics departments in the United States (Carrell, Figlio, and Lusher 2022).

[2] 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 2022).

[3] Pre-results review involves the reviewing and acceptance of detailed proposals for research studies prior to results being collected. Consequently, the journal commits to publishing the subsequent paper regardless of the study's results. The few journals adopting pre-results review in economics include the *Journal of Development Economics* and *Experimental Economics*.

We first find that the distribution of test statistics among submitted articles displays a hump around 10 and 5 percent significance thresholds, providing direct evidence that the distribution among published statistics cannot be fully attributed to a publication bias in peer review.[4] We then find that the distribution of desk rejections display greater bunching than those sent for review, suggesting that on average, false positives are filtered out during the initial desk review. We also find that this result is partially explained by author characteristics, which correlate with both desk rejection outcomes and propensity to produce marginally significant estimates.[5] Recommendations from anonymous reviewers, on the other hand, are positively associated with statistical significance: as we move from rejection recommendations to strong positive recommendations, the distributions of test statistics display excess mass around significance thresholds. Finally, we find that the distribution of statistics from the final draft of accepted manuscripts is similar to its initial draft counterpart.[6]

In total, by comparing the final draft of accepted manuscripts against all rejected submissions, we find that the peer review process does not significantly influence the distribution of test statistics; i.e., the issues of p-hacking are not exacerbated (nor attenuated) by the full peer review process. We further track papers after rejection to find that approximately 60 percent 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 less bunching at the 10 percent threshold in favor of greater statistical bunching at the 5 percent threshold (i.e., *greater* significance). This evidence suggests the prevalence of publication biases is perhaps not as prominent as feared, though concern remains about reviewers, as marginal significance is associated with positive recommendations.

Rather, our results suggest that the statistical bunching observed among published manuscripts cannot be (entirely) explained by the peer review process. Instead, they suggest that authors engage in actions that 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 percent of authors have stopped a research study or refrained from submitting a paper after finding null results within the past 5 years. We provide suggestive evidence that this

[4] One caveat worth noting is that papers submitted to the JHR may have already been influenced by editor and reviewer recommendations from prior journal submissions. Thus, the distribution of initial submissions at the JHR may reflect authors adjusting their main estimates to reflect feedback from the peer review process (which itself may display a publication bias).

[5] In general, our study is limited to providing descriptive evidence of editor and reviewer behavior. In particular, it may be that differences in distributions across the peer review process could be driven by both direct editor/reviewer preference for/against statistical significance, and/or papers with marginal significance tend to have other (unobserved) characteristics that editors/reviewers differentially evaluate (e.g., unobservedly "bad" papers may be more likely to contain marginally significant estimates).

[6] Still, given that 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).

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 percent 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 percent), excluding or recategorizing data after seeing the effects of doing so (18 percent), and selecting regressors after looking at the results (26 percent). Finally, nearly 30 percent 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 pre-registering studies and developing pre-analyses plans.[7] 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 documenting the sources 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 (2022) and Franco, Malhotra, and Simonovits (2014). DellaVigna and Linos (2022) find that results from RCTs published in academic journals have significantly larger treatment effects (8.7 pp) compared to RCTs conducted at a larger scale via "Nudge Units" (1.4 pp). Franco, Malhotra, and Simonovits (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 (i) editor decisions closely follow referee recommendations; (ii) papers by highly published authors receive more subsequent citations conditioning on referee recommendations and publication status; and (iii) 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, Figlio, and Lusher (2022) document signaling and network effects in how reviewers evaluate papers written by authors

[7] See Casey, Glennerster, and Miguel (2012) for an in-depth example and analyses of a pre-analysis plan and Brodeur et al. (2022) and Ofosu and Posner (2020) for analyses of the impact of pre-analysis plans on p-hacking and publication rates, respectively.

of matching characteristics: for example, the authors find evidence that reviewers positively evaluate research by authors who went to their same PhD program.

I. Data Sources and Background

Our data consist of two parts. The first are collected from the *Journal of Human Resources*. The JHR is often regarded as a highly selective applied microeconomics field journal. The editorial process at the JHR is similar to that at most other peer-reviewed economics journals. Papers submitted for review are first handled by the head editor. The head editor then either handles the paper themselves or assigns a coeditor to handle the paper. The editor handling the paper then decides whether to reject the paper or to send the paper to reviewers.[8] After receiving reports from reviewers, the editor chooses to either reject the paper or grant a "revise and resubmit."[9] Revised manuscripts are then resubmitted for further review, potentially by the same or additional reviewers; our analyses focus strictly on initially submitted manuscripts, initial reviewer recommendations, and the final draft of accepted manuscripts. 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 percent of the time, while less than 10 percent of reviewer recommendations are strong positives. The overall acceptance rate at the journal is 6 percent.

Our sample of data from the JHR contains all manuscripts submitted for review from 2013 to 2018. During this time frame, there were 2,365 submissions that were desk rejected, 1,018 submissions that were rejected after receiving reviewer recommendation (i.e., "reviewer rejections"), and 223 (eventually) accepted manuscripts. We then keep a random sample of 250 desk rejections, 250 reviewer rejections, and all 223 accepted manuscripts, stratified by year of submission. Lastly, upon reading the paper, we removed manuscripts that did not contain a clear experimental or quasi-experimental statistical inference (difference-in-differences, instrumental variables, regression discontinuity, and/or randomized control trials and experiments); this process closely followed that of Brodeur, Cook, and Heyes (2020).[10] Then, we included initial drafts of accepted papers into our sample. Our final analytic sample contains 705 manuscripts handled across 28 editors: 171 desk rejections, 210 rejections after receiving reviews, 162 drafts of eventually accepted manuscripts, and 162 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, Cook, and Heyes 2020), we only collect estimates from main results

[8] In very rare cases, papers can be accepted without receiving reviewer reports at the JHR. These occurred when the authors provided reviewer reports from previous journals and which the handling editor effectively used to substitute for JHR reviewers. These papers are dropped from our sample.

[9] In rare circumstances, authors of rejected manuscripts may revise their manuscript and submit again to the JHR (i.e., "reject and resubmit"). Our data do not distinguish these manuscripts, instead classifying the manuscript as being rejected (for the first submission), then submitted as a separate manuscript for any subsequent submission. These cases are dropped from our sample.

[10] Examples of omitted papers include literature reviews, methodology papers, descriptive exercises, structural estimations, and other identification strategies such as synthetic control and propensity score matching. Though some papers possess multiple identification strategies, we still coded each paper into a single identification strategy based on what we identified as the "primary" identification; for example, a paper that uses a fuzzy regression discontinuity design was coded as a regression discontinuity (as opposed to an instrumental variable).

tables. Estimates from summary statistics, appendixes, robustness checks, and placebo tests were not collected, nor were results from figures.[11] 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 20,206 test statistics.

Coefficients and standard errors are reported for the vast majority of tests, while *p*-values and *t*-statistics are reported for 2.2 percent and 2.1 percent of tests, respectively. 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. We then transform these *z*-statistics into their corresponding *p*-values. One issue discussed in the literature is the 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 2, 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 *de-rounded* *z*-statistics.[12] We also rely on a second de-rounding method developed by Kranz and Putz (2022), which omits observations that are too coarsely rounded.[13]

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 web page, ideas.repec.org web page, and the National Bureau of Economic Research (NBER) web page: gender, institution of PhD, PhD graduation year, tenure status, prior publication history, and NBER affiliation. Rankings for the prestige of the author's PhD program were also collected from the department productivity rankings on ideas.repec.org.[14]

A. *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

[11] Though some papers provided their "main" results via figures, nearly all these papers provided the corresponding point estimates and standard errors via table as well and thus were coded as "main" estimates as well for our data collection process.

[12] We also collected information on the number of stars reported for all *z*-statistics coded as being equal to two. In total, we have 232 test statistics with *z*-statistics equal to 2. The authors report stars to denote statistical significance for 220 of these 232 tests. Of these, the authors report 1 star for 61 tests, 2 stars for 132 tests, and 3 stars for 10 tests. The authors report no stars for only 17 tests. This means that the majority of tests with $z = 2$ are properly coded as having 2 stars, and most of these tests are statistically significant either at the 10 or 5 percent level, meaning that we are slightly underestimating bunching at the 10 percent level.

[13] We follow Kranz and Putz (2022) and omit all observations whose standard error has a significand below a threshold of 37. The significand consists of the significant digit(s) written as an integer. We also improved Kranz and Putz's (2022) method by making it more demanding; we take into account rounding issues for $z = 1.5$, which may be important for the 10 percent significance level.

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

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	21.15	26.42	23.73	24.80	30.78	38.60	28.27	32.77
Solo authored	0.39	0.49	0.30	0.46	0.25	0.44	0.25	0.44
Share of authors tenured	0.28	0.34	0.28	0.33	0.30	0.33	0.30	0.33
Share of authors female	0.32	0.38	0.34	0.40	0.39	0.38	0.39	0.38
Author avg. years since PhD	6.96	7.50	7.18	6.74	8.85	7.09	8.85	7.09
Oldest author (years since PhD)	11.12	13.67	10.98	10.83	14.25	13.72	14.25	13.72
Author avg. PhD rank	134.19	96.45	93.36	78.25	72.14	74.40	72.14	74.40
Authors highest PhD rank	95.47	104.60	60.07	78.19	38.36	60.64	38.36	60.64
Paper w/ T5 author	0.12	0.33	0.19	0.39	0.31	0.47	0.31	0.47
Paper w/ NBER author	0.08	0.27	0.14	0.35	0.30	0.46	0.30	0.46
Identification strategy								
–Difference-in-differences	0.37	0.48	0.40	0.49	0.40	0.49	0.41	0.49
–Instrumental variables	0.37	0.48	0.21	0.41	0.25	0.43	0.24	0.43
–Regression discontinuity	0.17	0.38	0.23	0.42	0.17	0.38	0.17	0.38
–Randomize control trial	0.09	0.29	0.16	0.37	0.18	0.38	0.18	0.38
Observations	171		210		162		162	

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.

comments, first drafts of accepted manuscripts, and final drafts of accepted manuscripts. Desk-rejected papers tend to have fewer main estimates (21) compared to those sent out for review. Additionally, accepted submissions tend to contain slightly more estimates (31) than submissions rejected at the reviewer stage (24). 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, our summary statistics reveal 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 39 percent rate, 30 percent of those rejected after review are solo authored, and 25 percent of accepted manuscripts are solo authored. Those who published in a "top five" economics journal previously tend to experience more positive outcomes. More experienced authors (measured as years since PhD) and those who came from better-ranked PhD programs also experience more positive paper outcomes. These correlations are unsurprising since these characteristics are generally associated with higher-quality papers.[15] Lastly, turning to identification strategy, randomized control trials appear to have a higher likelihood of getting past the desk and subsequently published relative to instrumental variables strategies.

[15] Of course, it is also possible that these characteristics alone influence paper outcomes through status signaling. For example, Huber et al. (2022) find that 20 percent of the reviewers recommended accept when a Nobel laureate is shown as the paper's author, while less than 2 percent did so when a relatively unknown junior coauthor was shown as the paper's author. Carrell, Figlio, and Lusher (2022) uncover differential outcomes for authors of varying status (e.g., NBER) based on "matches" (e.g., both the author and the reviewer belonging to the NBER).

B. *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 V) across 143 applied microeconomists who listed which journals they had submitted to in the previous 5 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 random sample of rejected manuscripts. For this sample, the eventual publication rate was roughly 59 percent: 58 percent for desk-rejected manuscripts and 60 percent for manuscripts rejected after receiving reviewer recommendations. The most common eventual publication outlets include *Economics of Education Review* (10 percent of eventual publications), *Journal of Health Economics* (7 percent), *Labour Economics* (6 percent), *Journal of Economic Behavior and Organization* (4 percent), *Economic Inquiry* (4 percent), *Health Economics* (4 percent), and *Education Finance and Policy* (4 percent).

II. Methods

Here, we briefly discuss the methods used in the following main results section. We simultaneously present a visual inspection of distribution plots, and results from econometric tests for potential bunching across statistical significance thresholds. Our figures plot the distribution of *p*-values using bins with a width of 0.0025 along the interval $[0.0025, 0.1500]$ for a total of 59 bins.

We then employ two different econometric methods. First, Elliott, Kudrin, and Wüthirch (2022) show, under a broad set of conditions and for many tests, that for any distribution of true effects, the p-curve should be nonincreasing and continuous under the null of no p-hacking. We focus on the tests derived by Elliott, Kudrin, and Wüthirch (2022), which allow for the examination of abnormalities in *individual* distributions of test statistics. Second, we borrow from several studies including Brodeur, Cook, and Heyes (2020) and Gerber and Malhotra (2008a) to conduct what is commonly referred as the "Caliper" test. The Caliper test allows us to quantify and conduct inference on the extent of the statistical bunching and to directly compare two (or more) different distributions of test statistics (e.g., desk rejections against non–desk rejections). Moreover, the Caliper test also allows

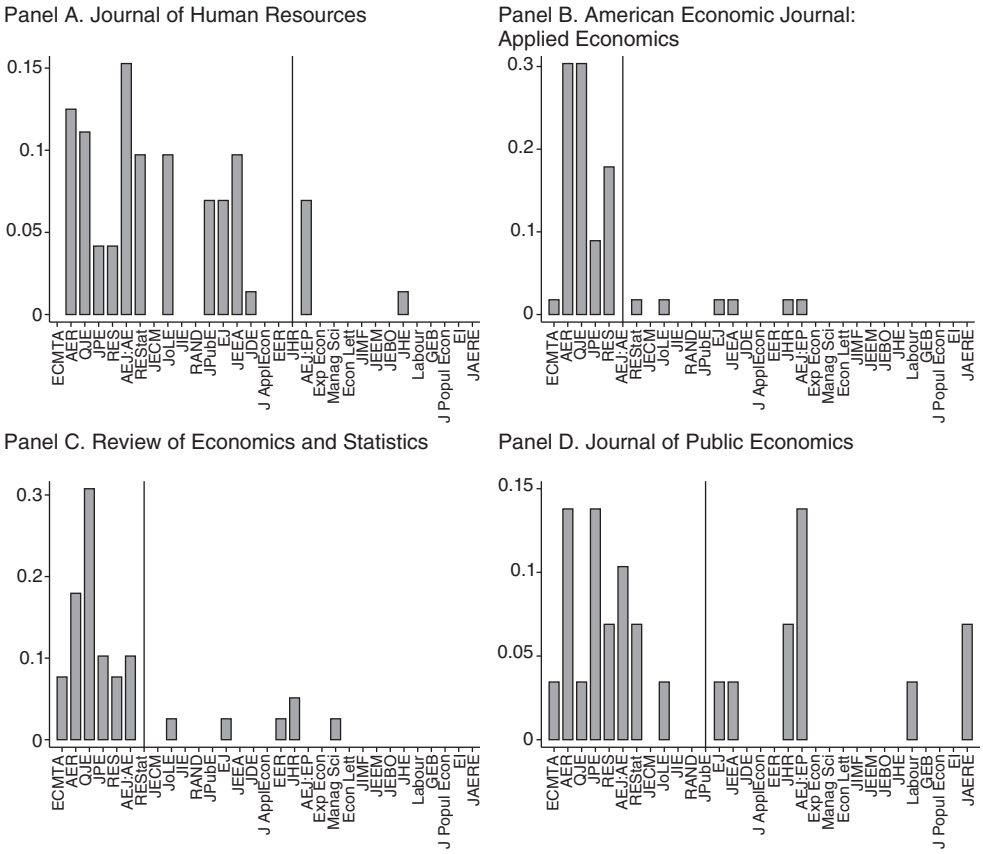

FIGURE 1. DISTRIBUTION OF JOURNAL SUBMISSIONS MADE PRIOR TO A SUBMISSION AT A PARTICULAR JOURNAL

Notes: Results based on survey data described in Section V. 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, Panel 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.

us to identify other factors/heterogeneities that potentially mediate the observed bunching, such as differential assignment to coeditor or author characteristics. See online Appendix 1 for more details on the Caliper test.

III. Main Results

A. *Initial Submissions*

We start with Figure 2, which plots the distribution of *p*-values and *z*-statistics for our full sample of initial submissions. Starting with the *z*-statistics in the first panel, the distribution displays a two-humped shape, with one hump for test statistics below 1 and another around the 5 percent statistical threshold. Approximately 51, 42, and 28 percent of test statistics are significant at the 10, 5, and 1 percent

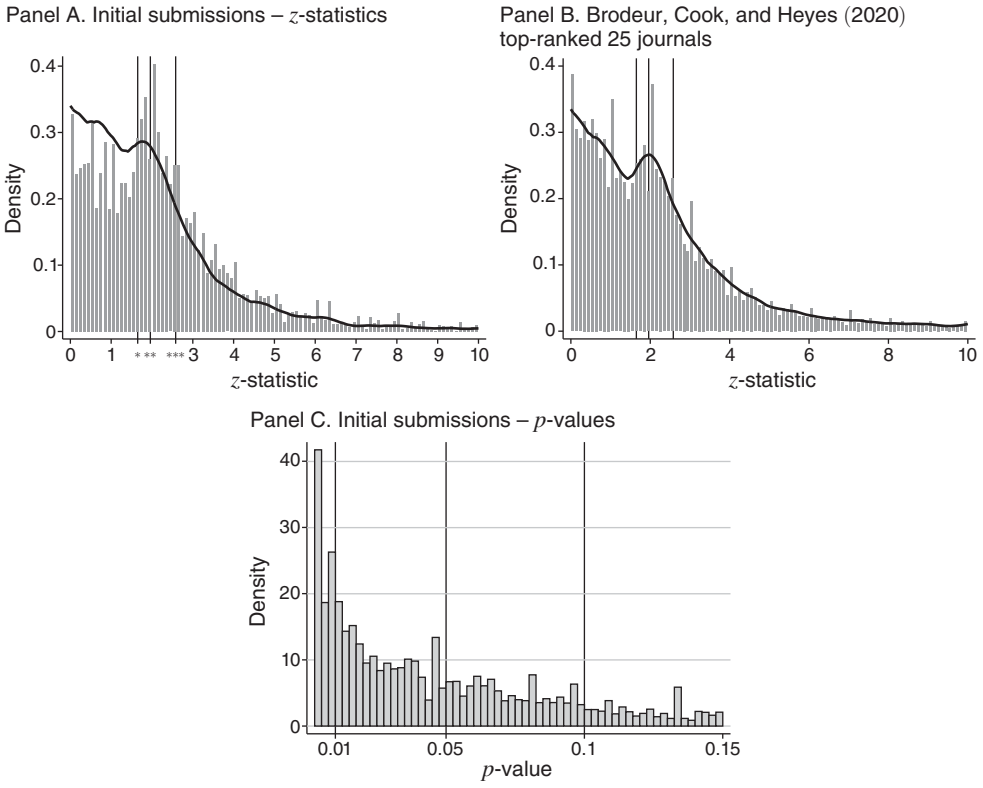

FIGURE 2. DISTRIBUTIONS OF z-STATISTICS AND p-VALUES FOR INITIAL SUBMISSIONS VERSUS z-STATISTICS FROM BRODEUR, COOK, AND HEYES (2020)

Notes: The first figure displays a histogram of test statistics for $z \in [0, 10]$, with bins of width 0.1, among all initial submissions in our dataset. As a comparison, the second figure plots the corresponding histogram of z-statistics from the top-ranked 25 economics journals published in 2015 and 2018 (from Brodeur, Cook, and Heyes 2020). The third figure displays a histogram of test statistics for p-values $\in [0.0025, 0.1500]$, with bins of width 0.0025. Vertical reference lines are displayed at conventional two-tailed significance levels. For the first two histograms, we superimpose an Epanechnikov kernel density curve. We use the inverse of the number of tests presented in the same article to weight observations.

levels, respectively.[16] In the second panel of Figure 2, we copy the distribution of published test statistics among 25 leading economics journals from Brodeur, Cook, and Heyes (2020). Overall, the distribution from our study closely reflects the distribution from Brodeur, Cook, and Heyes (2020), although with more bunching just after the 10 percent level threshold for initial submissions. As seen in the third panel of Figure 2, there are corresponding jumps in p-values below 0.01, 0.05, and 0.10.

[16] Online Appendix Figure A1 presents corresponding distributions using alternate binwidths of 0.05 and 0.15. As shown in online Appendix Figure A7, overall, using de-rounded p-values smooths potential discontinuities in histograms but does not change the shape of the distributions. Bunching just below the 10 percent level slightly increases, while the extent of bunching around 5 percent slightly decreases.

TABLE 2—ELLIOTT, KUDRIN, AND WÜTHIRCH'S (2022) TESTS

Threshold:	5 percent significance		10 percent significance				
Sample:	Bin.	Discont.	Bin.	Discont.	CS1	CS2B	LCM
Initial submissions	0.000	0.743	0.000	0.004	0.000	0.000	0.004
Desk reject stage							
Desk rejections	0.005	0.599	0.014	0.337	0.524	0.225	0.338
Not desk rejected	0.000	0.003	0.403	0.141	0.000	0.000	0.038
Reviewer stage							
Recommended rejection	0.000	0.380	0.906	0.001	0.005	0.003	0.077
Recommend against outright rejection	0.000	0.028	0.813	0.472	0.002	0.000	0.248
Recommend accept as is or with minor edits	0.001	0.162	0.028	0.848	0.000	0.000	0.703
Accepted manuscripts							
Initial submission	0.000	0.000	0.012	0.531	0.464	0.001	0.404
Published version	0.000	0.031	0.000	0.970	0.000	0.000	0.348
Peer review							
All rejections	0.000	0.485	0.466	0.005	0.009	0.003	0.029
Accepted manuscripts	0.000	0.031	0.000	0.970	0.000	0.000	0.348
After rejection							
Published elsewhere	0.000	0.878	0.376	0.052	0.023	0.001	0.374
Failed to publish	0.001	0.678	0.671	0.046	0.145	0.004	0.210

Notes: Each panel is a direct application of Elliott, Kudrin, and Wüthirch's (2022) binomial, discontinuity, and nonincreasingness tests to a subsample. The first two columns focus on the 5 percent significance threshold, while the next two columns focus on the 10 percent significance threshold. The remaining columns focus on the nonincreasingness tests. We do not weight observations.

Additionally, the mass of p-values increases as we move from right to left (to relatively "rarer" p-values).

To more formally test for evidence of p-hacking and/or publication bias, we next conduct various tests that have been implemented in the recent literature. First, Table 2 presents results from five tests of Elliott, Kudrin, and Wüthirch (2022): the binomial and discontinuity tests and three tests based on the expected nonincreasingness of the p-curve. First, rather than examining the distribution of test statistics on either side of statistical thresholds, as an alternative, the binomial test examines the null hypothesis that the p-curve is nonincreasing just below a significance cutoff. For the 5 percent significance threshold, we follow Elliott, Kudrin, and Wüthirch (2022) and split $[0.04, 0.05]$ into two subintervals $[0.040, 0.045]$ and $(0.045, 0.050]$. Under the null of no p-hacking, the fraction of p-values in $(0.045, 0.050]$ should be smaller than or equal to one-half; i.e., the fraction of p-values in the bin closer to the cutoff should be weakly smaller than the fraction in the bin farther away. This is in contrast to our caliper tests in which we compare the mass above and below 0.050.[17] For the 10 percent significance threshold, we again follow Elliott, Kudrin, and Wüthirch's (2022) decision and split $[0.09, 0.010]$ into two subintervals $[0.090, 0.095]$ and

[17] Of note, our caliper test analyses compare the just-above to just-below significance masses of test statistics. In contrast, Elliott, Kudrin, and Wüthirch's (2022) binomial test attempts to distinguish p-hacking from publication bias, assuming that p-hacking always (weakly) favors smaller p-values.

$(0.095, 0.010]$.[18] We also repeat this test for the 1 percent significance level in online Appendix Table A2. The *p*-values for the 5 and 10 percent levels are 0.000, confirming the visual result of bunching just above marginally significant thresholds. In contrast, we find no evidence of p-hacking for the 1 percent significance threshold.

Second, we provide discontinuity tests for each significance threshold, which test for a violation of the continuity of the p-curve around a threshold with data-driven bandwidth selection. This is an application of the density discontinuity test from Cattaneo, Jansson, and Ma (2020). Again, we find evidence of bunching with the test at the 10 percent level significance threshold. On the other hand, we find no evidence of p-hacking for the 5 and 1 percent thresholds.

Third, the three tests based on the expected nonincreasingness of the p-curve reject the null of no p-hacking.[19] First, CS1 is an application of the conditional chi-squared test introduced in Cox and Shi (2022). Second, CS2B is a histogram-based test for 2-monotonicity and bounds on the p-curve and its first two derivatives. The third test is based on the least concave majorant (LCM). The rationale for this last test is that the CDF of *p*-values is concave. The *p*-values are all 0.00 for the initial submissions.

So far, the visual inspection and econometric tests suggest that initial submissions suffer from p-hacking, especially around the 10 percent significance threshold. As a robustness check, we rely on Elliott, Kudrin, and Wüthirch's (2022) tests but use de-rounded *p*-values. The *p*-values are reported in online Appendix Tables A3 and A4. As in Elliott, Kudrin, and Wüthirch (2022), we find virtually no evidence for p-hacking and publication bias when using de-rounded *p*-values. The only exception is the histogram-based test for 2-monotonicity and bounds on the p-curve and its first two derivatives, which detects p-hacking and publication bias.

Overall, our results provide suggestive evidence that the distribution of test statistics faced by the journal is already skewed toward statistical thresholds. In other words, we can rule out the case that the distribution of test statistics initially faced by editors is "free of p-hacking" and then a process of publication bias skews the distribution toward statistical significance.[20] Thus, the observed distributions from prior studies cannot be solely attributed to the peer review process. Even if, for example, the papers received by the JHR were influenced by selection from "upstream" journals, one would assume that if a preference for statistical significance were at work (i.e., a publication bias at "upstream" journals), then we would expect "too few" marginally significant results to be submitted to the JHR. 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.

Next, we use the Caliper test to investigate selective reporting by author and paper characteristics near statistical significance thresholds among initially submitted

[18] We report the number of observations for each stage and decision for the binomial and discontinuity tests in online Appendix Table A1.

[19] We omit Fisher's Test, as it almost always yields a *p*-value of 1 as in Elliott, Kudrin, and Wüthirch (2022).

[20] Our data do not, however, observe the distribution of test statistics among submissions made to any journals prior to the authors submitting to the JHR. Thus, it is possible that for the same research study, the distribution of submitted test statistics to journals prior to the JHR differs from the distribution submitted to the JHR. This could happen if, for example, authors adjust their estimates in response to comments from editors and reviewers after journal rejection.

Table 3—Caliper Test, Author Heterogeneity in Initial Submissions

	10 percent significant	5 percent significant	1 percent significant
Solo authored	−0.016	−0.012	−0.040
	(0.050)	(0.051)	(0.056)
Share tenured	0.030	0.072	−0.012
	(0.072)	(0.075)	(0.081)
Share female	0.046	−0.014	−0.011
	(0.050)	(0.047)	(0.057)
Author avg. years since PhD	−0.012	0.009	−0.003
	(0.005)	(0.007)	(0.007)
max{Author years since PhD}	0.004	−0.004	0.003
	(0.002)	(0.003)	(0.004)
Author avg. PhD rank	−0.000	0.000	−0.000
	(0.001)	(0.001)	(0.001)
Authors highest PhD rank	0.001	−0.001	0.001
	(0.001)	(0.001)	(0.001)
Paper w/ T5 author	0.101	0.023	0.064
	(0.054)	(0.056)	(0.057)
Paper w/ NBER author	−0.048	−0.093	−0.055
	(0.058)	(0.059)	(0.052)
Identification strategy			
Diff-in-diff	0.019	0.155	0.044
	(0.049)	(0.050)	(0.052)
IV	0.094	0.089	−0.051
	(0.057)	(0.050)	(0.061)
RCT	−0.009	0.107	−0.049
	(0.058)	(0.062)	(0.075)
Observations	2,027	2,047	1,361
z sample bounds	$[1.35, 1.95]$	$[1.66, 2.26]$	$[2.28, 2.88]$

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.

papers. Table 3 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 $z \in [1.35, 1.95]$, $z \in [1.66, 2.26]$, and $z \in [2.28, 2.88]$ 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 heterogeneity comes from the paper's identification strategy, where difference-in-differences and instrumental variables, and some evidence for experimental papers, tend to contain more marginally significant estimates compared to regression discontinuities. This result is comparable to that of Brodeur, Cook, and Heyes (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, Cook, and Heyes (2020) cannot be driven

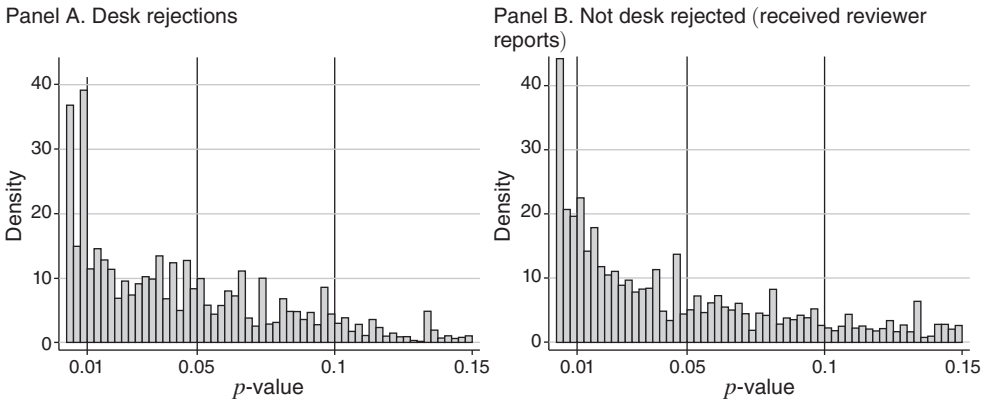

FIGURE 3. EDITOR'S FIRST DECISION—DISTRIBUTIONS OF *p*-VALUES BY DESK REJECTION

Notes: This figure displays histograms of test statistics for p-values $\in [0.0025, 0.1500]$ by editor's first decision. Histogram bins are 0.0025 wide. Reference lines are displayed at conventional two-tailed significance levels. We use the inverse of the number of tests presented in the same article to weight observations.

by the peer review process being biased simultaneously toward (i) marginal significance and (ii) particular identification strategies. Finally, other considered heterogeneities do not appear to be significantly associated with marginal significance, including solo authorship, author tenure, gender, years since PhD, the author's PhD ranking, prior publication in a "top five" journal, and author NBER affiliation.

B. *Results by Desk Rejection*

In Figure 3, 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.[21] We then use the Caliper test to formally examine whether submissions that are desk rejected are more likely to report marginally (in)significant estimates.[22] 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 4 (see online Appendix Table A5 for the 1 percent statistical significance threshold).[23]

[21] Online Appendix Figure A2 plots the corresponding smoothed distributions of z-scores by desk rejection into a single panel. Online Appendix Figures A3–A6 plot the remaining comparative distributions along the peer review process. Online Appendix Figure A8 plots the de-rounded distribution of p-values by desk rejection status. See online Appendix Figures A9–A12 for the remaining de-rounded distributions along the peer review process.

[22] We briefly discuss tests from Elliott, Kudrin, and Wüthirch (2022) here. Of note, these tests do not directly compare the two distributions of test statistics. They also do not allow for the inclusion of covariates, which may be an important issue in our context. For the binomial tests, we find that both initial submissions that are desk rejected and sent out for review suffer from p-hacking at the 5 percent level, but we only find evidence of p-hacking at the 10 percent level for desk-rejected manuscripts. On the other hand, the discontinuity tests provide little evidence of p-hacking at the 10 percent significance level and only evidence of p-hacking at the 5 percent significance level for submissions sent out for review. For the three tests based on the expected nonincreasingness of the p-curve, we reject the null of no p-hacking for manuscripts sent out for review, while the p-values are larger than 0.2 for desk-rejected manuscripts.

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

Table 4—Desk Rejection: Caliper Test, Significant at the 10 Percent and 5 Percent Levels

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. 10 percent significant						
Desk rejected	0.142	0.139	0.125	0.119	0.083	0.078
	(0.042)	(0.048)	(0.047)	(0.048)	(0.056)	(0.055)
Observations	2,027	2,027	2,027	2,027	957	957
z sample bounds	$[1.35, 1.95]$	$[1.35, 1.95]$	$[1.35, 1.95]$	$[1.35, 1.95]$	$[1.50, 1.80]$	$[1.50, 1.80]$
Coeditor fixed effect		Y	Y	Y	Y	Y
Identification strategy			Y	Y		Y
Paper-author controls				Y		Y
Panel B. 5 percent significant						
Desk rejected	−0.002	0.022	0.019	0.022	0.004	0.007
	(0.045)	(0.045)	(0.046)	(0.047)	(0.068)	(0.068)
Observations	2,042	2,042	2,042	2,042	1,062	1,062
z sample bounds	$[1.66, 2.26]$	$[1.66, 2.26]$	$[1.66, 2.26]$	$[1.66, 2.26]$	$[1.81, 2.11]$	$[1.81, 2.11]$
Coeditor fixed effects		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 $z \pm 0.30$. Columns 5 and 6 restrict the sample to $z \pm 0.15$. 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.

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 of Table 4, we restrict the sample to $z \in [1.35, 1.95]$ and $z \in [1.66, 2.26]$ 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 over 13 percentage points more likely to be statistically significant at the 10 percent level than estimates in manuscripts that are not desk rejected. The estimate is statistically significant at the 1 percent level (online Appendix Tables A6 and A7 provide similar conclusions using de-rounded p-values). This provides evidence that on average, marginally significant estimates are more likely to be desk rejected (desk-rejected papers display significantly more bunching at the 10 percent level). In contrast, desk-rejected estimates are not statistically more likely than non-desk-rejected estimates to be marginally statistically significant at the 5 percent level. This "filtering out" of marginally significant estimates could be driven by responses in editor behavior and/or 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. Though we cannot fully disentangle these two possibilities (editor behavior from unobserved correlates of paper quality and p-hacking), in the

following analyses, we enhance our model with observable characteristics to test for whether paper characteristics are associated with desk rejection propensity and the likelihood of containing marginally significant estimates.

At the *Journal of Human Resources*, each manuscript is assigned one handling coeditor, each of whom have complete autonomy over rejection, revision, and publication decisions.[24] One plausible explanation for our findings is that coeditors may have been differentially assigned papers with marginally (in)significant estimates and that coeditors may have different propensities to desk reject papers. More specifically, it may be that coeditors 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 coeditor fixed effects (column 2 of Table 4). The point estimate in panel A changes only slightly and remains statistically significant at the 1 percent level. The point estimate for the 5 percent significance level increases slightly but remains statistically insignificant.

Moving to column 3, we similarly test whether controlling for differences in the paper's identification strategy can explain the findings. 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 coeditors, the statistical bunching at the 10 percent level of significance cannot be explained by the paper's identification strategy. These results remain statistically significant at the 5 percent level, though the magnitude of the effect drops slightly (1.4 percentage points). Results for 5 percent significance remain positive but insignificant.

In column 4, we include the full vector of author characteristics. Again, we find that author characteristics are not simultaneously indicative of both (i) greater tendency to have marginally significant estimates and (ii) increased likelihood of being desk rejected.

As a final robustness check, in columns 5 and 6, we replicate columns 2 and 4, respectively, but for a narrower bandwidth of test statistics. Though our point estimates slightly decrease and lose their statistical significance, magnitudes for marginal significance at the 10 percent level remain large.

This finding, that the desk rejection stage picks up valuable information, is in line with findings from Card and DellaVigna (2020), who find that all else equal, desk-rejected papers end up with fewer ex post citations compared to papers rejected after review. As such, this result further supports the argument that desk rejection decisions from editors are informative, even after controlling for author and paper characteristics. Hence, to the extent that our vector of controls account for "paper quality," our results suggest that desk rejection decisions filter out false positives, on average. At a minimum, our results do not suggest that editors have a bias *toward* marginal significance.

C. *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

[24] Our sample spans 28 individual coeditors. In total, these 28 coeditors have served across over 40 journal editorial boards in economics.

statistic-paper-reviewer level, where each paper appears in the data proportional to the number of reviewers assigned. Estimates are 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 4 presents the distribution of *p*-values split by rejection recommendations, nonrejection recommendations (ranking of 2+), and strong positive recommendations (ranking of 4 or 5). The mass of *p*-values around significance thresholds becomes more pronounced as we move from the first figure (rejections) to the third figure (strong positive recommendations).

We now turn to the Caliper test to formally test for differences in these distributions in Table 5 (see online Appendix Table A8 for the 1 percent statistical significance threshold and online Appendix Tables A9 and A10 for de-rounded *p*-values). Similar to Table 4, we sequentially add more control variables (including reviewer controls) across columns. 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, the results provide further evidence that marginal statistical significance is associated with positive reviewer recommendations. From column 1 in panel A of Table 5, we see that papers that received either a weakly positive or strongly positive review were more likely to be marginally significant at the 10 percent level than negative reviews (though these estimates are not statistically significant). The results do not change much as we sequentially add additional covariates through column 5. Importantly, much like in our previous analysis where we included coeditor fixed effects to account for potential correlations in an editor's set of manuscripts and the editor's propensity for rejection, in column 3 we include a vector of reviewer-level covariates to account for potential correlation in (i) the assignment of manuscripts with marginally significant results to (ii) reviewers with a higher propensity to review manuscripts positively. We find little difference in our estimates between columns 2 and 3, suggesting editors do not choose reviewers based on both the paper's marginal significance and the reviewer's propensity to review papers positively or negatively. Finally, in panel B, we turn to potential reviewer preference for statistical significance at the 5 percent level, where we find significant differences in the likelihood a reviewer gives a positive review based on the paper's estimates. Estimates for our full specified model in column 5 are statistically significant at the 10 percent level for both weak positive and strong positive reviews (relative to negative reviews).

In total, the evidence suggests that reviewer recommendations are positively influenced by marginal statistical significance.[25] Furthermore, differences in reviewer recommendations by marginal significance are not explained by the paper's coeditor, reviewer controls, the paper's identification strategy, or author

[25] For the discontinuity test, we find some evidence of p-hacking at the 5 percent level for nonrejection recommendations and a *p*-value of 0.16 for strong positive recommendations. Again, these results should be viewed with caution, as we are not controlling for observable author and article characteristics. For the binomial tests, which attempt to distinguish p-hacking from publication bias, we obtain *p*-values that are all below 0.001 for the 3 reviewer recommendations for the 5 percent significance level.

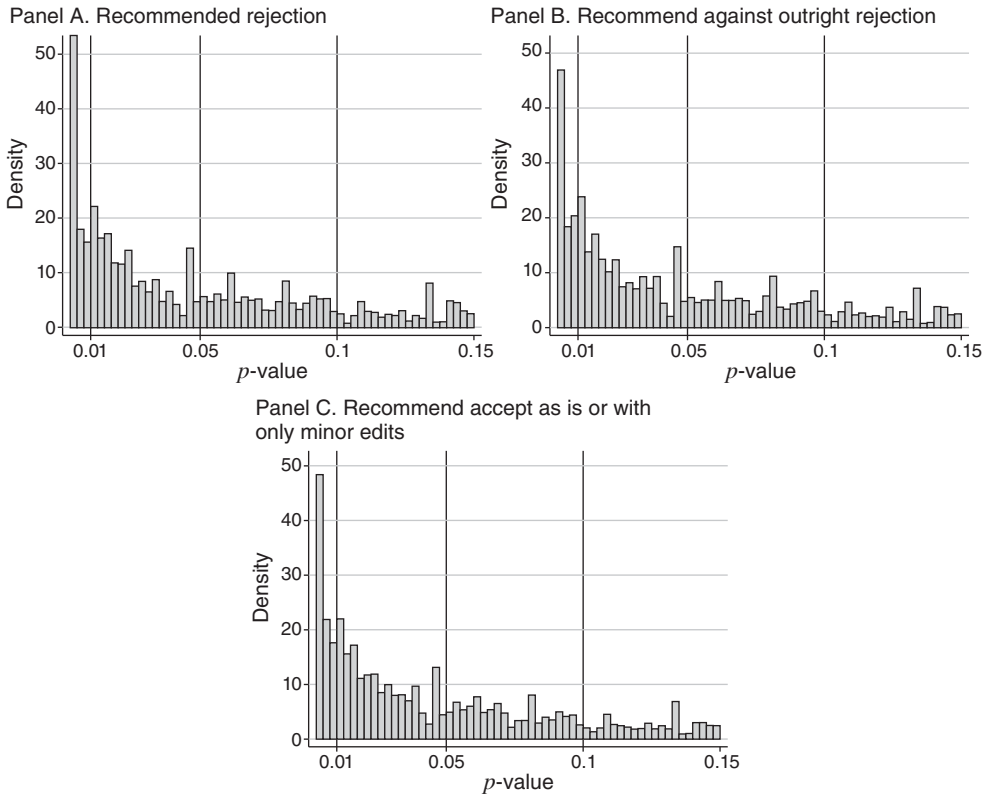

FIGURE 4. REVIEWER STAGE—DISTRIBUTIONS OF *p*-VALUES BY REVIEWER RECOMMENDATION

Notes: This figure displays histograms of test statistics for p-values $\in [0.0025, 0.1500]$ for the reviewer stage. Histogram bins are 0.0025 wide. Reference lines are displayed at conventional two-tailed significance levels. We use the inverse of the number of tests presented in the same article to weight observations.

characteristics. Note that this result differs from the desk rejection results, where marginally significant results were filtered out via desk rejection and this filtering out was partially explained by the paper's identification strategy.

D. *Comparing Initial versus Final Drafts of Accepted Papers*

In Figure 5 we juxtapose the distribution of p-values 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 see a hump among initial submissions below the 5 percent level. We then present estimates from Caliper tests in Table 6 in the same manner as in Table 4 (see online Appendix Table A11 for the 1 percent statistical significance threshold and online Appendix Tables A12 and A13 for de-rounded p-values). Similar to the graphical evidence, with positive coefficients, we see that initial submissions were more likely to display marginally significant results. These estimates are, however, imprecisely estimated, and the magnitudes of the effects are rather small at around 2 percentage

Table 5—Reviewer Rejection: Caliper Test, Significant at the 10 Percent and 5 Percent Levels

	(1)	(2)	(3)	(4)	(5)
Panel A. 10 percent significant					
–Weakly Positive Recommendation	0.049	0.030	0.037	0.036	0.039
	(0.036)	(0.032)	(0.031)	(0.030)	(0.028)
–Minor Edits or Accept As Is	0.053	0.035	0.035	0.038	0.046
	(0.051)	(0.044)	(0.045)	(0.045)	(0.043)
Observations	3,151	3,151	3,151	3,151	3,151
z sample bounds	[1.35, 1.95]	[1.35, 1.95]	[1.35, 1.95]	[1.35, 1.95]	[1.35, 1.95]
Coeditor fixed effects		Y	Y	Y	Y
Reviewer controls			Y	Y	Y
Identification strategy				Y	Y
Paper-author controls					Y
Panel B. 5 percent significant					
–Weakly Positive Recommendation	0.060	0.067	0.056	0.053	0.053
	(0.036)	(0.035)	(0.034)	(0.033)	(0.031)
–Minor Edits or Accept As Is	0.091	0.096	0.084	0.075	0.083
	(0.058)	(0.052)	(0.050)	(0.051)	(0.050)
Observations	3,142	3,142	3,142	3,142	3,142
z sample bounds	[1.66, 2.26]	[1.66, 2.26]	[1.66, 2.26]	[1.66, 2.26]	[1.66, 2.26]
Coeditor fixed effects		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.

points. Thus, we conclude that there is little difference between the first and final drafts of accepted manuscripts.[26] Still, the lack of a negative effect reveals that the peer review process among accepted papers does *not* push papers toward marginally significant estimates.

E. *Overall Impact of Peer Review—Accepted versus Rejected Manuscripts*

Lastly, in Figure 6 and Table 7 (see online Appendix Table A14 for the 1 percent statistical significance threshold and online Appendix Tables A15 and A16 for de-rounded p-values), we compare the distribution of test statistics from the final

[26] 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.

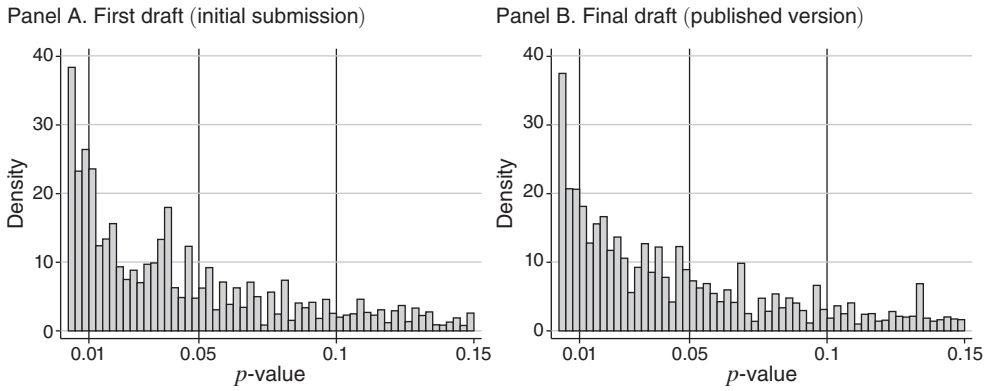

FIGURE 5. DISTRIBUTIONS OF *p*-VALUES BY DRAFT VERSIONS OF ACCEPTED MANUSCRIPTS

Notes: This figure displays histograms of test statistics for p-values $\in [0.0025, 0.1500]$ for published manuscripts against their corresponding first drafts (initial submissions). Histogram bins are 0.0025 wide. Reference lines are displayed at conventional two-tailed significance levels. We use the inverse of the number of tests presented in the same article to weight observations.

TABLE 6—INITIAL VERSUS FINAL (ACCEPTED) SUBMISSIONS: CALIPER TEST, SIGNIFICANT AT THE 10 PERCENT AND 5 PERCENT LEVELS

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. 10 percent significant						
Initial Draft	0.014	0.015	0.017	0.010	−0.038	−0.035
	(0.044)	(0.039)	(0.039)	(0.038)	(0.054)	(0.050)
Observations	1,539	1,539	1,539	1,539	728	728
z sample bounds	[1.35, 1.95]	[1.35, 1.95]	[1.35, 1.95]	[1.35, 1.95]	[1.50, 1.80]	[1.50, 1.80]
Coeditor fixed effects		Y	Y	Y	Y	Y
Identification strategy			Y	Y		Y
Paper-author controls				Y		Y
Panel B. 5 percent significant						
Initial Draft	0.024	0.020	0.015	0.027	0.045	0.057
	(0.048)	(0.043)	(0.043)	(0.039)	(0.053)	(0.046)
Observations	1,589	1,589	1,589	1,589	836	836
z sample bounds	[1.66, 2.26]	[1.66, 2.26]	[1.66, 2.26]	[1.66, 2.26]	[1.81, 2.11]	[1.81, 2.11]
Coeditor fixed effects		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 $z \pm 0.30$. Columns 5 and 6 restrict the sample to $z \pm 0.15$. 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.

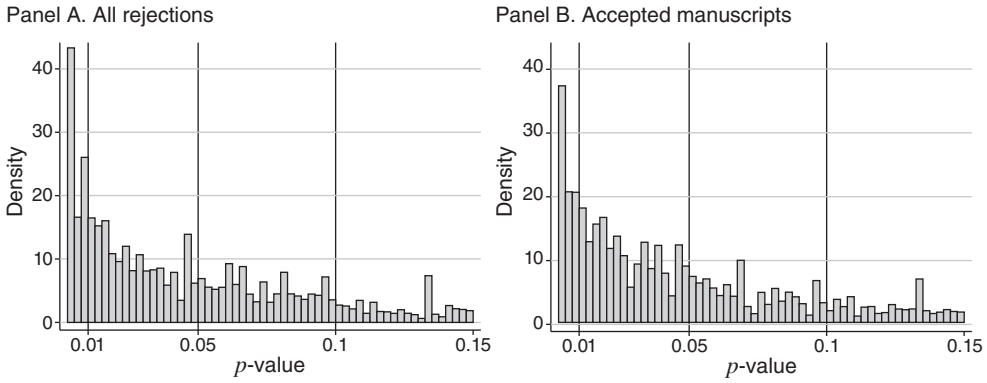

FIGURE 6. PEER REVIEW—DISTRIBUTIONS OF *p*-VALUES BY REJECTED AND FINAL DRAFT OF ACCEPTED MANUSCRIPTS

Notes: This figure displays histograms of test statistics for p-values $\in [0.0025, 0.1500]$ for all rejected manuscripts versus the final draft of published manuscripts. Histogram bins are 0.0025 wide. Reference lines are displayed at conventional two-tailed significance levels. We use the inverse of the number of tests presented in the same article to weight observations.

TABLE 7—ACCEPTED VERSUS REJECTED MANUSCRIPTS: CALIPER TEST, SIGNIFICANT AT THE 10 PERCENT AND 5 PERCENT LEVELS

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. 10 percent significant						
Accepted Manuscripts	−0.060	−0.030	−0.026	−0.013	−0.070	−0.051
	(0.039)	(0.040)	(0.039)	(0.040)	(0.046)	(0.047)
Observations	1,985	1,985	1,985	1,985	923	923
z sample bounds	$[1.35, 1.95]$	$[1.35, 1.95]$	$[1.35, 1.95]$	$[1.35, 1.95]$	$[1.50, 1.80]$	$[1.50, 1.80]$
Coeditor fixed effects		Y	Y	Y	Y	Y
Identification strategy			Y	Y		Y
Paper-author controls				Y		Y
Panel B. 5 percent significant						
Accepted Manuscripts	0.045	0.008	0.006	0.000	0.002	−0.009
	(0.044)	(0.040)	(0.039)	(0.039)	(0.053)	(0.057)
Observations	1,968	1,968	1,968	1,968	1,006	1,006
z sample bounds	$[1.66, 2.26]$	$[1.66, 2.26]$	$[1.66, 2.26]$	$[1.66, 2.26]$	$[1.81, 2.11]$	$[1.81, 2.11]$
Coeditor fixed effects		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 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 $z \pm 0.30$. Columns 5 and 6 restrict the sample to $z \pm 0.15$. 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.

draft of accepted manuscripts against all rejections (desk rejections plus rejections after reviews). This comparison allows us to evaluate the overall impact of the peer review process by comparing the net effect of the prior three sections: First, as

previously shown, marginally significant estimates are more likely to be desk rejected (Figure 3 and Table 4). Second, among non–desk rejections, statistically significant estimates are more likely to receive positive recommendations from reviewers (Figure 4 and Table 5). Editors then take reviewer recommendations and decide which papers to accept, which produces little change in the distribution of estimates across first and final drafts (Figure 5 and Table 6).

Overall, the results from Figure 6 and Table 7 reveal little difference between rejected and accepted manuscripts. Without the inclusion of covariates (i.e., graphically and in column 1 of Table 7), we observe slightly *more* bunching at the 10 percent significance level for rejected manuscripts.[27] However, once we account for covariates, we find little difference in the propensity for marginal significance in accepted manuscripts versus rejections. Importantly these results suggest that the peer review process does *not* exacerbate (nor attenuate) issues of p-hacking.

F. *After Journal Rejection—Eventually Published versus Never Published Manuscripts*

The prior sections examine changes in the distribution of test statistics at each stage of the peer review process. In this section, we investigate what happens to papers that were rejected in our sample, with a particular focus on whether a rejected manuscript eventually publishes elsewhere and whether these publications exhibit differences in p-hacking behavior. We do so to help determine whether our results are generalizable to the broader profession. For instance, if the distribution of eventually published manuscripts displays greater heaping, this suggests a larger publication bias in the profession overall. As such, our finding from the JHR may either be negligible or simply anomalous. To do so, we turn to our dataset that matches rejected papers to their (potential) eventual publication outlet.

Figure 7 compares the distributions of *p*-values for previously rejected manuscripts from the JHR that published elsewhere versus those that failed to publish. Likewise, Table 8 (see online Appendix Table A17 for the 1 percent statistical significance threshold and online Appendix Tables A18 and A19 for de-rounded *p*-values) tests for statistically significant differences using the Caliper test. Visually, eventually published manuscripts appear to have a sharper jump around the 10 percent threshold, while never published manuscripts tend to have more statistically significant estimates at the 5 percent level. The Caliper tests confirm these observations: never published manuscripts are less likely to have significant estimates at the 10 percent level (though most estimates are noisily estimated) in favor of containing more significant estimates at the 5 percent level.

Though noisy, these results appear inconsistent with the idea that there is a "graveyard" of working papers with null results that fail to publish, and suggest the peer review phenomena identified in the prior sections are likely applicable to the broader economics profession. That is, the net effect of the peer review process has a negligible effect on statistical bunching, and/or (observably) "bad" papers tend to p-hack more.

[27] This result is supported by the discontinuity and nonincreasingness tests presented in Table 2 and online Appendix Table A2.

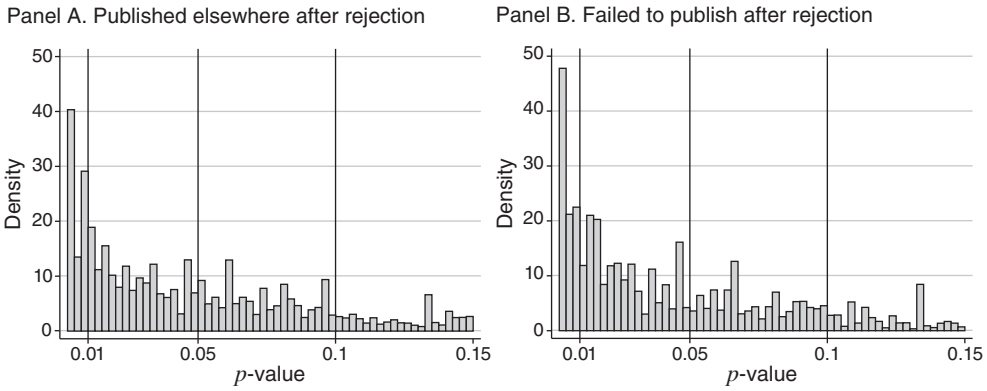

FIGURE 7. AFTER REJECTION—DISTRIBUTIONS OF *p*-VALUES BY WHETHER THE PAPER EVENTUALLY PUBLISHED ELSEWHERE

Notes: This figure displays histograms of test statistics for p-values $\in [0.0025, 0.1500]$ for rejected manuscripts that eventually published elsewhere versus rejected manuscripts that failed to publish anywhere else. Histogram bins are 0.0025 wide. Reference lines are displayed at conventional two-tailed significance levels. We use the inverse of the number of tests presented in the same article to weight observations.

However, since this exercise only tracks papers submitted to the JHR and their subsequent long-run outcomes, we recognize these findings may not be applicable to the peer review process at other journals, particularly those ranked higher than the JHR. We note, however, that Brodeur, Cook, and Heyes (2020) find similar levels of bunching at papers published in the Top 5 journals in comparison to journals ranked 6–25, which includes the JHR.

In summary, we find the following: initial submissions display significant bunching; papers sent for review display less bunching than desk-rejected papers; reviewer recommendations, in contrast, have a positive bias toward marginally significant results; and papers never published possess more marginally significant results. Hence, these results suggest that researchers engage in p-hacking prior to submitting their papers to academic journals, possibly in response to beliefs about preferences of editors and reviewers for significant results. Our results also suggest that many papers with insignificant results are likely never submitted for publication consideration. We later discuss these issues in the context of our survey results in Section V.

IV. Robustness Checks

For robustness, we conduct additional Caliper tests in the online Appendix to address the sensitivity of our estimates to various coding and modeling decisions. First, online Appendix Tables A20–A24 mirror our primary tables and show that our findings are not sensitive to including the test statistics that we had coded as "ambiguous" during the data collection phase. Second, in online Appendix Tables A25–A29, we show our results are not sensitive when accounting for the possibility that papers across different phases of the peer review process have differing quantities of main results tables. To do so, we conduct Caliper tests for our primary bandwidths while restricting our sample to the first main results table for each manuscript. Next, online Appendix Tables A30–A34 show our results are not sensitive to alternative (wider) bandwidths.

Table 8—Never Published versus Published Elsewhere: Caliper Test, Significant at the 10 Percent and 5 Percent Levels

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. 10 percent significant						
Never Published	−0.043	−0.046	−0.050	−0.072	−0.092	−0.097
	(0.050)	(0.047)	(0.045)	(0.043)	(0.053)	(0.051)
Observations	1,239	1,239	1,239	1,239	572	572
z sample bounds	[1.35, 1.95]	[1.35, 1.95]	[1.35, 1.95]	[1.35, 1.95]	[1.50, 1.80]	[1.50, 1.80]
Coeditor fixed effects		Y	Y	Y	Y	Y
Identification strategy			Y	Y		Y
Paper-author controls				Y		Y
Panel B. 5 percent significant						
Never Published	0.052	0.094	0.084	0.100	0.067	0.059
	(0.050)	(0.045)	(0.045)	(0.045)	(0.068)	(0.066)
Observations	1,209	1,209	1,209	1,209	609	609
z sample bounds	[1.66, 2.26]	[1.66, 2.26]	[1.66, 2.26]	[1.66, 2.26]	[1.81, 2.11]	[1.81, 2.11]
Coeditor fixed effects		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 variable of interest "Never published" equals one if the rejected manuscript failed to publish elsewhere. In columns 1–4, we restrict the sample to $z \pm 0.30$. Columns 5 and 6 restrict the sample to $z \pm 0.15$. 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.

V. 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 make in the process of conducting research and writing up the 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 1 of the 4 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, Cook, and Heyes (2020). We then dropped authors with an invalid or missing email address. 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.

Results are presented in Table 9. Of particular interest, we first find that approximately 30 percent of authors have stopped a research study or refrained from

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 (1.46)	7.97 (1.55)	7.93 (1.60)	8.04 (1.44)	8.12 (1.30)	7.56 (1.89)	8.19 (1.40)	8.35 (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, Cook, and Heyes (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.

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 percent), excluding or recategorizing data after seeing the effects of doing so (17 percent), and selecting regressors after looking at the results (24 percent). Finally, around 26 percent 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).[28]

Next, though imprecisely estimated, we find some evidence that authors with a greater number of publications and with a top five publication are more likely to stop or withhold their study after finding a null result. Assuming well-published authors write better papers on average, this suggests higher-quality null result papers are less likely to be submitted for review compared to their lower-quality counterparts. As such, our estimates likely understate the true "filtering out" effect by editors since these higher-quality papers would be more likely to get past the desk. In doing so, this would further smooth the distribution of non-desk-rejected papers.

To address the generalizability of our findings, we compare differences in author behavior for those who submitted to the *Journal of Human Resources* relative to other journals as well as other authors in the same field. In the subsequent columns of Table 9, we report mean responses for authors whose research specialty was either public, labor, education, or health economics who submitted a paper to the *Journal of Human Resources* at least once in the prior five years, as well as 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, results show that author behavior at the *Journal of Human Resources* is largely consistent with overall author behavior in the field as well as with behavior of authors who submitted to other journals. Hence, this suggests that our findings cannot likely be explained by a unique set of authors who engage in differential behavior at the *Journal of Human Resources* relative to authors submitting to other journals.

As is the case with most surveys, it's important to note that responses could be subject to several potential biases. For one, social desirability issues could drive respondents to underreport their p-hacking behaviors. Even though the survey was anonymous, respondents still could have been biased in their own recollection of whether they engaged in certain behaviors. Thus, it is likely that the "true" fraction of authors engaging in various behaviors is higher than reported in our survey.

VI. 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*

[28] Interpreting the coefficient, increasing a respondent's belief in publication bias along the 10-point scale by 1 unit increases the likelihood they stopped or withheld their study after finding null results by 8 percentage points.

Resources, a journal largely regarded as a top applied microeconomics journal. Test statistics were collected from a random sample of over 700 manuscripts submitted from the year 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 likely cannot be strictly attributed to the peer review process. Then, we find that papers sent for review display *less* bunching than desk-rejected papers; i.e., marginally statistically significant estimates are less likely to get past the desk. Anonymous reviewers, on the other hand, appear to be influenced by statistical significance: papers with (strong) positive recommendations are more likely to possess marginally significant results. In total, estimates from rejected manuscripts versus the final draft of accepted manuscripts display similar distributions. 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 percent 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.

Though our study is limited to a single journal, we find evidence that our results are likely generalizable to the broader profession. First, we find that manuscripts rejected at the *Journal of Human Resources* that never published (around 40 percent) have slightly fewer marginally significant estimates at the 10 percent level in favor of significantly more marginally significant estimates at the 5 percent level. This suggests the peer review phenomenon we identify is likely broadly applicable to peer review in the economics profession overall (or at least the journals authors may have submitted to *after* rejection from the JHR). Second, our survey of author behavior confirms that authors who submit to the JHR act in a similar fashion to those who submit to other journals. That is, a large set of 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 nonsignificant results from journal submission (Franco, Malhotra, and Simonovits 2014).

Further work could shed additional light on this problem by investigating "papers" that are never observed due to author behavior in response to a belief of publication bias. Furthermore, our study does not distinguish between statistical significance and economic significance. In particular, a sufficiently powered study could estimate a "zero" effect while maintaining statistical significance (i.e., a "precisely estimated zero"). Finally, our study does not identify the "causal" effect of marginal statistical significance on peer review decisions since papers that attain marginal significance may differ in (unobserved) ways to those that are insignificant.[29] Still, our estimates

[29] Of course, identifying causality of marginal statistical significance is exceptionally challenging since it would require two groups of reviewed papers that are identical in all manners, except one happens to have statistical significance (i.e., different point estimates and/or standard errors), while the other does not.

show that for the papers that are submitted for peer review, the peer review process does not appear to be heavily biased in favor of statistical significance (and thus, cannot be the sole explanation of observed statistical bunching among published statistics).

REFERENCES

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

Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review* 109 (8): 2766–94.

Ashenfelter, Orley, Colm Harmon, and Hessel Oosterbeek. 1999. "A Review of Estimates of the Schooling/Earnings Relationship, with Tests for Publication Bias." *Labour Economics* 6 (4): 453–70.

Blanco-Perez, Cristina, and Abel Brodeur. 2020. "Publication Bias and Editorial Statement on Negative Findings." *Economic Journal* 130 (629): 1226–47.

Brodeur, Abel, Scott Carrell, David Figlio, and Lester Lusher. 2023. "Replication data for: Unpacking P-Hacking and Publication Bias." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E192271V1.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review* 110 (11): 3634–60.

Brodeur, Abel, Nikolai M. Cook, Jonathan S. Hartley, and Anthony Heyes. 2022. "Do Pre-registration and Pre-analysis Plans Reduce P-Hacking and Publication Bias?" IZA Discussion Paper 15476.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics* 8 (1): 1–32.

Bruns, Stephan B., Igor Asanov, Rasmus Bode, Melanie Dunger, Christoph Funk, and Sherif M. Hassan. 2019. "Reporting Errors and Biases in Published Empirical Findings: Evidence from Innovation Research." *Research Policy* 48 (9): 103796.

Card, David, and Stefano Dellavigna. 2020. "What do Editors Maximize? Evidence from Four Economics Journals." *Review of Economics and Statistics* 102 (1): 195–217.

Card, David, Stefano Dellavigna, Patricia Funk, and Nagore Iriberri. 2020. "Are Referees and Editors in Economics Gender Neutral?" *Quarterly Journal of Economics* 135 (1): 269–327.

Carrell, Scott E., David N. Figlio, and Lester R. Lusher. 2022. "Clubs and Networks in Economics Reviewing." NBER Working Paper 29631.

Casey, Katherine, Rachel Glennerster, and Edward Miguel. 2012. "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan." *Quarterly Journal of Economics* 127 (4): 1755–1812.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma. 2020. "Simple Local Polynomial Density Estimators." *Journal of the American Statistical Association* 115 (531): 1449–55.

Christensen, Garret, and Edward Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature* 56 (3): 920–80.

Cox, Gregory, and Xiaoxia Shi. 2022. "Simple Adaptive Size-Exact Testing for Full-Vector and Subvector Inference in Moment Inequality Models." *Review of Economic Studies* 90 (1): 201–28.

De Long, J. Bradford, and Kevin Lang. 1992. "Are all Economic Hypotheses False?" *Journal of Political Economy* 100 (6): 1257–72.

DellaVigna, Stefano, and Elizabeth Linos. 2022. "RCTs to Scale: Comprehensive Evidence from Two Nudge Units." *Econometrica* 90 (1): 81–116.

Doucouliagos, Chris, and T. D. Stanley. 2013. "Are All Economic Facts Greatly Exaggerated? Theory Competition and Selectivity." *Journal of Economic Surveys* 27 (2): 316–39.

Elliott, Graham, Nikolay Kudrin, and Kaspar Wüthirch. 2022. "Detecting p-Hacking." *Econometrica* 90 (2): 887–906.

Ferraro, Paul J., and Pallavi Shukla. 2020. "Is a Replicability Crisis on the Horizon for Environmental and Resource Economics?" *Review of Environmental Economics and Policy* 14 (2): 339–51.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345 (6203): 1502–05.

Frankel, Alexander, and Maximilian Kasy. 2022. "Which Findings Should Be Published?" *American Economic Journal: Microeconomics* 14 (1): 1–38.

Furukawa, Chishio. 2020. "Publication Bias under Aggregation Frictions: From Communication Model to New Correction Method." Unpublished.

Gerber, Alan S., and Neil Malhotra. 2008a. "Do Statistical Reporting Standards Affect what is Published? Publication Bias in Two Leading Political Science Journals." *Quarterly Journal of Political Science* 3 (3): 313–26.

Gerber, Alan S., and Neil Malhotra. 2008b. "Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?" *Sociological Methods and Research* 37 (1): 3–30.

Havránek, Tomás. 2015. "Measuring Intertemporal Substitution: The Importance of Method Choices and Selective Reporting." *Journal of the European Economic Association* 13 (6): 1180–1204.

Havránek, Tomás, 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, Tomás, T. D. Stanley, Hristos Doucouliagos, Pedro Bom, Jerome Geyer-Klingeberg, Ichiro Iwasaki, W. Robert Reed, and Katja Rost. 2020. "Reporting Guidelines for Meta-Analysis in Economics." *Journal of Economic Surveys* 34 (3): 469–75.

Huber, Jürgen, Sabiou Inoua, Rudolf Kerschbamer, Christian König-Kersting, Stefan Palan, and Vernon L. Smith. 2022. "Nobel and Novice: Author Prominence Affects Peer Review." *Proceedings of the National Academy of Sciences* 119 (41): 1–7.

Ioannidis, John P. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8): 696–701.

Ioannidis, John P., T. D. Stanley, and Hristos Doucouliagos. 2017. "The Power of Bias in Economics Research." *Economic Journal* 127 (605): F236–65.

Kranz, Sebastian, and Peter Putz. 2022. "Methods Matter: P-hacking and Publication Bias in Causal Analysis in Economics: Comment." *American Economic Review* 112 (9): 3124–36.

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73 (1): 31–43.

Lybbert, Travis J., and Steven T. Buccola. 2021. "The Evolving Ethics of Analysis, Publication, and Transparency in Applied Economics." *Applied Economic Perspectives and Policy* 43 (4): 1330–51.

McCloskey, Donald N. 1985. "The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests." *American Economic Review: Papers and Proceedings* 75 (2): 201–05.

Miguel, E., C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, et al. 2014. "Promoting Transparency in Social Science Research." *Science* 343 (6166): 30–31.

Ofosu, George K., and Daniel 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 (3): 309–45.

Stanley, T. D. 2008. "Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection." *Oxford Bulletin of Economics and Statistics* 70 (1): 103–27.

Vivalt, Eva. 2019. "Specification Searching and Significance Inflation Across Time, Methods and Disciplines." *Oxford Bulletin of Economics and Statistics* 81 (4): 797–816.