Journal of Accounting Research

DOI: 10.1111/1475-679X.12579
Journal of Accounting Research
Vol. No. xxxx 2024
Printed in U.S.A.

CHICAGO BOOTH

Discontinuous Distribution of Test Statistics Around Significance Thresholds in Empirical Accounting Studies

XIN CHANG,* HUASHENG GAO,† AND WEI LI‡

Received 15 August 2021; accepted 14 June 2024

ABSTRACT

Examining test statistics from articles in six leading accounting journals, we detect discontinuities in their distributions around conventional significance thresholds (p-values of 0.05 and 0.01) and find an unusual abundance of test statistics that are just significant. Further analysis reveals that these discontinuities are more prominent in studies with smaller samples and are more

*Nanyang Business School, Nanyang Technological University; †International School of Finance, Fudan University; ‡Department of Accountancy, City University of Hong Kong

Accepted by Luzi Hail. We are grateful for the insightful comments provided by an anonymous associate editor, two anonymous referees, Robert Bloomfield, Andrew Burton-Jones, Lei Chen, William Cready, Yun Dai, Michael Dougherty, Victor van Pelt, Christian Peters, Shyam Sunder, Hun Tong Tan, Bernard Yeung, Lu Zhang, Hailan (Flora) Zhou, Hua Zhou, seminar participants from the Central University of Finance and Economics, China Europe International Business School, Fudan University, Jilin University of Finance and Economics, Nanyang Technological University, the University of Adelaide, Shanghai University of Finance and Economics, Southern University of Science and Technology, Sun Yat-sen University, and Xiamen University, and participants in the First Fanhai Economics and Finance Workshop and the 2019 MIT Asia Conference in Accounting. We acknowledge financial support from the Ministry of Education (Singapore) (Tier 1 Research Integrity Grant No. RGI03/19), the National Natural Science Foundation of China (Grant Nos. 71973029, 72472028), and the 111 Project (B20094). Any errors are our own. The authors have no conflict of interest to declare. An online appendix to this paper can be downloaded at https://www.chicagobooth.edu/jar-online-supplements.

salient in experimental than in archival studies. The discontinuity discrepancy between experimental and archival studies relates to several proxies for researcher degrees of freedom. Nevertheless, this evidence does not imply that experimental research is more prone to questionable research practices than archival studies. Overall, our findings speak to the concern of whether accounting researchers could exercise undisclosed discretion to obtain and report statistically significant results. Based on our results, a healthy skepticism of some just-significant test statistics is warranted.

JEL codes: C12, C40, M40, M41, M42

Keywords: *P*-value discontinuity; experimental accounting; archival accounting; researcher degrees of freedom

1. Introduction

As competition for publication space in top journals intensifies, researchers are tempted to shift their focus from generating scientific findings to meeting publication standards (Harvey [2017]). Chief among the standards for publication is statistical significance (Fanelli [2012]), commonly determined using test statistics (e.g., *p*-values and *t*-statistics), which reflect the strength of the evidence against the null hypothesis. In search of greater statistical significance (e.g., lower *p*-values), empirical researchers have considerable freedom in data analysis and reporting (Simmons, Nelson, and Simonsohn [2011], Krawczyk [2015]).[1] To enhance the chance of publication, some researchers may exploit undisclosed discretion until insignificant results become significant. Such questionable research practices (QRPs) can cause discontinuity or "bunching" of test-statistic distributions (Masicampo and Lalande [2012]): an unusual abundance of just-significant test statistics around conventional significance thresholds (e.g., *p*-values of 0.05 and 0.01) in published studies (Brodeur, Cook, and Heyes [2020]). However, we are quick to add that many just-significant test statistics stem from legitimate research practices instead of QRPs and that not all QRPs are reflected in the bunching of test statistics around conventional thresholds (Bishop and Thompson [2016]). Thus, the overabundance of just-significant test statistics serves only as suggestive evidence of some researchers potentially exercising discretion to enhance statistical significance.

This paper examines the distribution of test statistics reported in top accounting journals to address three questions: (1) What is the extent of

[1] For example, researchers can exercise their discretion in terms of the tests used, sample size, the selection and construction of outcome variables, how to deal with outliers, the choice of control variables (covariates), and decimal points to include in test statistics. We stress that the use of researcher discretion is often rooted in sound empirical analysis, rather than merely aiming to boost statistical significance in a self-serving manner. The questionable research practice of running many tests and analyses until insignificant results become significant is known by various names, including data dredging, significance chasing, target beating, and *p*-hacking.

discontinuity in test statistics in the accounting literature? (2) How does the extent of discontinuity differ between experimental and archival accounting research?[2] (3) How does the extent of discontinuity in test statistics relate to sample sizes and researcher degrees of freedom in data analysis and reporting? To answer these questions, we construct a large sample of test statistics from both experimental and archival accounting studies published from 1990 to 2020 in six leading accounting journals: *the Journal of Accounting Research (JAR), The Accounting Review (TAR), the Journal of Accounting and Economics (JAE), the Review of Accounting Studies (RAST), Contemporary Accounting Research (CAR),* and *Accounting, Organizations and Society (AOS).*

For experimental studies, which primarily report *p*-values, our distribution analysis reveals an unusually high number of *p*-values immediately below each of the two essential significance thresholds (i.e., 0.05 and 0.01). After constructing unbiased counterfactual *p*-value distributions, we find that the frequencies of reported experimental *p*-values at 0.01 and 0.05 are 22.8% and 19.1% higher than expected, respectively. When interpreting these statistics, an important caveat is that they only apply locally to *p*-values just below *p*-value significance thresholds. For archival studies, which mainly report *t*- or *z*-statistics and standard errors, we follow Brodeur et al. [2016] and transform all archival test statistics into *z*-statistics. By comparing the observed and expected *z*-statistic distributions, we find a noticeable underrepresentation of *z*-statistics for $z < 1.96$ (i.e., the 5% threshold) and overrepresentations of *z*-statistics for $z \geq 1.96$ or $z \geq 2.58$ (i.e., the 1% threshold). In summary, both experimental and archival studies exhibit statistically significant discontinuities in the distribution of test statistics around significance thresholds. Section 5.2 compares the discontinuities in test statistics between accounting and other fields.

Next, we examine whether the degree of discontinuity of test statistics differs between archival and experimental studies. A priori, which research methodology should generate more just-significant test statistics around significance thresholds is unclear. On the one hand, Head et al. [2015] argue that the research fields conducting laboratory experiments may have more researcher freedom than archival research, especially in the data collection, sample construction, and result reporting stages. Harvey [2017]

---

[2] Both archival and experimental research are empirical in nature. Experimental studies rely on data gathered through introducing treatments/interventions to participants (subjects), who are randomly allocated to control and treatment groups, whereas archival research uses observational data collected by researchers without interacting with research participants. Similar to experimental studies, which primarily focus on causal relations, most archival accounting studies also make causal claims in their main findings. By reading 100 randomly selected archival studies in our sample, we find that 83 articles draw causal inferences using various tests, including (quasi-)natural experiments, difference-in-differences, instrumental variables, Heckman correction procedures, and event studies. For the three identification strategies that have gained popularity in recent years (Brodeur, Cook, and Heyes [2020]), we identify 13 articles using difference-in-differences and eight papers relying on instrumental variables, but no articles employing regression discontinuity designs in the random sample.

argues that unlike most archival researchers using data readily available to other researchers (e.g., Compustat), experimentalists typically create the data for their research, resulting in a lower likelihood of discretionary actions being detected using the same data.[3] Moreover, most experimental studies have small samples. When a treatment effect truly exists, smaller sample sizes lower the likelihood of detecting the effect, resulting in larger expected *p*-values (Sackrowitz and Samuel-Cahn [1999]). In response, experimentalists may need to exercise greater discretion to obtain statistically significant results. Wicherts et al. [2016] also argue that smaller samples lead to greater sampling variability, making each analytical choice more influential. Therefore, leveraging researcher degrees of freedom to secure statistically significant outcomes can be more effective with smaller sample sizes. Collectively, these arguments imply that the test statistics of experimental studies may exhibit more bunching around significance thresholds than those of archival studies.

On the other hand, several arguments suggest that archival studies should have more discontinuous test statistics. Compared with experimental researchers, archival researchers may have more freedom in data analysis, especially in choosing regression specifications and variables. Mitton [2022] documents that archival finance researchers use considerable discretion in measuring corporate outcome variables (e.g., profitability, firm value, and investments) and selecting control variables. This finding is relevant to archival accounting research, given its overlap with financial research in outcome variables and research methods. Further, Brodeur et al. [2016] document that the distribution of test statistics published in three prestigious economics journals exhibits evident discontinuity around *p*-values of 0.05 and 0.10. However, such discontinuity is not discernible in articles using laboratory experiments and randomized control trials (RCTs). Relatedly, using test statistics generated by various causal identification methods in top economics journals, Brodeur, Cook, and Heyes [2020] reveal that studies utilizing RCTs—most of which are experimental—have less pronounced bunching of test statistics around conventional thresholds than those using instrumental variables (IV), difference-in-differences (DID), and regression discontinuity design (RDD) methods, most of which are archival. Based on these arguments and findings, one would expect archival researchers to take more discretionary actions to achieve statistical significance, resulting in more discontinuous test statistics.[4]

---

[3] Thus, some researchers may, for example, decide how to report their experimental conditions and whether to collect additional data after observing the initial results for the sake of enhancing statistical significance (Khan and Tronnes [2019]).

[4] Relatedly, surveying the invitees of the 2019 JAR Conference about their perceptions of the reproducibility of accounting research, Hail, Lang, and Leuz [2020] find that "*respondents were of two minds on whether archival research is more or less likely to replicate than behavioral or experimental research*" (p. 525). Some respondents believe that archival researchers' discretion in

To shed light on these issues, we compare the degree of discontinuity between experimental and archival studies in three ways. First, we identify excess test statistics at conventional significance thresholds based on the difference between the observed and expected distributions of test statistics. We find that excess test statistics are more frequent in experimental studies than in archival studies at all conventional levels of significance. Second, we convert experimental *p*-values to *z*-statistics and compare them with archival studies' *z*-statistics. Using *z*-statistics within narrow windows (± 0.1, ± 0.2, and ± 0.3) containing conventional significance thresholds, we calculate the differences in the proportion of significant test statistics between experimental and archival studies for all test-statistic windows. The results reveal that compared with archival studies, experimental studies report a significantly larger proportion of test statistics that meet or beat the statistical threshold within each window. Third, we use the Brodeur, Cook, and Heyes [2020] regression framework (i.e., caliper tests) to model the likelihood of significant test statistics being reported within a narrow range around conventional thresholds. This approach enables us to control for various paper and author characteristics. The results show that experimental *z*-statistics are significantly more discontinuous than archival ones around all significance thresholds. For example, for *z*-statistics in the 1.96 ± 0.1 interval (roughly equivalent to *p*-values in the 0.04–0.06 range), the likelihood of experimental accounting studies reporting test statistics that beat the 5% threshold is 21.2 percentage points higher than that for archival studies (55.1 percentage points). Overall, our comparisons reveal that discontinuities in the distribution of test statistics around significance thresholds are more pronounced in experimental studies than in archival studies.

Next, we explore the relation between sample size and discontinuities in test statistics by separately dividing archival and experimental studies into terciles according to their sample sizes. Caliper tests reveal that accounting studies with smaller samples display more discontinuous test statistics around significance thresholds. In particular, small-sample experimental studies are more likely to report test statistics that beat the 5% threshold than those with large samples. For archival studies, small samples exhibit higher likelihoods of beating the 1% significance threshold than medium and large samples.

Lastly, we investigate whether researcher degrees of freedom in data analysis and reporting are related to the differences in test-statistic distributions between the two types of accounting research. Empirically, we measure experimental researchers' discretion along four dimensions: (1) the number of experiment constructs, (2) the use of one-tailed or two-tailed tests, (3) the type of experiment participants, and (4) whether *p*-values are rounded to two decimal places. We find that the differences between experimental and archival studies widen when researchers' discretionary actions are

---

selectively reporting results makes published findings less replicable, whereas others reckon that experimental studies' small samples worsen irreproducibility.

more likely to result in significant test statistics, confirming the relevance of researcher degrees of freedom for the discontinuities in accounting test statistics around significance thresholds.

Two important caveats are in order. First, our finding that experimental studies have more discontinuous test statistics than archival ones should not be interpreted as evidence of experimentalists engaging more in QRPs than archival researchers. Our approach does not observe or detect QPRs. We only examine distributions of test statistics that may (or may not) be consistent with significance-seeking QPRs. Bishop and Thompson [2016] show that when QRPs involve including various dependent variables and reporting only those with $p \leq 0.05$, the p-value distribution depends on the correlations among the dependent variables attempted by researchers. Specifically, the simulated p-values display bunching just below 0.05 when the dependent variables are intercorrelated but not when they are uncorrelated. This finding implies that the lack of a bump in the p-value distribution does not necessarily imply no QRPs. Furthermore, in a blog article, Simonsohn [2020] defines slow (fast) p-hacking as opportunistic practices that change the p-value slightly (substantially) from analysis to analysis. His simulations show that only slow p-hacking causes bunching just above a z-statistic significance threshold because fast p-hacking shifts z-statistics outside the narrow windows around the threshold.[5] Thus, our caliper tests, which focus on narrow windows around significance thresholds, may fail to capture fast p-hacking sufficiently. Given the nature of the data, archival (experimental) research can be more subject to fast (slow) p-hacking.

Second, legitimate research practices often generate test statistics that happen to be just-significant. While the overabundance of just-significant test statistics suggests that some researchers may exercise discretion to achieve statistical significance, our analysis cannot separate just-significant test statistics caused by QRPs from those engendered by legitimate practices. In addition, among all QRPs, some QRP can be egregious, such as misreporting test statistics, censoring unfavorable data, or employing various dependent variables and selectively reporting them after analysis. Others may be relatively benign, reflecting researchers' human default. For example, ceasing data collection after finding a significant result may reflect confirmation bias—the tendency to search for evidence that confirms one's prior beliefs while ignoring unsupportive evidence and data.[6] Our

[5] For example, slow p-hacking includes removing several observations, transforming variables, and changing the length of an event window, whereas fast p-hacking includes trying different dependent variables and comparing alternative pairs of experiment condition to obtain significant results. Simonsohn expects that slow (fast) p-hacking should be more common in archival (experimental) studies. However, Brodeur and his coauthors argue the reverse based on several reasons, including that RCTs are more collaborative and are often preregistered. More detailed discussions can be found in the blog article https://datacolada.org/91.

[6] Simmons argues that if a researcher strongly believes a hypothesis, self-serving reasoning may lead them to equate the most significant result with the one they initially conjectured (https://www.wired.com/story/were-all-p-hacking-now/).

tests, however, cannot differentiate between the different motivations underlying the just-significant test statistics.

Our study contributes to the literature in two ways. First, our findings add to the burgeoning literature examining the distributions of reported test statistics in published studies. This strand of literature generally presents evidence consistent with the distribution patterns that one would expect if researchers engage in QRPs, which impede scientific progress in the long term by increasing false-positive rates. False positives may stimulate investments in fruitless research programs, motivate costly policies based on unwarranted findings, or even discredit an entire research field (Head et al. [2015]). Growing concerns over QRPs have accelerated research on the subject in many disciplines, including economics, finance, psychology, and medicine.[7] However, distribution patterns consistent with potential QRPs remain relatively underexplored in accounting. Given the substantial variations in QRPs across disciplines (Fanelli [2012], Head et al. [2015]), our research should be useful for various stakeholders who rely on accounting research for policymaking, investment decisions, education, and further scientific endeavors.

Our study is not the first to examine abnormal test-statistic distributions in accounting research. Basu and Park [2014] examine the distribution of *p*-values in all empirical accounting papers published in the top three accounting journals in 2011. They show that the frequency of *p*-values just below conventional significance thresholds is higher than that just above the thresholds. However, their sample includes only seven *p*-values of around 0.05 from experimental articles. Khan and Tronnes [2019] study the *p*-values reported in experimental auditing research articles published in top accounting and auditing journals and find that the number of *p*-values $\leq$ 0.05 or 0.10 is unusually large. Nevertheless, their sample does not include any archival studies. Relative to these previous studies, we extend the scope from focusing on a particular research topic or year to covering all topics in top accounting journals from 1990 to 2020, allowing us to take a panoramic view of the reported test statistics in both experimental and archival studies. In addition, we study the difference in the distribution of test statistics between experimental and archival accounting research and compare accounting with other fields regarding discontinuities in test statistics. Moreover, we show that discontinuities in test statistics around significance thresholds relate to sample sizes and researcher degrees of freedom.

Second, our study adds to the discourse on the reproducibility of accounting research. Hail et al.'s [2020] survey shows that most accounting researchers believe that the irreproducibility of accounting research is common but receives insufficient attention. If the statistical significance of some published findings is achieved through researchers exercising undisclosed

[7] See Simmons, Nelson, and Simonsohn [2011], John, Loewenstein, and Prelec. [2012], Brodeur et al. [2016], Adda, Decker, and Ottaviani [2020], and Chen [2021].

discretion, there should be a high chance that these findings are irreproducible. Thus, our analysis contributes to evidence-based discussions about the credibility of accounting research through the lens of the abnormal distribution of reported test statistics. However, because we are agnostic about the approaches that can deliver precise estimates of QRPs, we urge readers to exercise caution in interpreting our findings and avoid viewing discontinuities in the test-statistic distribution as a necessary condition for QRPs in accounting.

2. *Sample, Test Statistics, and Descriptive Statistics*

2.1 SAMPLE CONSTRUCTION

Our sample includes experimental and archival studies published between 1990 and 2020 in the six most influential accounting journals: *AOS, CAR, JAE, JAR, RAST,* and *TAR*.[8] Our sample period begins in 1990, when the Naveen Jindal School of Management at the University of Texas at Dallas (UTD) started to construct a database to track publications in leading business journals. We use this database to collect author and article characteristics.

Experimental studies emphasize treatment effects and typically report *p*-values instead of *t*- or *F*-statistics, which archival studies often report.[9] This reporting choice facilitates our data collection and analysis of experimental studies. Using a Python script to search all 6,459 articles published in the six journals, we identify an article as experimental if its abstract contains the keyword "experiment" but not "(quasi-) natural experiment." In total, 708 articles (about 11% of all articles) meet these criteria and are downloaded for our text-mining procedure to extract *p*-values.

The six accounting journals publish much more archival studies than experimental studies. Therefore, for each experimental paper, we randomly select an archival article published in the same year in one of the six journals, generating a total of 708 archival articles. Unlike experimental studies, archival studies typically perform multivariate regressions for hypothesis testing. Thus, our text-mining approach for collecting experimental *p*-values does not apply to archival studies. Instead, we read each archival article to identify its main hypothesis and manually collect the

[8] The SCImago Journal Rank (SJR) measures a journal's impact based on the average number of weighted citations received in a year across all articles published in the previous three years. As of December 2022, the following six accounting journals had the highest SJR indicators among all accounting journals: *JAE* (7.346), *JAR* (5.922), *TAR* (4.674), *RAST* (3.998), *CAR* (3.017), and *AOS* (2.204) (https://www.scimagojr.com/journalrank.php?category=1402).

[9] The *t*- or *F*-distributions are defined by degrees of freedom, which are often not disclosed in published articles and are difficult to infer, with various fixed effects in regressions. Therefore, a single reported *p*-value can be associated with different *t*- or *F*-statistics, and a reported *t*-statistic can correspond to different *p*-values.

corresponding test statistics.[10] Among the initial sample of 1,416 accounting studies, 603 archival and 653 experimental studies (1,256 articles in total) report test statistics and thus are retained in our final sample. Table A1 in the appendix tabulates the number of articles by year and by journal. *CAR*, *JAR*, and *TAR* collectively contribute more than 68% of the articles to our final sample, primarily because they publish more experimental articles than the other three journals.

2.2 TEST STATISTICS

We rely on different approaches to collect test statistics from experimental and archival studies. First, experimental researchers typically conduct the analysis of variance (ANOVA) and denote statistical significance using *p*-values for key results in the main text in a consistent manner.[11] We thus adopt a text-mining approach to extract *p*-values from experimental articles. Specifically, after transforming all downloaded pdf files into text format, we use a Python script to search for figures between zero and one that are immediately preceded by "$p =$," "p-value $=$," "$p_{\text{two-tailed}} =$," or "$p_{\text{one-tailed}} =$," regardless of spacing or capitalization. Extending previous studies (e.g., Head et al. [2015], Khan and Tronnes [2019]), we collect vague *p*-values (i.e., *p*-values reported after inequality symbols, such as "$p <$" or "$p \leq$") and analyze them in section 5.1. Our script cannot retrieve *p*-values from tables because the table format in pdf files is disorganized and varies substantially from article to article, rendering the search impractical. Moreover, by focusing only on *p*-values mentioned in the text, we ensure they are not recorded twice in our sample. Further, researchers are more likely to discuss the significance levels of their main results than those of peripheral findings (e.g., coefficients of covariates) in the text. Thus, *p*-values discussed in the text suit our purpose of detecting researchers' behavior undertaken to beat the conventional significance thresholds for key findings.

Our initial search yields 13,873 *p*-values for all experimental studies. Because we are specifically interested in any irregularities in the *p*-value distribution around conventional significance levels, our baseline *p*-curve analysis for experimental studies in section 3.1 excludes very large *p*-values ($p \geq 0.155$) and vague *p*-values, resulting in 6,078 exact *p*-values between 0 and 0.155.[12]

---

[10] We find that the data in around 92.5% of the archival studies are collected from publicly available databases, such as Compustat, Center for Research in Security Prices, and International Brokers' Estimate System (IBES). The remainder (7.5%) primarily rely on proprietary data, such as the salary and performance of divisional managers in companies, and individual investors' stock trading transactions from brokerage houses.

[11] For example, a typical experimental article states the significance level for its hypothesis testing as "we find that *x* has a significant impact on *y* if there is a strong versus weak manipulation condition ($p = 0.024$)."

[12] Specifically, we exclude 3059 *p*-values ≥ 0.155 from the baseline *p*-curve analysis because they are less likely to reflect researchers' QRPs. Our robustness check (see figure IA4 in the

Two benefits of our text-scraping method are worth mentioning. First, the method does not require us to have detailed knowledge about the articles in our sample (e.g., topic of interest, hypotheses tested, or experimental design), thereby allowing us to examine articles on any accounting topics. Second, it mitigates potential biases from manual data collectors such as research assistants, who may gather *p*-values that conform to our conjectures. To verify that our script captures *p*-values related to the key results in experimental accounting studies, we randomly select 30 experimental articles from our sample and identify all main hypotheses tested. We find that these articles collectively report 260 exact *p*-values, 251 (96.1%) of which are related to the significance levels of the main hypotheses tested. The remaining nine *p*-values belong to six articles, none of which discuss *p*-values for the main hypotheses. This verification indicates that our script can extract the vast majority of exact *p*-values that substantiate the main results in the experimental articles.

Given that archival studies primarily test their hypotheses using multivariate regressions and do not often report *p*-values, we read every archival article and manually collect data related to the main tests. In particular, we extract the reported coefficients, test statistics (i.e., *z*-statistics, *t*-statistics, standard errors, or *p*-values), and number of observations. If the main hypothesis is tested using between-group *t*-tests, we compute the tested differences and record them as coefficients. For tests based on the IV approach, we gather only the coefficients and test statistics for the second-stage regression. For DID tests, we collect the coefficients and test statistics of the primary interaction term and dynamic treatment effects. Our data collection does not include test statistics from auxiliary tests, robustness checks, and cross-sectional and mechanism/channel analyses. This procedure results in 9,010 test statistics, of which 71.4% are *t*-statistics, 14.9% are *p*-values, 7.8% are standard errors, and 5.9% are *z*-statistics from archival studies.

To make the different test statistics comparable, we follow Brodeur et al. [2016] and Brodeur, Cook, and Heyes [2020] and convert all test statistics to *z*-statistics. Specifically, *p*-values are transformed directly into equivalent *z*-statistics (e.g., when $p = 0.05$, $z = 1.96$). For papers reporting standard errors, we compute the ratio between the reported coefficient and the standard error. Because the degrees of freedom for most archival tests are unreported, we assume that *t*-statistics and the ratio of the coefficient to its standard error follow asymptotically standard normal distributions under the null hypothesis and make them equal to *z*-statistics (Brodeur et al. [2016]).[13]

---

online appendix) shows that our baseline results are basically unaffected if they are included in the *p*-curve analysis. The number of vague *p*-values is 4721. The analysis in section 5.1 reveals stronger evidence of discontinuous test statistics when vague *p*-values are included.

[13] Brodeur et al. [2016] also point out that unlike the standard distribution of *z*-statistics, the distribution of *t*-statistics depends on the degrees of freedom. Therefore, when the sample size is small, a *t*-statistic of 1.97 treated as a *z*-statistic may be inadequate to reject the null hypothesis

2.3 AUTHOR AND ARTICLE CHARACTERISTICS

We collect a set of author and article characteristics that may relate to researchers' incentives and opportunities to exercise discretion. We define *Authors* as the number of authors of each article. To account for gender differences among researchers related to research practices (e.g., Friesen and Gangadharan [2012]), we define *D*(*Female Author*) as an indicator variable that equals one if at least one author is female and zero otherwise. To measure authors' research experience, we count the number of years between PhD completion and article publication for each author and define *Experience* as the average value across all authors of an article.[14] To measure the academic reputation of authors' schools, we compute the proportion of authors affiliated with the top 20 business schools (*Top Institution*) and those who graduated from the top 20 business schools (*Top PhD*). We obtain school rankings from UTD's Top 100 Business School Research Rankings based on the number of publications in the top three accounting journals (*JAE, JAR*, and *TAR*) in a given year. We use the school ranking two years before the article's publication year for each article–author–school combination because Ellison [2002] and Wood [2016] document that the average time from a paper's initial submission date to its acceptance date is about two years for the top three accounting journals. By doing this, we aim to capture a school's ranking status while a paper is being written rather than when it is published. If a school is not ranked, we set its rank value equal to the lowest rank value on the ranking list in a given year.

Moreover, we count the number of test statistics (*Test Stats*) extracted from each article. Researchers' incentives to exercise discretion may be stronger for the main results, which determine the likelihood of an article being published (Krawczyk [2015]). The main results are typically presented before sections such as robustness checks or additional analyses. Thus, we use an indicator variable, *D*(*Main Results*), to indicate the first half of test statistics by their order of appearance in an article. For an experimental study, we retrieve the total number of participants (*Experiment Participants*) in its experiments. For an archival study, we record the average number of observations across all regressions (*Archival Obs*). *D*(*Two-tailed*) is an indicator variable that equals one if an article reports only two-tailed tests and zero if it reports any one-tailed tests. Further, we use the indicator variable *D*(*Top 3*) to denote articles published in the top three accounting journals. Finally, we gather Google Scholar citations that an article has accumulated up to the time of data collection (*Citations*). *D*(*High Citation*) equals one if the citations for an experimental (archival) paper exceed the median number of citations of all experimental (archival) papers published in the same year and zero otherwise.

---

at the 5% significance level. However, this should not be a major concern for archival studies, which typically have large samples (i.e., larger than 30 observations).

[14] Alternatively, we use the proportion of senior authors (i.e., associate professor or above) as a proxy for research experience and obtain similar results.

Table 1 shows the descriptive statistics of author and article characteristics at the article level for the full sample (panel A), archival studies (panel B), and experimental studies (panel C). On average, accounting articles in our sample are written by 2.5 authors and contain 18 test statistics. Around 25.6% of authors are affiliated with the top 20 schools, and 39.3% graduated from the top 20 schools. The average research experience across all authors is about ten years. There are also some notable differences between the archival and experimental subsamples. For instance, compared with experimental studies, archival studies have a smaller proportion of female authors (44.1% versus 51.0%), contribute fewer test statistics to our sample (14.9 versus 21.2), have a larger proportion of authors affiliated with the top 20 institutions (29.3% versus 22.2%), attract more citations (290 versus 87), and report more two-tailed tests (69.1% versus 37.6%).

3. Baseline Results

In this section, we first present graphical evidence of *p*-value discontinuity in the reported test statistics for experimental and archival studies separately. We then examine the differences in discontinuities between the two types of accounting studies.

3.1 *P*-VALUE DISCONTINUITIES IN EXPERIMENTAL ACCOUNTING STUDIES

To detect researchers' efforts to turn insignificant *p*-values into significant ones, previous studies (e.g., Simonsohn and Nelson [2014]; Head et al. [2015]) employ a *p*-curve analysis, which plots the distribution of *p*-values reported in a large sample of independent studies. The logic is that in the absence of selective publication bias and researchers' discretionary actions to obtain and report favorable *p*-values, the shape of the *p*-curve primarily depends on the true effect size of the tested relation. If the true effect size is nonzero, the expected distribution of *p*-values should be exponential with a right skew because lower *p*-values occur more frequently than higher ones, as illustrated by the solid curve in chart A of figure 1. [15] Selective publication bias alters the shape of the *p*-curve because papers with statistically insignificant findings (e.g., $p > 0.05$) are either not submitted to or not accepted by journals. This should lead to a notable drop in the frequency of published *p*-values > 0.05, as shown by the dashed lines in chart A of figure 1. Further, researchers may exercise discretion to obtain statistically

[15] That is, the probability of researchers observing a certain *p*-value decreases exponentially and monotonically as the *p*-value increases. In addition, as the true effect size becomes stronger, the *p*-curve becomes more right skewed. If the true effect size is zero, *p*-values should be uniformly distributed between zero and one because every *p*-value is equally likely to be observed by researchers. The shape of the *p*-value distribution also depends on the priors of the tested hypotheses being true. In the online appendix, we use the Bayesian hypothesis-testing framework to show that as researchers' prior beliefs of the null hypothesis being true become stronger, the distribution of *p*-values becomes less right skewed.

TABLE 1
Summary Statistics

	N	Mean	SD	Min	Q_1	Median	Q_3	Max
Panel A: Full Sample								
Authors	1256	2.494	0.956	1.000	2.000	3.000	3.000	5.000
D(*Female Author*)	1256	0.477	0.500	0.000	0.000	0.000	1.000	1.000
Experience	1243	10.133	5.776	0.000	5.667	9.500	14.000	41.000
Top Institution	1256	0.256	0.362	0.000	0.000	0.000	0.500	1.000
Top PhD	1256	0.393	0.376	0.000	0.000	0.333	0.667	1.000
Test Stats	1256	18.217	14.420	1.000	8.000	15.000	25.000	91.000
D(*Two-tailed*)	1254	0.527	0.499	0.000	0.000	1.000	1.000	1.000
D(*Top 3*)	1256	0.579	0.494	0.000	0.000	1.000	1.000	1.000
Citations	1254	184.653	457.582	1	32	71	179	11,127
Panel B: Archival Studies								
Authors	603	2.459	0.949	1.000	2.000	3.000	3.000	5.000
D(*Female Author*)	603	0.441	0.497	0.000	0.000	0.000	1.000	1.000
Experience	602	9.855	5.268	0.000	5.667	9.500	13.333	30.500
Top Institution	603	0.293	0.383	0.000	0.000	0.000	0.500	1.000
Top PhD	603	0.421	0.383	0.000	0.000	0.333	0.667	1.000
Archival Obs	499	41,837	217,100	40	399	3,034	13,410	3,780,948
Test Stats	603	14.942	13.655	1.000	6.000	11.000	20.000	91.000
D(*Two-tailed*)	602	0.691	0.462	0.000	0.000	1.000	1.000	1.000
D(*Top 3*)	603	0.652	0.477	0.000	0.000	1.000	1.000	1.000
Citations	603	290	635	2	44	110	307	11,127
Panel C: Experimental Studies								
Authors	653	2.527	0.963	1.000	2.000	3.000	3.000	5.000
D(*Female Author*)	653	0.510	0.500	0.000	0.000	1.000	1.000	1.000
Experience	641	10.394	6.208	0.000	5.667	9.333	14.500	41.000
Top Institution	653	0.222	0.338	0.000	0.000	0.000	0.333	1.000
Top PhD	653	0.368	0.368	0.000	0.000	0.333	0.667	1.000
Experiment Participants	626	147	205	4	73	102	158	2,676
Test Stats	653	21.240	14.457	1.000	11.000	18.000	29.000	91.000
D(*Two-tailed*)	652	0.376	0.485	0.000	0.000	0.000	1.000	1.000
D(*Top 3*)	653	0.511	0.500	0.000	0.000	1.000	1.000	1.000
Citations	651	87	100	1	25	53	107	927

This table reports the descriptive statistics of the author and article characteristics at the article level for the full sample (panel A), archival studies (panel B), and experimental studies (panel C). *Authors* is the number of authors for a paper. *D*(*Female Author*) is a dummy variable that equals one if the author team contains at least one female author and zero otherwise. *Experience* is the average number of years from the year of authors' PhD graduation to the year of the article's publication. *Top Institution* (*Top PhD*) indicates the fraction of authors affiliated with (graduated from) the top 20 institutions based on their number of publications in the top three accounting journals (*JAE*, *JAR*, and *TAR*) two years before a paper's publication year. *Archival Obs* (*Experiment Participants*) is the number of regression observations in archival studies (participants in experimental studies). *Test Stats* is the total number of test statistics collected from a paper. *D*(*Two-tailed*) flags articles that report only two-tailed *p*-values. *D*(*Top 3*) indicates articles published in the *Journal of Accounting and Economics*, the *Journal of Accounting Research*, or *The Accounting Review*. *Citations* is the number of Google Scholar citations a paper had received at the time of our data collection in 2020.

Chart A: An illustration of *p*-value discontinuities around the significance threshold of $p = 0.05$

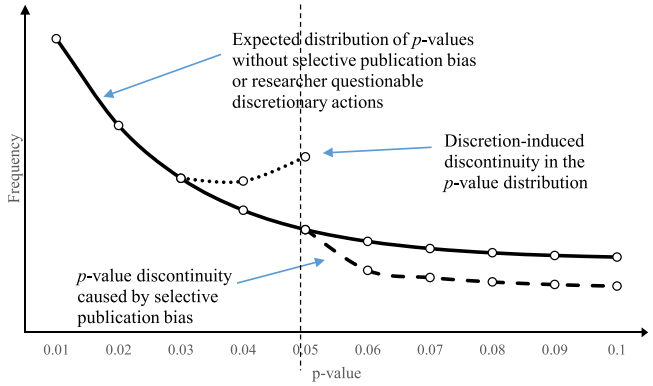

Chart B: The actual distribution of experimental p-values

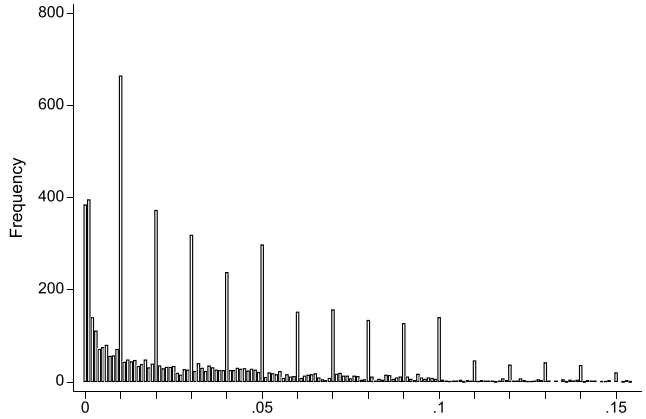

FIG. 1.—*P*-curve discontinuities and the actual distribution of experimental *p*-values. Chart A uses the significance level of 0.05 as an example to illustrate the potential discontinuities in the *p*-value distribution around $p = 0.05$. The solid curve represents the expected distribution of *p*-values when the effect size is nonzero and there is no selective publication bias or questionable discretionary actions by researchers that shift *p*-values from the insignificance region ($p > 0.05$) to the significance region ($p \leq 0.05$). The dashed line shows *p*-value discontinuity caused by selective publication bias, whereas the dotted line shows the discretion-induced discontinuity in *p*-value distribution. Chart B plots the actual distribution of 6,078 *p*-values from experimental studies with a bin width of 0.001. The *x*-axis represents *p*-values, and the *y*-axis represents frequencies.

significant results, shifting *p*-values from the nonrejection region ($p > 0.05$) to the rejection region ($p \leq 0.05$). Such practices should result in an overrepresentation of *p*-values just below 0.05 (see the dotted line in chart A of figure 1) and a noticeable drop in the number of *p*-values just above 0.05. To summarize, both selective publication bias and researchers' discretionary actions disturb the distribution of *p*-values and give rise to *p*-curve

discontinuities around conventional thresholds. However, only researchers' discretionary actions imply an overabundance of *p*-values in the tail of the distribution just below conventional thresholds (Head et al. [2015]).

Using *p*-values directly collected from experimental accounting studies, we examine the distribution of reported *p*-values by plotting a histogram based on all exact experimental *p*-values with a bin width of 0.001. We count the *p*-values in each bin and plot the frequency distribution in chart B of figure 1. Consistent with the *p*-curve with nonzero effect size in chart A, the frequency generally decreases as the *p*-value increases. About 44.7% of the *p*-values in our sample are reported to two decimal places, contributing to multiple frequency spikes in chart B.[16] There is a noticeable frequency spike in the interval immediately below the 5% significance threshold. The number of *p*-values between 0.049 and 0.050 is 299, about 24% higher than the number of *p*-values between 0.039 and 0.040 ($n = 242$). We observe similar discontinuities at the other two conventional significance thresholds ($p = 0.01$ and $p = 0.10$).[17]

To statistically quantify the frequency of excess experimental *p*-values just below significance thresholds, we follow Masicampo and Lalande [2012] and construct counterfactual distributions of *p*-values without potential bias.[18] By comparing the actual *p*-value distributions with the counterfactual ones, we can infer the magnitude of *p*-value discontinuity. Using Stata's curvefit command based on the least squares method, we first fit the histogram of reported *p*-values to an exponential distribution with a right skew (i.e., the trendline), which reflects the expected distribution of *p*-values with nonzero effect sizes. Because *p*-values are most frequently reported with two decimal places, we fit a trendline with all *p*-values rounded to two decimal places and present the results in chart A of figure 2. The exponential curve ($y = 1203.73e^{-21.18x}$) fits the *p*-value data well ($R^2 = 0.97$).

---

[16] Among the 6078 *p*-values, 20 are reported to one decimal place; 2715, 3022, 318, and three *p*-values are reported to two, three, four, and five decimal places, respectively.

[17] We conduct chi-square tests for differences in *p*-value frequencies across major *p*-value intervals. The results show that *p*-values in (0.049, 0.05] are significantly more frequent than those in (0.039, 0.04] (Pearson chi-square statistic = 6.006 and *p*-value = 0.014), *p*-values ($n = 665$) in (0.009, 0.01] are significantly more frequent than those ($n = 507$) in (0.000, 0.001] (Pearson chi-square statistic = 21.300 and *p*-value < 0.01), and *p*-values ($n = 141$) in (0.099, 0.10] are not significantly more frequent than those ($n = 128$) in (0.089, 0.09] (Pearson chi-square statistic = 0.628 and *p*-value = 0.428). The insignificant discontinuity around 0.10 is perhaps driven by the notion that $p = 0.10$ as a significance target is less important than $p = 0.05$. Supporting this interpretation, accounting researchers commonly label $p \leq 0.05$ as significant and $p \leq 0.10$ as marginally significant.

[18] Masicampo and Lalande [2012] extract 3627 *p*-values between 0.00 and 0.10 from three prominent psychology journals. They then use an exponential model that best fits the distribution of these *p*-values as the counterfactual distribution. Finally, they utilize chi-square analysis to compare the residual (the difference between observed and expected frequencies) between 0.045 and 0.05 with those in other intervals and compute excess *p*-values in the former interval.

Chart A: Rounding all *p*-values to two decimal places

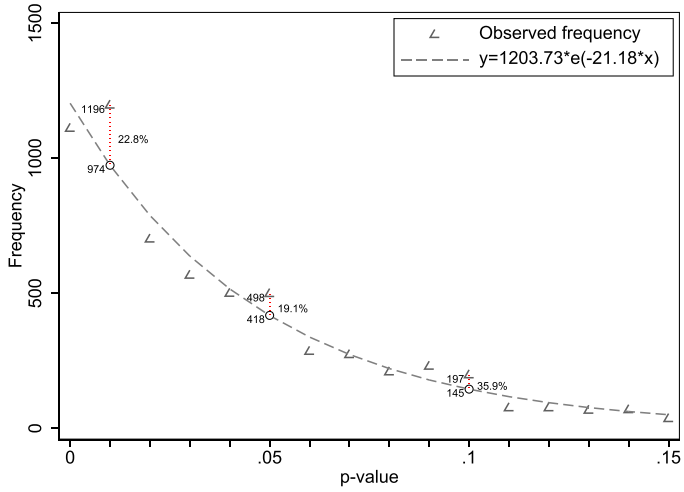

Chart B: Only including *p*-values reported to two decimal places

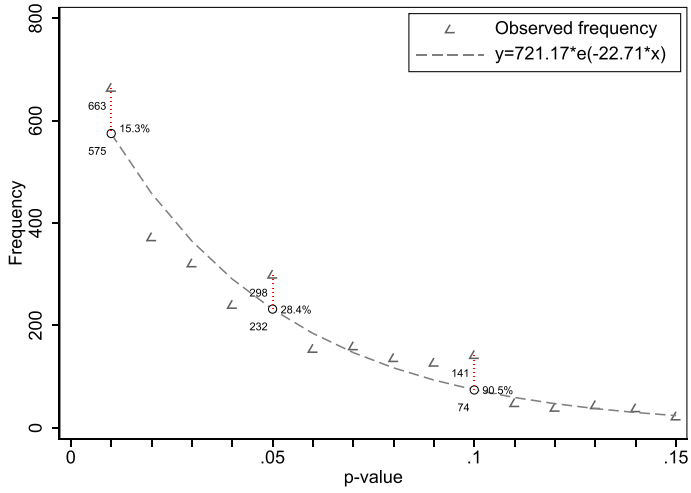

FIG. 2.—Fitting the histogram of experimental *p*-values to the exponential distribution. The dashed line represents the fitted exponential curve that minimizes the squared fitting errors. The triangles represent the frequencies of reported *p*-values. Chart A includes all *p*-values ($N = 6078$) rounded to two decimal places, and chart B includes *p*-values reported to two decimal places only ($N = 2{,}790$). The fitted exponential functions in charts A and B are $y = 1203.73e^{-21.18x}$ and $y = 721.17e^{-22.71x}$, respectively. The percentage frequency deviation for a *p*-value is the difference between the observed and expected frequencies implied by the fitted exponential curve divided by the expected frequency. The *x*-axis represents *p*-values, and the *y*-axis represents frequencies.

More importantly, we find that the frequencies of reported *p*-values at conventional thresholds (0.01, 0.05, and 0.10) sit well above the curve. We then compute the frequency deviation for a *p*-value (in percentage terms) as the difference between the observed and expected frequencies implied by the fitted exponential curve divided by the expected frequency. Chart A of figure 2 shows that the observed *p*-value frequencies at $p = 0.01$, 0.05, and 0.10, respectively, are 22.8%, 19.1%, and 35.9% higher than the expected frequencies. These excess test statistics are also in panel A of table 2. In addition, the frequencies of *p*-values slightly above conventional thresholds (0.02, 0.06, and 0.11) are well below the trendline, whereas the frequencies of *p*-values relatively more distant from conventional significance levels (e.g., $p = 0.03$, 0.07, and 0.12) are closer to those expected. This finding reflects that turning insignificant *p*-values more distant from conventional significance thresholds into significant ones may be more challenging. We conduct post hoc pairwise comparisons (i.e., Dunnett's test) and find that the frequency deviations of $p = 0.01$, 0.05, and 0.10 differ from the mean deviations of other *p*-values at the 1%, 5.7%, and 18.1% levels, respectively. We tabulate the comparison details in the online appendix. In chart B of figure 2, we retain only the *p*-values reported to two decimal places and repeat the curve-fitting analysis. Similar results ensue. In sum, the relative abundance of *p*-values at conventional thresholds and the relative scarcity of *p*-values slightly above conventional thresholds are consistent with researchers' target-beating behaviors.

3.2 DISCONTINUITIES IN Z-STATISTICS OF ARCHIVAL STUDIES

After transforming all reported test statistics into *z*-statistics to form a homogeneous sample (Brodeur et al. [2016]), we plot the histogram of *z*-statistics from archival studies in chart A of figure 3. We choose a bin width of 0.1 for *z*-statistics between 0 and 10 and highlight the conventional significance thresholds with reference lines (*z*-statistics = 1.65, 1.96, and 2.58 correspond to the 10%, 5%, and 1% significance levels, respectively). We observe an increase in frequency as the *z*-statistic exceeds 1.65 (i.e., the 10% significance threshold), consistent with significant results being more likely to be published. In addition, *z*-statistics cluster in the range of 1.65–3.00, and the frequency declines as the *z*-statistic increases above 3. The distribution exhibits a global maximum around $z = 2.00$, corresponding to *p*-values slightly less than 0.05. Further, a local minimum around $z = 1.40$ suggests that some *z*-statistics just below the 10% significance threshold may have been moved over the traditional significance threshold.

Next, we compare the observed distribution of *z*-statistics to a counterfactual distribution. In chart B of figure 3, we plot a *z*-curve by superimposing an Epanechnikov kernel density on the histogram in chart A, which smooths the observed distribution. In addition, we follow Brodeur, Cook, and Heyes [2020] and use a *t*-distribution as the counterfactual (or expected) distribution for archival *z*-statistics. Specifically, we calibrate a noncentral *t*-distribution that minimizes the $z > 5$ mass difference between the

TABLE 2
Univariate Analysis on Discontinuities in Test Statistics of Accounting Studies

Panel A: Excess Test Statistics around Significance Thresholds Based on the $p(z)$ Curve

	(1) Observed	(2) Expected	(3) = (1) − (2) Excess Significance	(4) = [(1) − (2)] / (2) % Excess Significance
Experimental *p*-values				
$p = 0.10$	197	145	52	35.9%
$p = 0.05$	498	418	80	19.1%
$p = 0.01$	1,196	974	222	22.8%
Archival *z*-statistics				
[0, 1.65)	0.317	0.365	−0.048	−13.2%
[1.65, 1.96)	0.078	0.086	−0.007	−8.4%
[1.96, 2.58)	0.169	0.141	0.027	19.4%
[2.58, 5)	0.282	0.257	0.025	9.7%

Panel B: Randomization Tests on the Proportion of Significant Test Statistics within a Narrow Range

	Experimental				Archival					
Windows	(1) *Total*	(2) *Sig*	(3) %	(4) *p*	(5) *Total*	(6) *Sig*	(7) %	(8) *p*	(9) = (3)−(7) *Diff*	(10) *z*-Statistics of *Diff* = 0
[1.96 ± 0.1]	910	642	0.705	0.000	439	242	0.551	0.018	0.154	5.585
[1.96 ± 0.2]	1445	880	0.609	0.000	795	415	0.522	0.114	0.087	3.989
[1.96 ± 0.3]	2211	1288	0.583	0.000	1186	644	0.543	0.002	0.040	2.218

(*Continued*)

TABLE 2—*(Continued)*

Panel B: Randomization Tests on the Proportion of Significant Test Statistics within a Narrow Range

	Experimental				Archival					
Windows	(1) *Total*	(2) *Sig*	(3) %	(4) *p*	(5) *Total*	(6) *Sig*	(7) %	(8) *p*	(9) = (3)−(7) *Diff*	(10) *z*-Statistics of *Diff* = 0
[2.58 ± 0.1]	763	643	0.843	0.000	374	215	0.575	0.002	0.268	9.861
[2.58 ± 0.2]	1022	761	0.745	0.000	694	379	0.546	0.008	0.199	8.546
[2.58 ± 0.3]	1579	884	0.560	0.000	1024	516	0.503	0.426	0.057	2.822

Panel A summarizes excess test statistics around significance thresholds based on the $p(z)$-curve analysis in figures 2 and 3 for experimental and archival studies, respectively. For experimental studies, the expected number of *p*-values is from the calibrated exponential distribution in panel A of figure 2. For archival studies, columns 1 and 2 report the mass of the area within each *z*-statistic region for the observed and expected distribution of *z*-statistics. The observed distribution is the *z*-curve obtained by superimposing an Epanechnikov kernel density on the histogram in chart A. The expected distribution is the calibrated non-central *t*-distribution plotted in chart B of figure 3. The mass is computed using the cumulative distribution function of the observed distribution $(\hat{F}(U) - \hat{F}(L))$ or the expected distribution $(F_{t(2,1.79)}(U) - F_{t(2,1.79)}(L))$, where U (L) represents the upper (lower) bound of a *z*-statistic region. Column 3 reports excess test statistics, which is the difference between the observed and expected test statistics. Column 4 shows excess test statistics as a fraction of expected test statistics. Panel B reports the results of randomization tests on the proportion of significant test statistics within a narrow range for experimental and archival studies. For this analysis, all experimental *p*-values are converted into *z*-statistics. *Total* is the number of test statistics in a *z*-statistic window. *Sig* is the number of test statistics that are significant at the conventional levels within a window. % denotes the proportion of significant test statistics within a window. *p* is the *p*-value from the binomial test on whether test statistics are equally likely to be significant and insignificant within a window. *Diff* is the difference in the proportion of significant test statistics between experimental and archival studies. *z*-statistics are from testing the equality of two proportions for each window using an asymptotically normally distributed test statistic.

Chart A: The distribution of archival *z*-statistics

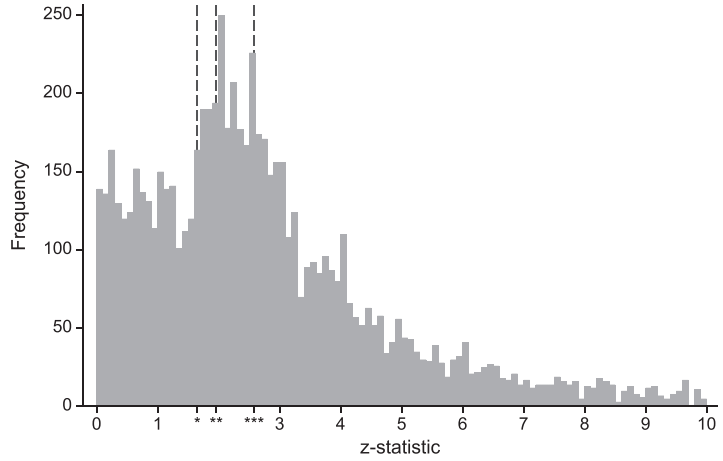

Chart B: Discontinuity in *z*-statistics from archival studies

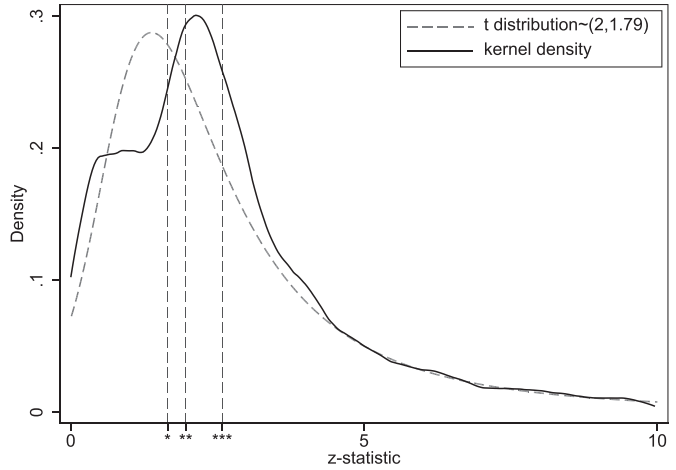

FIG. 3.—The distribution of and discontinuity in *z*-statistics from archival studies. Chart A plots the distribution of transformed *z*-statistics from archival studies for the range of *z* between 0 and 10 ($N = 8521$). Chart B presents the observed and expected distributions of archival *z*-statistics. The observed distribution is obtained by superimposing an Epanechnikov kernel density on the histogram in chart A. The expected distribution is the calibrated noncentral Student's *t*-distribution by minimizing the difference in the $z > 5$ mass between the observed and expected distributions.

observed and expected distributions. The underlying assumption is that observed *z*-statistics above five should be unaffected by researchers' QRPs. In other words, researchers' incentives to make highly significant results (i.e., *z*-statistics far above the traditional significance thresholds) even more significant should be weak. We optimize the expected distribution by cali-

brating the degrees of freedom and the noncentrality parameters of the *t*-distribution. The calibration details are in section III.C of Brodeur, Cook, and Heyes [2020].

Chart B of figure 3 shows that the optimized *t*-distribution (the dashed curve) matches well with the observed *z*-statistics (the solid curve) for the region of $z > 5$. The actual distribution of *z*-statistics has 15.4% of its mass in the tail, whereas the expected distribution has a mass of 15.1%. Thus, the calibration appears to be both visually and numerically successful. We then use the expected distribution as a benchmark to assess the extent to which the observed *z*-statistics are abnormally distributed around the significance thresholds. Specifically, we compute the mass differences between the observed and expected distributions for a given range of *z*-statistics. The mass of each distribution is separately computed using its cumulative distribution function. The results in panel A of table 2 show that the masses of z in [0, 1.65) and [1.65, 1.96) are both less than their counterparts under the expected distribution. In the [0, 1.65) region, the observed (expected) mass is 0.317 (0.365). The difference (0.048 or 4.8% of total mass) amounts to 13.2% of the expected mass in the region. In the [1.65, 1.96) region, its observed mass (0.078) is lower than expected (0.086) by 0.007, which is about 8.4% of the expected mass. These results suggest that a nontrivial fraction of $z < 1.96$ (below the 5% threshold) is missing.

Turning to regions with $z > 1.96$, we find that the differences between observed and expected masses are positive. Specifically, the [1.96, 2.58) region has an observed mass (0.169) of significant test statistics exceeding the expected mass (0.141) by 0.027 (19.4% of the expected mass). Finally, the [2.58, 5] region has an excess mass of 0.025 (0.282–0.257), amounting to 9.7% of the expected mass. Overall, the distribution of excess masses suggests that a nonnegligible proportion of insignificant test statistics in archival studies may have been shifted to significant regions, especially to [1.96, 2.58).

3.3 COMPARING DISCONTINUITIES IN TEST STATISTICS BETWEEN EXPERIMENTAL AND ARCHIVAL STUDIES

Having established that both experimental and archival studies exhibit discontinuous test statistics around all three significance thresholds, we compare the degree of test statistics discontinuity between the two types of accounting studies in several ways. First, in panel A of table 2, we summarize the proportion of excess test statistics around significance thresholds based on the $p(z)$-curve analysis conducted above. For the 1% significance threshold, we observe that the proportion of excess test statistics is 22.8% for experimental studies and 9.7% for archival studies. Experimental studies have a similar proportion of excess test statistics at the 5% level (19.1%) and a substantially higher proportion of excess test statistics at the 10% level compared with archival studies (35.9% versus −8.4%).

Second, we convert experimental *p*-values to their corresponding two-tailed *z*-statistics under the standard normal distribution (e.g., $p = 0.05$ is

converted to $z = 1.96$) and compute the fraction of z-statistics in a narrow window containing a conventional significance threshold. We choose three z-statistic windows around a significance threshold: ± 0.1, ± 0.2, ± 0.3. Within a given window (e.g., 1.96 ± 0.1), we count the total number of test statistics and compute the proportion of test statistics reported as significant ($z \geq 1.96$). This randomization test assumes that the underlying distribution of z-statistics is continuous and infinitely differentiable (Andrews and Kasy [2019], Brodeur, Cook, and Heyes [2020]). As a result, any discontinuity in observed z-statistics around a conventional threshold could arise from researchers' QRPs. Accounting researchers commonly label $p \leq 0.05$ as significant and $p \leq 0.10$ as marginally significant, making $p = 0.10$ a less important significance target than $p = 0.05$. Thus, to conserve space, we only tabulate the results related to the 1% and 5% significance levels in subsequent analyses and tabulate the 10% significance level findings in the online appendix.

Columns 3 and 7 in panel B of table 2 suggest that both experimental and archival studies have higher proportions of significant test statistics relative to insignificant test statistics within the narrow windows around conventional thresholds. For instance, for the window of half-width = 0.1 around the 5% significance threshold, 70.5% (55.1%) of experimental (archival) test statistics are reported as statistically significant. For both groups of accounting studies, the proportion of significant test statistics generally decreases with window width, probably because researchers do not see much harm in turning a marginally insignificant test statistic (e.g., $p = 0.052$) into a significant one (e.g., $p = 0.049$) given all the uncertainties of hypothesis testing in social science research. Meanwhile, such manipulation is not statistically substantial, and the $p = 0.05$ threshold is arbitrary with little relation to whether the underlying hypothesis is true or not.

We follow Brodeur, Cook, and Heyes [2020] and test whether the reported test statistics are binomially distributed around a significance threshold with equal probability. Researchers' QRPs primarily aim to inflate significance, resulting in too many rather than too few significant test statistics. Thus, we report p-values for binomial tests in columns 4 and 8 of panel B. Under the null hypothesis that test statistics are equally likely to be significant or insignificant within narrow windows, the probability of observing experimental studies' proportions of significant test statistics is close to zero (i.e., p-values of all binomial tests = 0.000 in column 4). The probability of observing archival studies' proportions of significant test statistics varies across different windows, ranging from 0.000 to 0.426. Column 8 shows that four out of six p-values from the binomial tests are < 0.050. These findings indicate that both experimental and archival studies have statistically significant discontinuities in the distribution of test statistics around significance thresholds.

Further, we compute the difference in the proportion of significant test statistics between experimental and archival studies for all test-statistic windows in column 9. We then test the equality of two proportions for each

window using an asymptotically normally distributed test statistic and report the resulting *z*-statistics in column 10. The results reveal that experimental articles report a significantly larger proportion of test statistics that beat the statistical threshold for each window. The differences are all statistically significant, with the minimum *z*-statistic = 2.218. Collectively, our findings in table 2 illustrate that experimental studies exhibit more pronounced discontinuities in the distribution of test statistics around conventional statistical thresholds than archival studies.

Third, we follow Brodeur, Cook, and Heyes [2020] and use a caliper test to compare the degree of *p*-value discontinuity between experimental and archival studies. Similar to the randomization test above, this test relies on transformed *z*-statistics and focuses on those in a narrow window around a statistical significance threshold. In addition, the test's multivariate regression framework allows us to account for author and article characteristics explicitly. Specifically, we run the following probit regression:

$$\Pr\left(Significant_{ij} = 1\right) = \Phi\left(\alpha + \beta\, Experiment_i + \gamma\, X_i + \epsilon_{ij}\right), \tag{1}$$

where $Significant_{ij}$ is an indicator of whether test statistic j from article i is statistically significant for a given threshold. For instance, $Significant_{ij}$ equals one for a *z*-statistic ≥ 1.96 and zero otherwise for the 5% threshold. The indicator variable $Experiment_i$ equals one for test statistics from experimental studies and zero for those from archival studies. The vector X_i includes the author and article characteristics defined in section 2.3 and a time trend variable (*Trend*) defined as the number of years between 1990 and a paper's publication year. We cluster the standard errors by article to ensure robustness to unspecified correlations among test statistics within an article.

Table 3 reports the marginal effects of the probit model. In columns 1–3, we examine the likelihood of beating the 5% significance threshold for transformed *z*-statistics within the narrow windows of 1.96 ± 0.1, 1.96 ± 0.2, and 1.96 ± 0.3, respectively. Across all three windows, the coefficients of *Experiment* are positive and significant at the 1% level, suggesting that experimental accounting studies are more likely to report test statistics that beat the 5% threshold than archival studies. For example, in column 1, where *z*-statistics are in the range of 1.96 ± 0.1, the coefficient for *Experiment* is 0.212 (*z*-statistic = 5.104). This result indicates that the likelihood of beating the 5% significance threshold in experimental studies exceeds that of archival studies by 21.2 percentage points. Such a difference is meaningful given that the probability of beating the 5% threshold for archival studies is 55.1 percentage points (panel B of table 2) for *z*-statistics $\in [1.86, 2.06]$. Moreover, similar to the results in panel B of table 2, we find that the coefficients of *Experiment* decrease from columns 1 to 3, indicating that the difference in the likelihood of reporting significant test statistics between experimental and archival studies decreases as the *z*-statistic range widens. Columns 4–6 present the likelihood of beating the 1% significance threshold for *z*-statistics within three ranges centered around $z = 2.58$. The coeffi-

TABLE 3

Caliper Tests Comparing the Extent of Discontinuity in Test Statistics between Experimental and Archival Studies

	(1)	(2)	(3)	(4)	(5)	(6)
Windows	[1.96 ± 0.1]	[1.96 ± 0.2]	[1.96 ± 0.3]	[2.58 ± 0.1]	[2.58 ± 0.2]	[2.58 ± 0.3]
Experiment	0.212	0.154	0.090	0.225	0.215	0.108
	(5.104)	(4.790)	(3.277)	(5.177)	(4.930)	(2.846)
Authors	−0.016	−0.000	−0.001	−0.004	0.021	0.006
	(−0.926)	(−0.015)	(−0.063)	(−0.213)	(1.171)	(0.387)
D(*Female Author*)	0.003	−0.017	−0.008	0.099	0.087	0.056
	(0.113)	(−0.696)	(−0.393)	(3.168)	(2.908)	(2.191)
Experience	−0.006	−0.013	−0.012	−0.012	−0.037	−0.028
	(−0.197)	(−0.555)	(−0.614)	(−0.353)	(−1.244)	(−1.152)
Top Institution	0.095	0.081	0.072	0.058	0.035	−0.002
	(2.234)	(2.108)	(2.225)	(1.227)	(0.778)	(−0.064)
Top PhD	−0.087	−0.085	−0.062	0.025	0.033	−0.013
	(−2.081)	(−2.422)	(−2.100)	(0.560)	(0.745)	(−0.353)
D(*Main Results*)	0.030	0.032	0.026	0.008	0.029	0.035
	(1.106)	(1.420)	(1.423)	(0.328)	(1.216)	(1.640)
Ln(*Sample Size*)	0.018	0.020	0.014	0.005	0.014	0.019
	(1.719)	(2.532)	(2.141)	(0.521)	(1.519)	(2.264)
Ln(*Test Stats*)	0.000	0.002	−0.006	0.016	0.022	0.014
	(0.014)	(0.123)	(−0.396)	(0.629)	(0.900)	(0.640)
D(*Top* 3)	−0.017	−0.030	−0.031	−0.015	−0.009	−0.004
	(−0.599)	(−1.171)	(−1.454)	(−0.435)	(−0.281)	(−0.151)
D(*High Citation*)	0.002	−0.012	−0.023	0.029	0.009	−0.001
	(0.087)	(−0.514)	(−1.190)	(0.951)	(0.297)	(−0.043)
Trend	−0.003	−0.002	−0.000	−0.004	−0.005	−0.006
	(−1.393)	(−1.016)	(−0.255)	(−1.992)	(−2.230)	(−3.440)
N	1349	2240	3397	1137	1716	2604

This table reports the marginal effects from probit regressions that estimate the likelihood of transformed statistics being statistically significant within narrow *z*-statistic windows. In columns 1–3 and 4–6, the dependent variables (*Significant*) are dummy variables that equal one if a test statistic is significant at the 5% and 1% levels, respectively. The indicator variable *Experiment* equals one for test statistics from experimental studies and zero for those from archival studies. Control variables are author and article characteristics defined in Table 1. *Sample Size* equals *Experiment Participants* for experimental studies and *Archival Obs* for archival studies. *Trend* is the number of years between 1990 and a paper's publication year. The figures in parentheses are *z*-statistics based on robust standard errors clustered at the article level.

cients of *Experiment* imply that experimental studies are 22.5, 21.5, and 10.8 percentage points more likely to report statistically significant test statistics compared with archival studies for the three ranges, respectively.

Turning to control variables, we find that articles with more authors affiliated with top business schools are more likely to report tests significant at the 5% level. These researchers may be under more publication pressure, compelling them to exercise discretion to beat the significance threshold. In addition, researchers with more rigorous PhD training (measured by graduating with doctorates from top schools) are less likely to report tests beating the 5% threshold. Articles with larger samples report a larger fraction of statistically significant test statistics around significance thresholds. Paper citation counts, researcher experience, the number of test statistics,

and *D*(*Main Results*) are not significantly associated with the likelihood of reporting statistically significant results.[19] The coefficients of *D*(*Top* 3) suggest that *p*-value discontinuity does not differ significantly between the top three and the other three accounting journals. Finally, the coefficients of *Trend* imply that in more recent years, accounting articles report test statistics that are less discontinuous at the 1% level.[20]

3.4 SAMPLE SIZE AND DISCONTINUITIES IN TEST STATISTICS

Experimental and archival accounting studies differ significantly in sample size. Table 1 reports that the mean (median) sample size is 147 (102) for experimental accounting studies and 41,837 (3,034) for archival research. The *p*-value in statistical hypothesis testing is a random variable, with its randomness arising from the variability inherent in sampling (Murdoch, Tsai, and Adcock [2008]). Sackrowitz and Samuel-Cahn [1999] demonstrate that, *ceteris paribus*, larger sample sizes increase the likelihood of detecting a tested effect if it truly exists, resulting in smaller expected *p*-values. To confirm this in our sample, we summarize archival and experimental *p*-values after converting archival z-statistics to *p*-values. Panel A of table 4 shows that compared with archival *p*-values, experimental *p*-values are, on average, larger, more likely to be closer to one than to zero (exhibiting higher skewness), and more concentrated around 0.05 (with a median of 0.054).

As experimental studies have larger average *p*-values due to smaller samples, within any narrow windows around conventional thresholds, they should have a lower proportion of significant test statistics (e.g., small *p*-values) than archival studies. However, our analyses in tables 2 and 3 reveal

---

[19] Ex ante, the relation between test-statistic discontinuity and a paper's citations is unclear. On the one hand, papers with more significant results may attract more attention and citations. On the other hand, if researchers understand that discontinuity increases the chance of false-positive findings, they may avoid citing those articles. Relatedly, Serra-Garcia and Gneezy [2021] point out that review teams tend to prefer "novel articles," which typically have a high number of citations and apply more lax robustness and reproducibility standards to their findings. Thus, nonreplicable publications are often cited more often than replicable ones. Last, Schafmeister [2021] finds that the publication of nonsupportive replication attempts does not significantly affect the citation pattern of the original studies, consistent with a lack of attention to and the limited communication of replication results among researchers.

[20] In an untabulated test, we divide the sample period into two subperiods (1990–2004 and 2005–2020) and define an indicator variable, *D*(*2005–2020*), which equals one for the 2005–2020 period and zero otherwise. We then use it to replace *Trend* in Equation (1). The results reveal that the coefficients on *D*(*2005–2020*) are negative in four of the six columns but marginally significant only in column 6 for the window of 2.58 ± 0.3 (z-statistic $= -1.869$). This additional test provides weak support for the conjecture that the test-statistic discontinuity in the latter period (2005–2020) is less pronounced than that in the earlier period (1990–2004). Moreover, we compute the frequencies of vague *p*-values in both subperiods for experimental studies. The 1990–2004 subperiod has 3538 *p*-values, 1471 of which are vague, whereas the 2005–2020 subperiod has 10,335 *p*-values, 3250 of which are vague. Thus, the fraction of vague *p*-values in the latter period ($31.4\% = 3250/10{,}335$) is lower than that in the earlier period ($41.5\% = 1471/3538$).

TABLE 4

Sample Size and Discontinuities in Test Statistics

Panel A: Summary Statistics of Archival and Experimental *P*-Values

P-values	Mean	Median	Skewness	*N*
Archival	0.154	0.020	1.820	8992
Experimental	0.196	0.054	1.474	9137

Panel B: Caliper Tests for Experimental *z*-Statistics

	(1)	(2)	(3)	(4)	(5)	(6)
Windows	[1.96 ± 0.1]	[1.96 ± 0.2]	[1.96 ± 0.3]	[2.58 ± 0.1]	[2.58 ± 0.2]	[2.58 ± 0.3]
D(*Medium Sample*)	0.022	−0.019	0.007	0.051	0.034	−0.021
	(0.466)	(−0.434)	(0.194)	(0.961)	(0.538)	(−0.365)
D(*Large Sample*)	−0.129	−0.147	−0.154	0.013	0.036	−0.024
	(−1.639)	(−2.179)	(−2.728)	(0.141)	(0.371)	(−0.261)
N	910	1445	2211	763	1022	1579

Panel C: Caliper Tests for Archival *z*-Statistics

	(1)	(2)	(3)	(4)	(5)	(6)
Windows	[1.96 ± 0.1]	[1.96 ± 0.2]	[1.96 ± 0.3]	[2.58 ± 0.1]	[2.58 ± 0.2]	[2.58 ± 0.3]
D(*Medium Sample*)	0.084	0.108	0.083	−0.220	−0.174	−0.085
	(0.897)	(1.632)	(1.481)	(−2.560)	(−2.672)	(−1.462)
D(*Large Sample*)	0.095	0.163	0.146	−0.392	−0.283	−0.157
	(0.634)	(1.522)	(1.616)	(−2.879)	(−2.848)	(−1.755)
N	439	795	1186	374	694	1025

Panel A summarizes archival and experimental *p*-values after converting archival *z*-statistics to *p*-values. Panels B and C report the marginal effects from probit regressions that estimate the likelihood of transformed statistics being statistically significant within narrow *z*-statistic windows for experimental and archival studies, respectively. In columns 1–3 and 4–6, the dependent variables (*Significant*) are dummy variables that equal one if a test statistic is significant at the 5% and 1% levels, respectively. Accounting articles are split into terciles according to sample size for experimental and archival studies separately. The indicator variables (*D*(*Medium Sample*) and *D*(*Large Sample*)) represent test statistics from the articles with medium and large samples. Other explanatory variables (excluding *Experiment*) are the same as those in table 3 and defined in table 1. Their coefficients are not tabulated for brevity. The figures in parentheses are *z*-statistics based on robust standard errors clustered at the article level.

that experimental studies report more statistically significant test statistics than archival studies for all narrow ranges around conventional thresholds. Consequently, the divergence in test statistics discontinuity between experimental and archival studies cannot be attributed to the disparity in average *p*-values induced by sample sizes.

To further examine how sample size relates to the discontinuity in accounting test statistics, we split accounting articles into terciles according to sample size separately for experimental and archival studies. We then construct three indicator variables ($D(\textit{Small Sample})$, $D(\textit{Medium Sample})$, and $D(\textit{Large Sample})$) representing test statistics from the articles with small, medium, and large samples. After excluding *Experiment* and adding $D(\textit{Medium Sample})$ and $D(\textit{Large Sample})$, we reestimate equation (1) and report the results in panels B and C of table 4 for experimental and archival studies, respectively. The coefficient of $D(\textit{Medium Sample})$ ($D(\textit{Large Sample})$) reflects the difference in the test-statistic discontinuity between medium (large) samples and small samples, whose discontinuity is captured by the intercept. Other explanatory variables in equation (1) are included in regressions, but their coefficients are not tabulated to conserve space.

Panel B presents the results for experimental studies. The coefficients of $D(\textit{Medium Sample})$ indicate no significant differences in test-statistic discontinuities around the 5% and 1% thresholds between small and medium samples. However, the coefficients of $D(\textit{Large Sample})$ are negative and significant for the narrow windows of 1.96 ± 0.2 and 1.96 ± 0.3, implying that experimental studies using large samples report test statistics with a lower degree of discontinuity around the 5% threshold than those with small samples. In the context of archival studies, panel C shows that both medium and large samples are associated with lower likelihoods of beating the 1% significance threshold for *z*-statistics centered around $z = 2.58$ compared with small samples. Collectively, our results in table 4 suggest that accounting studies with smaller samples display more discontinuous test statistics around significance thresholds.

4. Researchers' Discretion and Discontinuities in the Distribution of Test Statistics

To examine whether researcher degrees of freedom relate to the discontinuity in test statistics reported in accounting studies and explain the differences in *p*-value discontinuities between experimental and archival studies, we employ several characteristics of experimental design that may correlate with researcher degrees of freedom.

First, we record the experimental design of each article and count the number of variables (i.e., constructs) manipulated. A higher number of constructs naturally leads to more interaction terms in analyses. For example, a 2×2 ANOVA has only one interaction term, whereas a $2 \times 2 \times 2$ ANOVA has three two-way interactions and one three-way interaction. Researchers should have greater discretion to select significant

interaction terms or exclude insignificant terms if an experiment has many constructs.[21] Thus, we define the indicator variable *D*(*Fewer Constructs*) for experimental articles with the number of constructs equal to or lower than the sample median (two constructs) and *D*(*More Constructs*) for other experimental articles.

In panel A of table 5, we reestimate equation (1) with *D*(*Fewer Constructs*) and *D*(*More Constructs*). The coefficients suggest that experimental papers with a higher number of constructs have a more discontinuous distribution of test statistics relative to archival studies. Within the range 2.58 ± 0.1, the marginal effect of *D*(*More Constructs*) is 0.245, which is about 1.6 times as large as the marginal effect of *D*(*Fewer Constructs*) (0.151) on the likelihood of reporting significant results at the 1% level. The *F*-test on the equality of the two coefficients is significant at the 10% level ($p = 0.086$). However, the differences between the coefficients of *D*(*Fewer Constructs*) and *D*(*More Constructs*) for *z*-statistic ranges around the 5% threshold are statistically insignificant.

Second, we examine researchers' discretion regarding the choice between one-tailed and two-tailed tests. A one-tailed test is appropriate if a researcher predicts a difference between groups or conditions in a specific direction (e.g., group 1 scores higher than group 2) and has no interest in the possibility of the opposite direction being true (e.g., group 1 scores lower than group 2). The main advantage of a one-tailed test is its higher statistical power compared with its two-tailed counterpart at the same significance level because it focuses the critical regions of the distribution on one side (i.e., the predicted direction). However, some researchers may misuse this advantage by strategically framing their hypotheses as directional and conducting one-tailed tests to achieve lower *p*-values (Khan and Tronnes [2019]).

We define two indicators, *D*(*One-Tailed Experiment*) and *D*(*Two-Tailed Experiment*), to flag experimental articles with and without one-tailed tests, respectively. Panel B of table 5 reports the results obtained by reestimating equation (1) with the two indicators. We find that the coefficients of *D*(*One-Tailed Experiment*) are larger than those of *D*(*Two-Tailed Experiment*) for all three *z*-statistic ranges around the 5% threshold. The coefficient differences are statistically significant at the 1% level according to the *F*-

[21] The experimental papers in our sample include an average of about 147 participants. Combining this information with the number of constructs (or manipulations) allows us to compute the number of participants per cell and assess the power of a typical test in experimental studies. We tabulate in the online appendix that the most commonly used experimental design is 2×2 (40.0% of the experimental papers in our sample), followed by 2×3 (12.1%) and $2 \times 2 \times 2$ (9.0%). Thus, a typical paper has four conditions and about 37 subjects ($= 147 \div 4$) per condition. This implies that experimental tests in our sample are likely to be inadequately powered for small or medium effect sizes, which are the norm for behavioral experiments (Peterson, Albaum, and Beltramini [1985], McShane et al. [2019]). In particular, according to Cohen [2013], a two-sided *t*-test with 37 participants per cell has 56% (7%) power to detect a medium (small) effect size at the 5% significance level.

TABLE 5

Researcher Degrees of Freedom and the Extent of Discontinuity in Test Statistics

	(1)	(2)	(3)	(4)	(5)	(6)
Windows	[1.96 ± 0.1]	[1.96 ± 0.2]	[1.96 ± 0.3]	[2.58 ± 0.1]	[2.58 ± 0.2]	[2.58 ± 0.3]
Panel A: Fewer vs More Constructs						
D(*Fewer constructs*): *a*	0.126	0.100	0.071	0.151	0.152	0.066
	(3.163)	(2.996)	(2.521)	(3.505)	(3.458)	(1.873)
D(*More constructs*): *b*	0.197	0.127	0.055	0.245	0.240	0.165
	(3.754)	(2.843)	(1.384)	(4.365)	(4.065)	(3.337)
N	1349	2240	3397	1137	1716	2604
P-values of testing $a = b$	0.136	0.503	0.641	0.086	0.127	0.031
Panel B: One-tailed vs. Two-tailed Tests						
D(*One-tailed Experiment*): *a*	0.241	0.196	0.116	0.215	0.200	0.082
	(5.650)	(6.062)	(4.112)	(4.669)	(4.267)	(2.095)
D(*Two-tailed Experiment*): *b*	0.124	0.050	0.024	0.253	0.245	0.165
	(2.411)	(1.135)	(0.630)	(4.412)	(4.152)	(3.013)
N	1349	2238	3394	1136	1714	2602
P-values of testing $a = b$	0.005	0.000	0.005	0.461	0.435	0.109

(*Continued*)

TABLE 5—*(Continued)*

Panel C: High vs. Low Experiment Costs						
D(*High Subject Cost*): *a*	0.226	0.181	0.114	0.222	0.210	0.138
	(5.066)	(5.034)	(3.777)	(4.640)	(4.364)	(3.302)
D(*Low Subject Cost*): *b*	0.177	0.114	0.067	0.237	0.216	0.060
	(3.745)	(3.185)	(2.128)	(4.471)	(4.048)	(1.390)
N	1298	2168	3285	1101	1671	2528
P-values of testing $a = b$	0.186	0.041	0.089	0.755	0.904	0.043
Panel D: Rounded vs. Unrounded Experimental *p*-Values						
D(*Rounded P*): *a*	0.299	0.181	0.102	(.)	(.)	0.177
	(7.099)	(4.896)	(3.272)	(.)	(.)	(4.124)
D(*Unrounded P*): *b*	0.111	0.133	0.079	−0.017	0.002	0.044
	(2.502)	(3.868)	(2.687)	(−0.270)	(0.044)	(1.185)
N	1349	2240	3397	636	1215	2604
P-values of testing $a = b$	0.000	0.118	0.353	0.787	0.965	0.000

This table shows the association between researcher degrees of freedom and differences in *p*-value discontinuity between experimental and archival studies. Coefficients are marginal effects from probit regressions on the likelihood of transformed statistics being statistically significant within narrow *z*-statistic windows. In columns 1–3 and 4–6, the dependent variables (Significant) are dummy variables that equal one if a test statistic is significant at the 5% and 1% levels, respectively. In panel A, *D*(*Fewer Constructs*) and *D*(*More Constructs*) indicate experimental articles with the number of constructs equal to or smaller than the sample median and other experimental articles, respectively. In panel B, *D*(*One-Tailed Experiment*) and *D*(*Two-Tailed Experiment*) indicate experimental articles with and without one-tailed tests, respectively. In panel C, *D*(*High Subject Cost*) and *D*(*Low Subject Costs*) indicate experimental articles with high-cost (practitioners and Master of Business Administration students) and low-cost (undergraduate students, online participants, and others) participants, respectively. In panel D, *D*(*Rounded P*) and *D*(*Unrounded P*) indicate *p*-values reported to two or more decimal places, respectively, in experimental articles. Control variables are included but not reported in all panels. Figures in parentheses are *z*-statistics based on robust standard errors clustered at the article level.

tests in columns 1–3. These findings imply that experimental articles with one-tailed tests are more likely to report results significant at the 5% level than those with two-tailed tests. In other columns, the coefficients of *D*(*One-Tailed Experiment*) are generally smaller than those of *D*(*Two-Tailed Experiment*), but the differences are not statistically significant. This indicates that researchers' discretion in reporting one- or two-tailed tests is related to *p*-value discontinuity mainly for test statistics around the 5% threshold.

Third, we examine researchers' discretion related to the type of experiment participants. To conduct experiments, researchers need access to a suitable group of participants that can be placed under similar experimental conditions. There are typically four types of participants in experimental accounting studies: (1) participants from online experiment platforms (e.g., MTurk (Amazon Mechanical Turk) and Qualtrics), (2) undergraduate students, (3) Master of Business Administration (MBA) students, and (4) professional subjects such as partners in auditing firms. The first two types of participants are relatively easy to access, given their abundant supply, whereas the latter two are relatively scarce and costly to access (Abdel-Khalik [1974], Libby, Bloomfield, and Nelson [2002]). Experiments conducted with high-cost participants are more difficult for other researchers to replicate. This lack of external replication may increase researchers' incentives to exercise discretion in their experiments (Lacetera and Zirulia [2011], Craig et al. [2020]). Conversely, low experiment costs may facilitate researchers' QRPs. For example, some researchers run numerous studies on MTurk due to the low marginal cost of data collection (Calin-Jageman [2018], Brodeur, Cook, and Heyes [2022]) and may selectively report studies with statistically significant findings, especially when replications are scarce.[22]

We use the indicator *D*(*High Subject Cost*) to flag experimental articles with MBA students or professionals as participants and the indicator *D*(*Low Subject Cost*) to flag experimental articles with online participants or undergraduates as participants. Panel C of table 5 reports the results of *D*(*High Subject Cost*) and *D*(*Low Subject Cost*). We find that the differences in *p*-value discontinuity between experimental and archival studies are larger for experimental papers with high-cost subjects in four out of six regressions. Among the four regressions, the coefficient differences are statistically significant at the 5% or 1% levels for three regressions, supporting the argument that the degree of *p*-value discontinuity is positively related to the cost of conducting experiments.

[22] Brodeur, Cook, and Heyes [2022] analyze all MTurk experiments published in leading journals between 2010 and 2020 and find that the average remuneration per participant is only US$1.30. Moreover, most MTurk experiments use small samples: the median number of subjects is 249. The test-statistic distributions of MTurk articles exhibit discontinuities around conventional significance thresholds, consistent with considerable QRP and publication bias. When partitioning their sample based on the cost of experiments, Brodeur, Cook, and Heyes [2022] find that low-cost experiments report more discontinuous test statistics around significance thresholds than high-cost experiments.

Fourth, we consider researchers' flexibility in choosing how many decimal places to use when reporting *p*-values. In particular, researchers have the discretion to round *p*-values to two decimal places instead of reporting them with three or more. Under null hypothesis significance testing (NHST), reporting rounded *p*-values can raise the chance of meeting statistical significance thresholds because *p*-values just above the thresholds could be rounded down (e.g., $p = 0.054$ rounded down to 0.05).[23] Although such a discretionary action does not drastically inflate statistical significance or exacerbate irreproducibility, it can influence false-positive rates and how the academic community perceives the results. Gelman and Stern [2006] demonstrate that empirical researchers often fail to appreciate that the difference between "significant" ($p = 0.05$) and "not significant" ($p = 0.054$) is not itself statistically significant. McShane et al. [2019] point out that the NHST paradigm encourages researchers to interpret *p*-values dichotomously rather than continuously. Such dichotomous thinking is reinforced by the practice of marking test statistics with eye-catchers, such as stars (asterisks). For example, researchers often use double stars (**) for $p = 0.05$ to indicate 5% significance and a single star (*) for $p = 0.054$ to represent 10% significance (Brodeur et al. [2016]).

To examine whether the rounding choice is related to the discontinuity in test statistics, we define an indicator *D*(*Rounded P*) to flag *p*-values reported with two decimal places and an indicator *D*(*Unrounded P*) to flag *p*-values reported with three or more decimal places in experimental articles. We then reestimate the caliper model in equation (1) by replacing *Experiment* with the two indicators. The results in panel D of table 5 show that the likelihood of reporting significant tests is lower for experimental articles reporting unrounded *p*-values around the 5% threshold. For example, column 1 shows that the marginal effect of *D*(*Rounded P*) is 0.299, about 2.7 times as large as the marginal effect of *D*(*Unrounded P*) (0.111) on the likelihood of reporting significant test statistics in the range of 1.96 ± 0.1. We conduct *F*-tests on the equality of these two coefficients and obtain a *p*-value of 0.000 ($\chi^2 = 37.22$), suggesting that the difference between the two coefficients is statistically significant. Note that in columns 4 and 5, *D*(*Rounded P*) is automatically excluded from the regressions because of a lack of variation in *D*(*Rounded P*) (all rounded *p*-values are 0.01 for $z \in 2.58 \pm 0.01$).[24]

---

[23] Although archival researchers typically report *t* or *z*-statistics rather than *p*-values, they may round up some of these statistics (e.g., rounding 1.955 up to 1.96) to achieve statistical significance. In our archival statistics sample, 84.7% of the *t* and *z*-statistics are reported with three or more decimal places.

[24] In addition, we use a bootstrapping approach to conduct a derounding analysis of the effect of rounded test statistics on discontinuity. To undo rounding, for test statistics reported to two decimal places, we randomly assign values from a range of potential unrounded values. For example, $p = 0.05$ can receive any value between 0.045 and 0.055. We repeat the derounding procedure 500 times and reestimate equation (1) using the bootstrapped samples. We find

In sum, using several proxies for researcher discretion in data analysis and result reporting, we show that the discontinuity gap between experimental and archival studies widens when researchers have the opportunity to take discretionary actions in experimental studies, suggesting that the unusual abundance of just-significant test statistics around significance thresholds is related to researcher degrees of freedom. Most discretion proxies have weak explanatory power for the discontinuity gaps around the 1% threshold, possibly because some researchers' primary target is the 5% significance threshold. Last, we emphasize and reiterate that none of these findings implies that experimentalists engage more in QRPs than archival researchers for reasons outlined in section 1.

5. *Additional Analyses and Further Discussions*

5.1 VAGUE *P*-VALUES

Our baseline analyses exclude 4721 vague *p*-values in experimental articles. For example, *p*-values reported as $p \leq 0.05$ can be any value between 0.00 and 0.05. This subsection examines whether our baseline results are robust to the inclusion of these vague *p*-values, which are mostly relevant for experimental studies. Chart A of figure 4 shows that vague *p*-values are concentrated at conventional significance thresholds ($p \leq 0.001$, $p \leq 0.01$, $p \leq 0.05$, and $p \leq 0.1$). These four cases account for 77% of all vague *p*-values. Given that the first two cases denote high significance, we focus mainly on $p \leq 0.05$ because these *p*-values could directly affect the degree of *p*-value discontinuity documented in previous sections.

We read experimental articles in our sample and calculate the exact *p*-values for those reported as $p \leq 0.05$ using the test statistics and degrees of freedom information provided in the articles.[25] Despite such information being unavailable in most cases, we uncover the exact *p*-values for 80 cases reported as $p \leq 0.05$. Chart B of figure 4 shows that five of the 80 exact *p*-values reported as $p \leq 0.05$ have an exact *p*-value of ≥ 0.06 when rounded to two decimal places. These findings are consistent with those of John et al. [2012], who show that some psychology researchers use vague *p*-values to inflate significance. More importantly, chart B of figure 4 indicates that researchers are more likely to report $p \leq 0.05$ when the *p*-value is closer to 0.05

---

that after derounding, the coefficients of *Experiment* become smaller, especially around the 5% and 1% thresholds. This result confirms that a considerable part of the difference in the likelihood of reporting significant results between experimental and archival studies arises from experimental articles that round off *p*-values. We include this analysis in the online appendix.

[25] *P*-values can be calculated using the cumulative distribution function for test statistics. For example, given a two-tailed *t*-test, *p*-value $= 2 \times cdf_{t,d}(-|t_{score}|)$, where $cdf_{t,d}$ denotes the cumulative distribution function of a *t*-distribution with *d* degrees of freedom. Thus, if a paper discloses *d* and t_{score} for a vague *p*-value, the formula can be used to uncover the exact *p*-value. The *p*-values based on *F*-statistics can be calculated in a similar fashion. We find that papers published in earlier years of our sample are more likely to report vague *p*-values.

Chart A: The distribution of vague *p*-values

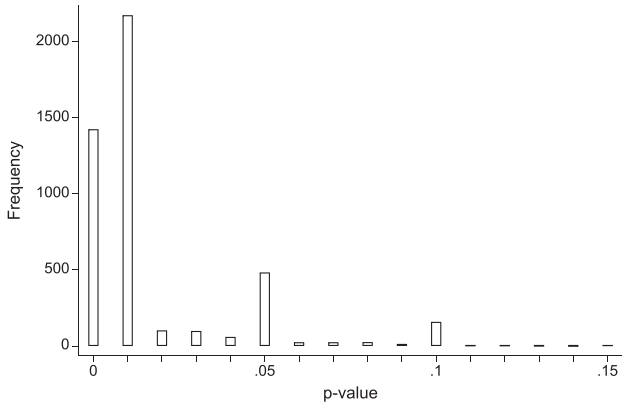

Chart B: The distribution of 80 exact *p*-values underlying vague *p*-values reported as $p \leq 0.05$

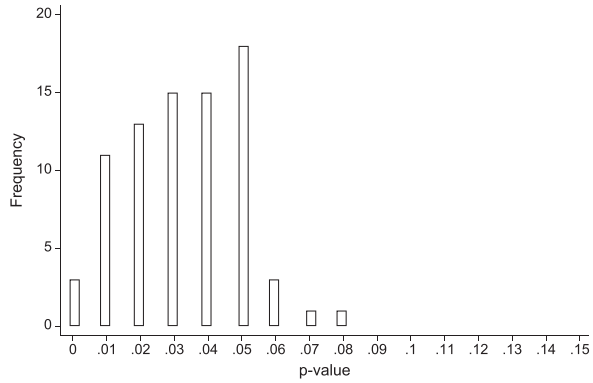

FIG. 4.—The distribution of vague *p*-values. Chart A shows the distribution of 4721 *p*-values from experimental articles reported by means of inequalities (i.e., vague *p*-values). Chart B shows the distribution of 80 exact *p*-values that we are able to compute among all vague *p*-values reported as $p \leq 0.05$ using the test statistics and the degree of freedom information provided in the articles. The *x*-axis represents *p*-values, and the *y*-axis represents frequencies.

(e.g., 0.048) than to 0.01 (e.g., 0.014). We obtain similar results for vague *p*-values reported as $p \leq 0.10$ (see the online appendix). These findings suggest that the remaining vague *p*-values, for which we have no information to calculate exact values, should have their exact *p*-values concentrated near conventional thresholds. We then impose the distribution in chart B of figure 4 on the remaining vague *p*-values and stack the estimated distribution of all vague *p*-values at 0.05 and 0.10 with that of the exact *p*-values used in our baseline analysis. Our results in the online appendix show that *p*-value discontinuities around the 5% and 10% significance thresholds in experimental studies become more prominent.

5.2 COMPARING ACCOUNTING WITH OTHER FIELDS REGARDING DISCONTINUITIES IN TEST STATISTICS

Discontinuities in test statistics are known to exist in many scientific disciplines, such as clinical medicine, materials science, psychology, and economics (Fanelli [2010]). Using a *p*-curve analysis similar to ours in section 3.1, Masicampo and Lalande [2012] estimate that about 30% of *p*-values (based on our calculation using their figure 1) immediately below the $p = 0.05$ threshold are excessive in experimental psychology. In contrast, our analysis reveals that the observed *p*-value frequency at $p = 0.05$ is about 19% higher than the expected frequencies (panel A of table 2) for experimental accounting studies.

Moreover, we employ the randomization test (panel B of table 2) to compare discontinuities in test statistics across accounting and other disciplines. In particular, we focus on test statistics reported in the window of 1.96 ± 0.2 (i.e., $1.76 < z < 2.16$), which Brodeur, Cook, and Heyes [2020] also use for comparisons. The ratio of tests just above and below 1.96 is 1.37 (= [880 + 415] / [1445 + 795 – 880 – 415]), which is higher than the corresponding ratio (1.10) for economics but lower than those for political science (2.62) and sociology (2.21) (Brodeur, Cook, and Heyes [2020]). In other words, *p*-value discontinuity is more pronounced in accounting than in economics studies but less severe than in political science and sociology studies. The ratio of tests just above and below 1.96 is 1.09 (= 415 / [795 – 415]) for archival accounting studies, similar to economics, whereas the ratio takes the value of 1.56 (= 880 / [1445 – 880]) for experimental accounting articles. However, these simple comparisons should be interpreted with caution because they are primarily based on our randomization test results for accounting studies and those reported by Brodeur, Cook, and Heyes [2020].[26]

Next, we reconcile our findings with those of Brodeur, Cook, and Heyes [2020], who assess the trustworthiness of four causal identification methods (RCT, DID, IV, and RDD) using test statistics from top economics journals. They document that RCT, a prominent experimental approach, is less subject to researchers' QRPs than other identification methods, consistent with the common view that RCT is the gold standard for studying causal relations. In contrast, we study two main research methodologies in accounting, which differ according to whether researchers collect data with (experimental) or without interactions with research subjects (archival). Most of our archival test statistics are from tests such as ordinary least squares and probit rather than DID, IV, and RDD.

Apart from differences in research focus, there is a considerable discrepancy between the RCTs of economics and accounting studies regarding pre-

[26] Brodeur, Cook, and Heyes [2020] report the test results in table A11 of their online appendix. Test results for political science are from Gerber and Malhotra [2008a], and those for sociology are from Gerber and Malhotra [2008b].

registration plans, which requires researchers to outline essential aspects of their research study without advance knowledge of the research outcomes (Nosek et al. [2018]). Usually, researchers need to submit a study protocol to an independent registry or platform, which preserves the document and makes it discoverable. With preregistration, researchers cannot easily change their research course after observing the data because deviations from the registered plans are knowable.[27] Thus, preregistration limits researcher degrees of freedom in data collection and analysis, constraining their discretionary actions to meet or beat significance thresholds. Brodeur, Cook, and Heyes [2020] point out that preregistration is expected for RCTs in economics. However, this practice is rare in accounting.

We search for keywords related to preregistration plans in our 654 experimental accounting articles.[28] Other than seven articles published under the *JAR*'s Registration-based Editorial Process (REP) in 2017, no articles mention preregistration. Bloomfield, Rennekamp, and Steenhoven [2018] find that these seven articles are more likely to report statistically insignificant results than other papers published in the *JAR* and comparable journals, implying that preregistration with guaranteed publication can limit researchers' significance-inflating QRPs and publication bias.[29] More importantly, by comparing the discontinuities in test statistics between experimental accounting and economics articles, we infer that preregistration of RCTs in economics may partially explain the difference between our findings and those of Brodeur, Cook, and Heyes [2020].

5.3 BRIEF DISCUSSIONS ON REMEDIES FOR QUESTIONABLE RESEARCHER PRACTICES

While preregistration is a valuable tool for promoting transparency and research integrity, it is also subject to a range of criticisms, including vague

---

[27] The scope of preregistration differs across research fields. For instance, preregistration in psychology typically means the comprehensive preregistration of hypotheses, data inclusion/exclusion, and detailed pre-analysis plans, among others. In addition, information on how pre-analysis plans are followed or deviated is often provided. However, in economics, preregistration is typically less comprehensive and does not require pre-anlaysis plans (Ofosu and Posner [2019]).

[28] The keywords are pre-analysis plan, RCT registry, preanalysis plan, analysis plan, preregistration, pre-registration, pre-registered, and preregistered.

[29] In the Registration-based Editorial Process (REP), authors submit data collection and analysis proposals, with the assurance of publication for successful proposals as long as they fulfil their commitments, regardless of whether the results align with their predictions. Similar to the REP, Registered Reports (RRs) are a publication form, in which authors preregister a study plan with hypotheses, methods, and analysis details before collecting data. This plan undergoes peer review, and if approved, the journal commits to publishing the final article, whether or not the hypotheses are supported. After analyzing data, authors create the final report, which is also peer-reviewed, mainly to check if they have followed the registered plan and justified their conclusions. Scheel, Schijen, and Lakens [2021] find that published RRs have significantly fewer positive results than hypothesis-testing studies in the standard psychology literature, consistent with RRs alleviating publication bias and type I error inflation.

registered research plans, increasing researchers' upfront investments, discouraging investments in unplanned but potentially important analyses, generating less thorough and refined articles, and causing nontransparent deviations from the preregistered plan (Bloomfield, Rennekamp, and Steenhoven [2018], Ofosu and Posner [2020]). Thus, Bloomfield, Rennekamp, and Steenhoven [2018] conclude that "*no system is perfect*" and propose that the editorial process should address the above criticisms by, for example, encouraging researchers' follow-up investments based on their preplanned analyses. Nosek et al. [2018] discuss the potential of preregistration to curb postdiction and address the challenges of implementing preregistration.

Pre-analysis plans (PAPs) constitute an essential practice closely related to, yet distinct from, preregistration. Compared with preregistration, which commonly serves as a time-stamped record of a study's basic characteristics (Banerjee et al. [2020]), PAPs require researchers to register more explicitly which hypotheses will be tested and how data will be analyzed (Brodeur et al. [2023]).[30] Olken [2015] summarizes the common features of PAPs, including primary/secondary outcome variables, variable definitions, inclusion/exclusion rules, statistical model specifications, subgroup analysis, and other issues (e.g., data monitoring plans and stopping rules). Hence, a PAP can substantially limit researchers' flexibility in data analysis and ability to cherry-pick hypotheses. Analyzing RCT test statistics published in 15 leading journals, Brodeur et al. [2023] find no significant difference in the test-statistic distribution between preregistered and non-preregistered studies. However, preregistered studies with fully fledged PAPs exhibit less discontinuous test statistics around significance thresholds than those without PAPs. These findings highlight the effectiveness of PAPs, rather than preregistration *per se*, in enhancing research credibility. In another vein, PAPs have been criticized for discouraging exploratory work and falling short in mitigating publication bias (Ofosu and Posner [2019]). Coffman and Niederle [2015] argue that the benefit of PAPs is reduced when replications of published studies are feasible.

Another remedy proposed is to lower the p-value threshold for statistical significance to, for example, 0.005 (Benjamin et al. [2018]). Altmejd et al. [2019] and Camerer et al. [2018] show that the replicability of social science experiments is negatively correlated with the p-value of the original study, implying that lower p-value thresholds can help enhance research credibility, promote higher-quality work, and reduce false positives

[30] For example, as of December 2023, the American Economic Association (AEA) Registry for RCTs has 17 required fields, including the trial title and status, abstract, trial start and end dates, intervention start and end dates, primary outcomes, experimental design, randomization method, expected sample size, and Institutional Review Boards approval. A PAP is optional rather than required for preregistration. Banerjee et al. [2020] report that as of February 2019, about one-third of trials on the AEA RCT Registry had uploaded a PAP in addition to the required field completion.

(Benjamin et al. [2018]). However, McShane et al. [2019] argue that lower *p*-value thresholds may lead to adverse outcomes, including overconfidence in published results with very low *p*-values, the devaluation of important discoveries that fail to meet lower thresholds, and increases in false negatives. Hence, lowering the *p*-value threshold entails a tradeoff between positive and negative outcomes. Relatedly, Pütz and Bruns [2021] suggest reporting exact *p*-values rather than adding asterisks to coefficients as eye-catchers. Because there is no sharp line between significant and insignificant differences, researchers should treat *p*-values as a continuous variable to avoid dichotomizing results. Several journals (e.g., the *American Economic Review and Econometrica*) have required authors to report either standard errors or exact *p*-values without asterisks. Moreover, Brodeur et al. [2016] document that the distribution of test statistics in three top economics journals shows less noticeable discontinuity around *p*-values of 0.05 and 0.10 in articles without eye-catchers, which are defined as the presence of stars or bold printing in tables to highlight statistical significance. This finding implies that not using eye-catchers may make researchers less influenced by dichotomous thinking with regard to statistical significance.

Research disclosure and data-sharing policies have received increasing attention across various disciplines as a promising way to facilitate replication and curb researchers' QRPs (Christensen and Miguel [2018], Nosek et al. [2018]). In December 2014, the *JAR* implemented a data-sharing policy requiring authors to disclose detailed computer programs and data used in their published articles. The policy aims to facilitate the replication of published findings and hold researchers more accountable for their analyses and results. To examine whether the data-sharing policy affects the distribution of accounting test statistics, we conduct a DID analysis based on the caliper model and find no evidence that the policy weakens the extent of discontinuity in test statistics from articles published in the *JAR* relative to those in the other five journals (see the online appendix). One possible explanation for our finding is that our DID test is inadequately powered because of the small sample in the post-policy period, which includes only 26 *JAR* articles. Another possible explanation is that while researchers may become more truthful about the results they report in published papers, they may not disclose all tests conducted or all results arising from alternative variable definitions and methodologies. Hence, hidden discretionary actions can still contribute to the discontinuous distribution of test statistics around significance thresholds in accounting studies. Consistent with this explanation, Brodeur, Cook, and Neisser [2024] document that policies on data and code sharing have no significant impact on test-statistic discontinuities around significance thresholds in top economics journals.

6. *Conclusions*

Based on a large sample of articles published in six top accounting journals from 1990 to 2020, we document considerable discontinuities in

the distribution of reported test statistics around conventional significance thresholds for both archival and experimental accounting studies. That is, there is an excessive proportion of just-significant test statistics in the accounting literature. Moreover, accounting studies with smaller samples display more discontinuous test statistics around significance thresholds. Further analysis reveals a significant difference between methods: Experimental test statistics are more discontinuous around conventional significance thresholds than archival ones. This difference is associated with researcher degrees of freedom in experimental studies.

We emphasize that our findings do not serve as evidence of opportunistic or deliberate fraudulent behavior committed by accounting researchers or imply that accounting research is collectively unreliable. On the contrary, we find that the vast majority of test statistics reported in accounting journals conform to expected distributions without researchers exercising discretion to meet or beat significance thresholds. The nontrivial fraction of excess test statistics near significance thresholds suggests that it is essential to approach reported just-significant test statistics with a healthy dose of skepticism. Yet, one should seek more substantial evidence before levying allegations of QRPs against accounting researchers.

Potential QRPs underlying the discontinuity of test statistics can give rise to false-positive results. Thus, it is important to remain cautious about drawing scientific conclusions or making business decisions based only on whether a test statistic such as a *p*-value passes a specific threshold. Given the increasing attention to research integrity in accounting research, our study contributes to open, evidence-based discussions of this issue.

APPENDIX

TABLE A1

The Distribution of the Empirical Accounting Articles in Our Sample

Publication Year	*Total*	*Archival*	*Experimental*	AOS	CAR	JAE	JAR	RAST	TAR
1990	39	19	20	2	9	7	8	0	13
1991	35	18	17	4	11	3	4	0	13
1992	30	14	16	3	4	4	3	0	16
1993	24	13	11	0	4	2	3	0	15
1994	30	16	14	4	7	7	7	0	5
1995	30	12	18	3	5	3	8	0	11
1996	32	16	16	7	3	8	8	1	5
1997	38	16	22	6	4	7	10	0	11
1998	18	10	8	3	2	2	6	2	3
1999	41	19	22	9	6	7	11	1	7
2000	27	13	14	6	3	5	5	3	5
2001	36	15	21	4	8	0	10	2	12
2002	25	11	14	2	3	4	6	3	7
2003	18	7	11	2	6	1	2	0	7

(*Continued*)

TABLE A1—*(Continued)*

Publication Year	*Total*	*Archival*	*Experimental*	AOS	CAR	JAE	JAR	RAST	TAR
2004	30	14	16	1	9	5	2	1	12
2005	44	22	22	3	14	2	3	2	20
2006	36	17	19	5	8	7	2	3	11
2007	29	13	16	1	7	4	6	5	6
2008	43	19	24	4	8	6	6	5	14
2009	30	14	16	2	7	1	3	2	15
2010	60	28	32	10	12	6	7	5	20
2011	52	23	29	8	9	6	8	5	16
2012	45	21	24	4	12	4	8	1	16
2013	31	14	17	3	4	3	7	3	11
2014	38	20	18	4	10	2	7	3	12
2015	65	33	32	14	15	3	7	8	18
2016	62	32	30	13	15	6	4	6	18
2017	46	23	23	7	16	1	6	5	11
2018	84	42	42	14	12	11	10	10	27
2019	62	32	30	9	16	6	5	5	21
2020	76	37	39	10	25	8	7	7	19
Total	1256	603	653	167	274	141	189	88	397

The table reports the distribution of the empirical accounting studies in our sample by publication year and by articles. Our sample includes 1256 articles published from 1990 to 2020 in the top six accounting journals: *Accounting, Organizations and Society* (AOS), *Contemporary Accounting Research* (CAR), *Journal of Accounting and Economics* (JAE), *Journal of Accounting Research* (JAR), *Review of Accounting Studies* (RAST), and *The Accounting Review* (TAR). The sample does not include articles that are primarily comments on or discussions of other papers. In addition, *TAR* published book reviews in the earlier years of our sample period, which are also excluded from our analysis. *Total*, *Archival*, and *Experimental* indicate the yearly number of articles, archival articles, and experimental articles, respectively. The number of experimental articles for the six journals is 233 (*TAR*), 165 (*CAR*), 142 (*AOS*), 94 (*JAR*), 12 (*RAST*), and 7 (*JAE*), respectively.

References

Abdel-Khalik, A. R. "On the Efficiency of Subject Surrogation in Accounting Research." *The Accounting Review* 49 (1974): 743–50.

Adda, J.; C. Decker; and M. Ottaviani. "P-hacking in Clinical Trials and How Incentives Shape the Distribution of Results Across Phases." *Proceedings of the National Academy of Sciences* 117 (2020): 13386–92.

Altmejd, A.; A. Dreber; E. Forsell; J. Huber; T. Imai; M. Johannesson; M. Kirchler; G. Nave; and C. Camerer. "Predicting the Replicability of Social Science Lab Experiments." *PLoS ONE* 14, 12 (2019): e0225826.

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

Banerjee, A.; E. Duflo; A. Finkelstein; L. F. Katz; B. A. Olken; and A. Sautmann. "In Praise of Moderation: Suggestions for the Scope and Use of Pre-analysis Plans for RCTs in Economics." Working Paper, 2020. https://www.nber.org/papers/w26993.

Basu, S., and H.-U. Park. "Publication Bias in Recent Empirical Accounting Research." Working Paper, 2014. https://ssrn.com/abstract=2379889.

Benjamin, D. J.; J. O. Berger; M. Johannesson; B. A. Nosek; E.-J. Wagenmakers; R. Berk; K. A. Bollen; B. Brembs; L. Brown; and C. Camerer. "Redefine Statistical Significance." *Nature Human Behaviour* 2 (2018): 6–10.

Bishop, D. V., and P. A. Thompson. "Problems in Using *p*-curve Analysis and Text-mining to Detect Rate of *p*-hacking and Evidential Value." *PeerJ* 4 (2016): e1715.

Bloomfield, R.; K. Rennekamp; and B. Steenhoven "No System Is Perfect: Understanding How Registration-Based Editorial Processes Affect Reproducibility and Investment in Research Quality." *Journal of Accounting Research* 56 (2018): 313–62.

Brodeur, A.; N. Cook; J. Hartley; and A. Heyes. "Do Pre-Registration and Pre-Analysis Plans Reduce p-Hacking and Publication Bias?: Evidence from 15,992 Test Statistics and Suggestions for Improvement." Working Paper, 2023. https://ssrn.com/abstract=4180594

Brodeur, A.; N. Cook; and A. Heyes. "Methods Matter: *P*-hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review* 110 (2020): 3634–60.

Brodeur, A.; N. Cook; and A. Heyes. "We Need to Talk about Mechanical Turk: What 22,989 Hypothesis Tests Tell Us about Publication Bias and p-Hacking in Online Experiments." Working Paper, 2022. https://ssrn.com/abstract=4188289

Brodeur, A.; N. Cook; and C. Neisser. "P-Hacking, Data Type and Data-Sharing Policy." *The Economic Journal* 134 (2024): 985–1018.

Brodeur, A.; L. Mathias; S. Marc; and Z. Yanos. "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics* 8 (2016): 1–32.

Calin-Jageman, R. "The Perils of MTurk, Part 1: Fuel to the Publication Bias Fire?" (2018). https://thenewstatistics.com/itns/2018/05/02/the-perils-of-mturk-part-1-fuel-to-the-publication-bias-fire/.

Camerer, C.F.; A. Dreber; F. Holzmeister; T.H. Ho; J. Huber; M. Johannesson; M. Kirchler; G. Nave; B.A. Nosek; T. Pfeiffer; and A. Altmejd. Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour* 2 (2018): 637–44.

Chen, A. Y. "The Limits of p-Hacking: Some Thought Experiments." *The Journal of Finance* 76 (2021): 2447–80.

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

Coffman, L. C., and M. Niederle. "Pre-analysis Plans Have Limited Upside, Especially Where Replications Are Feasible." *Journal of Economic Perspectives* 29 (2015): 81–98.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences.* Routledge, 2013.

Craig, R.; A. Cox; D. Tourish; and A. Thorpe. "Using Retracted Journal Articles in Psychology to Understand Research Misconduct in the Social Sciences: What Is to Be Done?" *Research Policy* 49 (2020): 103930.

Ellison, G. "The Slowdown of the Economics Publishing Process." *Journal of Political Economy* 110 (2002): 947–93.

Fanelli, D. "'Positive' Results Increase Down the Hierarchy of the Sciences." *PLoS ONE* 5 (2010): e10068.

Fanelli, D. "Negative Results are Disappearing From Most Disciplines and Countries." *Scientometrics* 90 (2012): 891–904.

Friesen, L., and L. Gangadharan. "Individual Level Evidence of Dishonesty and the Gender Effect." *Economics Letters* 117 (2012): 624–26.

Gelman, A., and H. Stern. "The Difference Between 'Significant' and 'Not Significant' Is Not Itself Statistically Significant." *The American Statistician* 60 (2006): 328–31.

Gerber, A., and N. Malhotra. "Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals." *Quarterly Journal of Political Science* 3 (2008a): 313–26.

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

Hail, L.; M. Lang; and C. Leuz. "Reproducibility in Accounting Research: Views of the Research Community." *Journal of Accounting Research* 58 (2020): 519–43.

Harvey, C. R. "Presidential Address: The Scientific Outlook in Financial Economics." *Journal of Finance* 72 (2017): 1399–440.

Head, M. L.; L. Holman; R. Lanfear; A. T. Kahn; and M. D. Jennions. "The Extent and Consequences of *p*-hacking in Science." *PLoS Biology* 13 (2015): 1–15.

John, L. K.; G. Loewenstein; and D. Prelec. "Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling." *Psychological Science* 23 (2012): 524–32.

Khan, M. J.; P. C. Tronnes; B. Street; and P. C. Tronnes. "P-hacking in Experimental Audit Research." *Behavioral Research in Accounting* 31 (2019): 119–31.

Krawczyk, M. "The Search for Significance: A Few Peculiarities in the Distribution of *p*-Values in Experimental Psychology Literature." *PLoS ONE* 10 (2015): 1–19.

Lacetera, N., and L. Zirulia. "The Economics of Scientific Misconduct." *The Journal of Law, Economics, & Organization* 27 (2011): 568–603.

Libby, R.; R. Bloomfield; and M. W. Nelson. "Experimental Research in Financial Accounting." *Accounting, Organizations and Society* 27 (2002): 775–810.

Murdoch, D. J.; Y. L. Tsai; and J. Adcock. "P-values are Random Variables." *The American Statistician* 62 (2008): 242–45.

Masicampo, E. J., and D. R. Lalande. "A Peculiar Prevalence of *p*-values Just Below .05." *Quarterly Journal of Experimental Psychology* 65 (2012): 2271–79.

McShane, B. B.; D. Gal; A. Gelman; C. Robert; and J. L. Tackett. "Abandon Statistical Significance." *The American Statistician*, 73 (2019): 235–45.

Mitton, T. "Methodological Variation in Empirical Corporate Finance." *The Review of Financial Studies* 35 (2022): 527–75.

Nosek, B. A.; C. R. Ebersole; A. C. DeHaven; and D. T. Mellor. "The Preregistration Revolution." *Proceedings of the National Academy of Sciences* 115 (2018): 2600–06.

Ofosu, G. K., and D. N. Posner. "Pre-analysis Plans: An Early Stocktaking." *Perspectives on Politics* (2019): 1–17.

Ofosu, G. K., and D. N. Posner. "Do Pre-analysis Plans Hamper Publication?" *AEA Papers and Proceedings* 110 (2020): 70–74.

Olken, B. A. "Promises and Perils of Pre-analysis Plans." *Journal of Economic Perspectives* 29 (2015): 61–80.

Peterson, R. A.; G. Albaum; and R. F. Beltramini. "A Meta-analysis of Effect Sizes in Consumer Behavior Experiments." *Journal of Consumer Research* 12(1985): 97–103.

Pütz, P., and S. B. Bruns. "The (Non-) Significance of Reporting Errors in Economics: Evidence from Three Top Journals." *Journal of Economic Surveys* 35 (2021): 348–73.

Sackrowitz, H., and E. Samuel-Cahn. "P Values as Random Variables—Expected P Values." *The American Statistician* 53 (1999): 326–31.

Schafmeister, F. "The Effect of Replications on Citation Patterns: Evidence from a Large-scale Reproducibility Project." *Psychological Science* 32 (2021): 1537–48.

Scheel, A. M.; M. R. Schijen; and D. Lakens. "An Excess of Positive Results: Comparing the Standard Psychology Literature with Registered Reports." *Advances in Methods and Practices in Psychological Science* 4 (2021). doi: 10.1177/25152459211007467

Serra-Garcia, M., and U. Gneezy. "Nonreplicable Publications are Cited More than Replicable Ones." *Science Advances* 7 (2021): eabd1705.

Simmons, J. P.; L. D. Nelson; and U. Simonsohn. "False-positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (2011): 1359–66.

Simonsohn, U. "*P*-hacking Fast and Slow: Evaluating a Forthcoming AER Paper Deeming Some Econ Literatures Less Trustworthy." Datacolda (blog), September 15, 2020, https://datacolada.org/91

Simonsohn, U., and L. D. Nelson. "*P*-curve: A Key to the File Drawer." *Journal of Experimental Psychology: General* 143 (2014): 534–47.

Wicherts, J. M.; C. L. S. Veldkamp; H. E. M. Augusteijn; M. Bakker; R. C. M. van Aert; and M. A. L. M. van Assen. "Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid *p*-hacking." *Frontiers in Psychology* 7 (2016): 01832.

Wood, D. A. "Comparing the Publication Process in Accounting, Economics, Finance, Management, Marketing, Psychology, and the Natural Sciences." *Accounting Horizons* 30 (2016): 341–61.