P-hacking in Experimental Accounting Studies

Xin Chang, Huasheng Gao, Wei Li*

This version: Nov 2018

Abstract

We provide evidence that accounting experimental studies are prone to *p*-hacking. Extracting *p*-values from experimental accounting studies published on the three top accounting journals from 1990 to 2016, we find an unusual abundance of *p*-values that are just significant: the frequency of *p*-values equal to 0.05 far exceeds what would be expected based on the frequency of neighboring *p*-values. The discontinuity at 0.05 is more pronounced for articles written by untenured authors, male authors, sole author, and authors from higher-ranking schools. We also document a positive association between the extent of *p*-hacking and future citations. In contrast, we do not find a similar discontinuity from archival accounting articles. Lastly, we suggest several potential solutions to guard against *p*-hacking in experimental accounting studies.

Keywords: *P*-hacking; Experimental accounting; Statistical significance

* Xin Chang (changxin@ntu.edu.sg) and Wei Li (wli022@e.ntu.edu.sg) are from Nanyang Technological University, Singapore; Huasheng Gao (huashenggao@fudan.edu.cn) is from Fanhai International School of Finance, Fudan University, Shanghai. We are grateful for the helpful comments from the seminar participants at the Nanyang Technological University, University of Adelaide, Shanghai University of Finance and Economics, Fudan University, seminar participants in the First Fanhai Economics and Finance Workshop, and participants at the 2018 FMA Asia/Pacific conference. All errors are our own.

"*Given the competition for top journal space, there is an incentive to produce 'significant' results*".
Campbell Harvey (2017)

1. Introduction

There is growing consensus among various disciplines that due to intensifying competition for top journal spaces, researchers may shift their focus away from producing scientific findings to hit publication standards (e.g., Simmons, Nelson, and Simonsohn, 2011; Masicampo and Lalande, 2012; Head et al., 2015; Krawczyk, 2015; Harvey, 2017). Chief among the standards for publication are statistical significance, which help journals attract more readers and citations (Fanelli, 2012). When it comes to search for statistical significance, researchers can be quite resourceful. For instance, researchers can: try many tests and statistical methods, collect less or more data, combine or transform specific measures, delete or keep certain observations, employ more or fewer dependent variables or control variables, and so on. The behavior of exploiting these discretions to search and report the most significant results is, however, often considered controversial or even unethical, and is generally referred to as *p*-hacking.

P-hacking may benefit researchers and their affiliated institutions by inflating the chance of top journal publications, but it impedes progress in a research area in the long run by increasing the false positive rates. Furthermore, higher false positives rates, combined with the undisclosed decisions made by researchers during the data analysis process, will lead to higher likelihood that these findings are irreproducible in the future. Such concern is not groundless. Open Science Collaboration (2015) initiated a large-scale replications of 100 experimental studies in psychology. The main finding is somewhat surprising: only 36% replications have significant results, a decrease from 97% in the original articles.[1] Yet reproducibility is regarded as "a core principle of scientific progress" (Open Science Collaboration, 2015). Consequently, discussions of *p*-hacking and its potential remedies have attracted considerable attention, and evidence of *p*-hacking has been found in many disciplines (e.g.,

[1] Moreover, the mean effects size from the replication is merely half the mean effects size of the original results.

psychology, communication, biomedical research, finance, and economics). However, there are important cross-disciplinary variations in the extent of *p*-hacking.

P-hacking is most notable in fields that rely on laboratory experiments to conduct analyses. Such fields include psychology, pharmacology and toxicology (e.g., Head et al., 2015; Harvey, 2017; Fanelli, 2012). Several reasons have been offered to explain this phenomenon. First, unlike areas that rely on public (i.e., archival) data sources, experimental studies often create their own data sets. This gives experimentalists additional discretions regarding sample selection (Wicherts et al., 2016). Second, to the extent that these data sets are often kept proprietary, follow-up replications are costly as it requires constructing a new sample. The difficulties in replication may *ex ante* enhance the incentive to engage in *p*-hacking, because results are less likely to be cross-validated later. Lastly, studies with smaller sample size are more subject to *p*-hacking, because low statistical power makes it harder to obtain significant results (Head et al., 2015). Furthermore, results based on a small sample is particularly sensitive to data manipulations, rendering the effect of *p*-hacking on significance level more powerful.

In this paper, we investigate the prevalence of *p*-hacking among experimental accounting studies. Aside from the above reasons that experimental accounting studies may be prone to *p*-hacking, our focus on this specific area is motivated by two additional reasons. First, experimental studies in accounting has been growing and taken a stronghold in the U.S. accounting academy.[2] This is perhaps due to an important comparative advantage experimental studies have over archival-empirical research, that is, by randomizing the treatments and controlling for cofounding factors, experimentalists can easily draw causal conclusion and even identify the underlying mechanism behind a causal relationship. The growing importance of experimental accounting research strengthens the need to guard against *p*-hacking in this area. Second, a methodological advantage in focusing on experimental studies is that,

[2] Libby, Bloomfield, and Nelson (2002) discuss several factors that promote progress in experimental accounting research since 1990s.

papers in this area commonly report *p*-values, instead of *t- or F- statistics*, in text. This greatly facilitate our data collection and analyses (introduced later).

We examine the extent of *p*-hacking by revealing a discontinuity in the pooled distribution of reported *p*-values. The reliance on the *p*-value distribution to detect *p*-hacking exploits an important feature of the Null Hypotheses Significance Testing (NHST): for an effect to be considered significant in experimental accounting studies, its corresponding *p*-value must be ≤ 0.05. This arbitrary standard for significance is made to facilitate the discrimination between rejected and accepted hypotheses, a selection also dubbed as *publication bias* (see e.g., Sterling, 1959; Greenwald, 1975). However, a potential consequence of this standard is that researchers may consider it as a hurdle for their papers to be published, and rely on *p*-hacking to turn insignificant results into significant ones, just like managers engage in earnings management to turn small losses into small gains.

How does *p*-hacking introduce discontinuity into the distribution of *p*-values? Consider a researcher who works on a project that may yield results with uncertain *p*-values. Under the condition that her null hypothesis is no effect, the researcher will clearly prefer and keep *p*-values below or equal to 0.05. From the publication point of view, the remaining *p*-values can be divided into two groups, those that are "promising", i.e., close to 0.05 but not there yet, and those "beyond hope", i.e., much larger than 0.05. Suppose the researcher has some discretion in data collection and analyses, for example, she can conduct another experiment to gather more data, or delete some seemingly-outlier participants from the sample. Now if she got a *p*-value of 0.06, the marginal cost of adjusting her sample or methods to achieve a *p*-value below 0.05 should be much lower than the marginal benefit of doing so, which is a higher chance of publication. However, this may not be the case if the initial *p*-value is very large. If such a practice is widespread, we should observe an unusual high frequency of *p*-values right below or equal to 0.05, followed by a significant decline in the frequencies of *p*-values slightly larger than 0.05..

Using a sample of 2,404 *p*-values reported in 240 experimental accounting articles published in three top accounting journals (i.e., Journal of Accounting and Economics, Journal of Accounting Research, and The Accounting Review), we document evidence consistent with this prediction: the distribution of these *p*-values exhibits a sharp discontinuity at 0.05. Moreover, the frequency of "just" significant *p*-value (0.05) is higher than the frequencies of 0.04 and 0.03, as if researchers prefer "just" significant results to more significant ones. Interestingly, the frequency of 0.06 is unusually low: it is lower than the frequency of 0.07, suggesting that at least part of the marginally significant results from the articles are inflated into significant results, resulting in an over-representation of just-significant *p*-values.

It is worthwhile to point out that publication bias cannot explain the non-monotonic pattern of reported p-values, because the underlying assumption of publication bias is that lower *p*-values should be strictly preferred, resulting in a monotonic decreasing frequency of *p*-values as we move from 0 to 1. This monotonic selection turns out to be a useful benchmark that portrays the "natural" distribution of *p*-values in the absence of *p*-hacking, allowing us to decompose the observed frequency of *p*-values into two parts: the part due to publication bias, and the part due to *p*-hacking.[3]

Theoretically, any distribution that is monotonically decreasing in $[0, +\infty)$ can be potential candidates for the benchmark, we follow Masicampo and Lalande (2012) and Harvey, Liu , and Zhu (2016) to choose an exponential distribution to fit our distribution. The exponential curve matches well with the empirical distribution of *p*-values in our sample: the R-squared from the fitting exercise is on average above 98%. The discontinuity around 0.05 generates a large residual that cannot be explained by publication bias, we attribute this residual to inflation from *p*-hacking. After removing the inflation component, our conservative estimate from the fitting shows that expected frequency of 0.05 should

[3] There may be also other potential selection procedures that shape the distribution of *p*-values. For examples, authors might simply censor themselves; there is also a potential feedback effect as outlined by Henry (2009): editors and referees prefer significant results, pushing authors to generate more stars, this may further increase the requirement from editors and referees. However, the monotonic decreasing benchmark distribution are able to capture all those non-increasing selections (i.e., lower *p*-values are more likely to be selected).

be 23% less than observed. The residual of fitting for 0.05 is also much larger than the mean residuals of other *p*-values.

Having showed the existence and extent of *p*-hacking, we turn to examine what predict the extent of *p*-hacking. By exploiting the cross-sectional variations in authors' incentive and costs to conduct *p*-hacking, we identify five important author and paper characteristics. For example, we find that *p*-hacking is more prevalent among articles written by authors facing higher publication pressure (untenured researchers), authors with stronger risk-taking incentive (male researchers); moreover, our results are also stronger among the sample of sole-authored papers (lower costs of *p*-hacking), and the sample of *p*-values appeared earlier in articles (*p*-hacking for main results). These cross-sectional variations in the extent of *p*-hacking again confirm that the discontinuity is indeed caused by *p*-hacking. However, we also document stronger results among articles written by authors working in (or graduated as PhD from) higher ranking schools. Additional test also shows that more cited articles exhibit a greater extent of *p*-hacking.

While we have argued that, due to the public availability of archival data, the extent of *p*-hacking may be lower in archival research than in experimental accounting studies. We empirically investigate whether archival accounting studies also exhibit *p*-hacking. Our results show a highly right-skewed histogram of *p*-values from archival accounting studies, but we do not observe a significant discontinuity at 0.05 based on these articles. This result provides suggestive evidence that *p*-hacking problem in archival accounting studies (if any) is not as severe as experimental accounting studies. However, we still cannot rule out the possibility that archival studies regard p = 0.01, rather than 0.05, as the *p*-hacking target.

Lastly, we provide some solutions to help guard against false positives resulted from *p*-hacking. We first introduce three broad solutions developed from the statistics literature to adjust for multiple testing, and then generate simulated "experiments" with to demonstrate the application and effectiveness of these methods. The simulation results show that most of the significant findings from

p-hacking are unlikely to survive after adjusting for multiple testing or applying the Bayesian perspective. We also introduce the Bayesian approach and discuss its potential application, and provide some guidelines for journals that try to improve the disclosure transparency from authors.

The rest of this article proceeds as follows. In Section 2, we clarify a basic confusion on the definition of *p*-value and review the related literature on *p*-hacking. We introduce our sample construction and empirical strategy in Section 3, and present evidence on the existence and extent of *p*-hacking in Section 4; Section 5 further explores the factors that shape researchers' incentive to conduct *p*-hacking, Section 6 presents additional tests. Section 7 includes the proposed remedies for *p*-hacking, and we conclude in Section 8.

2. Background information and related literature

2.1. Understanding p-values: what are we hacking?

The use of *p*-value null hypothesis testing is widespread in science. Originally introduced by Fisher (1925), it is designed to separate findings from noises by measuring the extent to which the data contradicts a hypothesis (i.e., the null hypothesis, and no alternative hypothesis are defined). Formally, the Fisherian *p*-value is the probability of observing an outcome or more extreme outcomes given the *null hypothesis being true*. Users of the Fisherian approach may evaluate a hypothesis based the corresponding *p*-values, with smaller *p*-values indicating stronger evidence against null. However, such evaluation can only be applied to the *particular* data set under examination, implying that inferences based on this approach may not generalize. To draw from Neyman and Pearson (1933): "as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis".

Aware of the above limitation of the *p*-value, Neyman and Pearson (1933) came up with an alternative framework. The key argument of Neyman and Pearson is as follows. While we may not be able to learn from a single *p*-value (or any other test statistics) on the truth or falsehood of a given hypothesis, there is one way to make sure that *in the long run*, we make the right decisions most of the

time. That is, if we always follow the rule of rejecting the null when *p*-value is lower than a predetermined Type I error rate, say 0.05, then *over repetitive experiment*, we will reject the null when it's true around 5% of the time. However, over the years after the Fisherian *p*-value is combined with the Neyman and Pearson (1933) framework to form the current Null Hypothesis Significance Testing framework, *p*-value is often misinterpreted even among scientists in many fields (see, Pashler and Harris, 2012). Specifically, in NHST, the *p*-value of a specific test is calculated, and then compared with a predetermined Type I error rate (e.g., α=0.05). If *p*-value < α, it is usually interpreted as evidence rejecting the null hypothesis and accepting the alternative; otherwise, accept the null and reject the alternative. This comparison creates one confusion: people often treat *p*-value as equivalent to the Type I error rate (Harvey, 2017). The confusion would temper many to think that a *p*-value gives the probability that the null hypothesis is true, or the probability that a result is driven by chance alone.[4] This is the most basic mistake in the use of *p*-value.

A better understanding of *p*-value enables us to say more on the nature and consequence of *p*-hacking. Consider the following practices of *p*-hacking from a researcher, for example, running many experiment sessions and cherry-pick the most significant session to report, collecting multiple (related) outcome variables and choose the one that gives the more significant results. Finally, the researcher end up with a *p*-value of 0.01. This low *p*-value resulted certainly does *not* mean that there is only 1% percent probability that the null hypothesis is true, or that if other researchers repeat the same experiment for a large number of times, they will obtain significant result 99% of the time. Quite differently, the *p*-value of 0.01 merely means that, given the null being true, there is only 1% probability that we observe the result or more extreme results. However, in this example, the extreme result obtained by the researcher is very likely due to *p*-hacking: because the researcher have sampled

[4] In mathematical terms, the definition of *p*-value can be written as $P(D|H_0)$, which is the probability of observing an effect (or more extreme) *D* given the null is true. However, *p*-value is often misinterpreted as $P(H_0|D)$, which is the probability of the null being true given an observed (or more extreme) effect *D*.

the data long enough and hard enough, he or she is almost bound to find some extreme observations.[5] Therefore, in the end, *p*-hacking will lead to higher false positive rates.

2.2. Literature on p-hacking

The literature generally defines *p*-hacking as the process of exploiting undisclosed flexibility in data collection and analyses to cheery-pick the most significant results, while filing away the rest (for examples of *p*-hacking, see e.g., Simmons, Nelson, and Simonsohn, 2011; Head et al., 2015; Harvey, 2017).[6] The concerns over the negative effect of *p*-hacking can be traced back to Sterling (1959), who point out an unconscious type of *p*-hacking. When journals only publish papers with significant results (i.e., the publication bias), those experiments that do not work will not be observable to other researchers, who may independently repeat such experiments until a significant finding occurs by chance. The consequence is a higher false positive error than expected.

While *p*-hacking has been recognized for many years, the concern that it can be a widespread problem has only recently aggravated. Two main reasons are discussed in the literature. First, there is increasing bias against negative findings over the recent decades. Fanelli (2010) documents that in social science (such as Economics and business, psychology), it is more difficult to see published results that do not support the main hypothesis developed in the paper than in other "hard" science (such as Space Science and Physics). Fanelli (2012) and Head et al. (2015) point out that negative results are disappearing and significant ones are more likely to be published in many fields. Second, the competition for top journal space may have intensified. Researchers and institutions are increasingly evaluated based on their top journal publications and the citations these papers generate (Young, Loannidis, and Al-Ubaydli, 2008; Statzner and Resh 2010), which enhances the incentive to publish novel and more significant findings in top journals. This bias against negative findings,

[5] There are also other reasons why *p*-value can be low. For example, it could be due to the poor ability of the hypothesis to explain the data, or a misspecified test.

[6] In addition to *p*-hacking, the practice of trying many tests and analyses until none significant results become significant has received other names, for example, data dredging or data fishing (Selvin and Stuart, 1966) , data snooping (White, 2000).

combined with the increasing competition for top journal space, likely contribute to researchers' stronger incentive to conduct *p*-hacking.

The growing concerns over the widespread of *p*-hacking have accelerated research on it in many areas. As we have outlined in the introduction, *p*-hacking induces kink around the significance threshold (0.05) in the cross-sectional distribution of *p*-values. Masicampo and Lalande (2012) first apply the distribution of *p*-values to examine *p*-hacking in experimental psychology. Simonsohn, Nelson, and Simmons (2014) term this method as the use of "*p*-curve" and further promote its usefulness to detect *p*-hacking, which is similar to the use of distribution of earning numbers to detect earnings management in accounting research (see, e.g., Burgstahler and Dichev, 1997). This method subsequently stimulates *p*-hacking research in many fields. Head et al. (2015) employ this method to study the extent of *p*-hacking in 14 disciplines and find that it is widespread. Krawcyzk (2015) documents too many "just significant findings" in experimental psychology from more than 5,000 papers. Vermeulen et al. (2015) take this method to communication science, and document similar evidence. Moreover, they also find that 8.8% of the 5,834 *p*-values in their paper are misreported, 74.5% of which are too low. Brodeur et al. (2016) apply similar methodology to articles published on top economics journals, and document strong evidence of *p*-hacking. They also explore how the extent of *p*-hacking varies across several author and article characteristics.

The paper closest to ours is Khan and Tronnes (2018). They specifically focus on experimental audit research and find a kink in the distribution of *p*-values. Our paper extents theirs in the following ways. First, we include all topics in experimental accounting research, and focus only on articles appeared in top accounting journals. It may be more unsettling that *p*-hacking is widespread in even the most prestigious journals. Second, Khan and Tronnes (2018) exclude vague *p*-values (e.g., $p < 0.05$) in their sample, which may inflate the relative frequency of 0.05 and extent of *p*-hacking if many such vague *p*-values have true values close to 0.04 or 0.03. Our paper addresses this issue by empirically estimating the distribution of vague *p*-values using their underlying test statistics and degree of freedom.

Third, as mentioned previously, we conduct rich cross-sectional analyses that helps to understand the determinants of *p*-hacking, and especially how *p*-hacking affects future citations. Lastly, we also briefly examine evidence of *p*-hacking in archival papers. It is reassuring that archival research does not display a discontinuity around 0.05 in the distribution of *p*-values.

Furthermore, our article extend the above works by making serious attempt to provide statistical solutions that may help guard against *p*-hacking. We demonstrate the usefulness of procedures to adjust for multiple comparisons when the chance of false positives are high, and also introduce the Bayesian approach as a potential alternative method. However, we also emphasize that transparent disclosure from researchers is necessary to facilitate the implementation of the strategies we have proposed.

3. Our sample and empirical strategy

3.1. Sample construction

Our aim is to include in our sample as many experimental accounting studies from influential accounting journals as possible. We thus use a computer program to search among all articles on three top publishers of experimental studies ─The Accounting Review (TAR), Journal of Accounting Research (JAR), and Journal of Accounting and Economics (JAE) ─ from 1990 to 2016.[7] We record an article if its abstract contains keyword "experiment", but not "(quasi-) natural experiment".[8] A total of 311 papers meet this criteria, we then download them for further data collection. We exclude articles that do not mention any *p*-values in their text (27 articles), and articles that are retracted due to Dr Hunton's fraud (6 articles). In the end, we have 281 articles in our final sample, with 193 from TAR, 84 from JAR, and 4 from JAE, accounting for 14.9%, 10.5%, and 0.3% percent of the total number of articles published in TAR, JAR, and JAE during our sample period, respectively.[9] In our main sample,

[7] Our sample period starts from 1990 because that's when Dallas Ranking for accounting divisions across the world starts to be available, which we will need to conduct subsample analysis.

[8] Alternatively, we can also rely on JEL Classification code to distinguish experimental studies from archival studies, but this information is not provided on the website of the two journals.

[9] When counting the total number of articles, we exclude any articles that are generated by the comments and/or discussions of existing papers. In addition, TAR frequently publishes book reviews in our earlier sample period, these are also excluded from the count.

we also exclude papers that only report vague *p*-values such as $p < 0.05$ (41 articles). The main sample thus contains 240 articles.

To analyze what affect the extent of *p*-hacking, we further collect a series of author and article characteristics that may be related to researchers' marginal benefit and cost of *p*-hacking. Researchers with greater career concern are likely to have stronger incentive to conduct *p*-hacking (Fanelli, 2012). We thus measure researchers' career concern by calculating the proportion of tenured authors for an article (*Tenured*), and by counting the average number of years since a researcher earns PhD at the time of the article's publication (*Experience*). Head et al. (2012) state that prestigious schools impose higher tenure hurdles and thus publication pressure on researchers. To measure academic ranking, we use the mean ranking of authors' affiliated institutions (*School Ranking*), as well as the mean ranking of the institutions where each author earns PhD degree (*PhD School Ranking*). We distinguish articles with no female author(s) from these with at least one (*Female*), as males tend to be more risk-taking than females (see, e.g., Byrnes, Miller, and Schafer, 1999). To account for the coordination costs among researchers, we also count the number of authors for each article (*N_ Authors*). We retrieve the total number of participants in an article's experiment(s) as various *p*-hacking "techniques" involves sample manipulation (*N_ Participants*). Because researchers may revise their hypotheses *ex post* as directional in order to report lower one-tailed *p*-value, we construct a dummy variable equals to one if all *p*-values in an article are two-tailed, and zero otherwise (*Two-tailed*). Lastly, we collect the number of citations for all articles from both Google Scholar (*Google_Citation*) and Web of Science (*Web_Citation*) to examine the relation between *p*-hacking and citations.

The variables described above come from various sources. To obtain tenure, graduation, and school ranking information for each researcher, we search for "author name" + "affiliated school" on google. Most of these searches deliver an author's latest CV, from which we can determine the year of tenure and graduation. We then collect information on school ranking from The UTD Top 100 Business School Research Rankings. Since our focus is experimental accounting studies, we are interested in

the ranking of accounting division of a specific university, we thus choose to rank the schools by the number of accounting top journal publications in a given year.[10] Note that for each article-author-school combination, the ranking for that school is two years before the publication of that article. We introduce this two-year lag to capture the ranking status while a paper is written rather than when it's published. If a school is not on the ranking list, we replace its ranking with the lowest ranking number on the ranking list in a specific year. Finally, if we cannot find CV for an author, we rely on other credible sources to obtain as much information as possible.[11] We manually check each article to obtain the number of participants, in articles where multiple experiments are conducted, we sum the number of subjects from the experiments. Lastly, we read all articles to determine whether all *p*-values in a paper are based on two-tailed tests.

Table 1 offers a glimpse of the characteristics of experimental studies in our sample. An average article in our sample has 2.2 authors, 117 subjects in the experiments, and mentions 18 *p*-values in its text. About 50% of the articles contain at least one female author, 34.3% of the articles report only two-tailed *p*-values, and 28.7% are sole works. In addition, an average paper is published 9 years after its authors obtain their PhD degree. When we break the total sample by journals, we notice that about two third of the articles in our sample comes from TAR, which publishes more behavioral research than JAR does.[12] We also find that the variables do not exhibit big differences between two journals.

[Insert Table 1 about here]

3.2. Methodology

After transforming all downloaded pdf files into txt format, we use a Python script to search for numbers between 0 and 1 immediately preceded by "p =", "p-value =", "$p_{two-tailed}$ =", and "$p_{one-tailed}$ =", regardless of spacing or capitalization. Unlike prior studies such as Head et al. (2015) and Khan and Tronnes (2018), we also collect *p*-values after inequality operators, such as "p <", or "p

[10] We choose three top journals in the accounting field: TAR, JAR, and Journal of Accounting Economics (JAE).
[11] Those sources include school website, LinkedIn, and BYU Accounting author search, etc.
[12] Differences in number of observations among variables are due to missing values.

≤". Importantly, we omit any numbers that indicate significance thresholds instead of real *p*-values in an article; for example, numbers appeared in sentences such as "*, **, *** Indicates p< 0.10, p<0.05, and p<0.01, respectively" are discarded. Our script cannot retrieve *p*-values within tables, because the formatting of tables in pdf is messy and varies from article to article, rendering the search impractical. However, by focusing only on *p*-values mentioned in the text, we ensure that a *p*-value is not recorded twice in our sample; furthermore, researchers are more likely to mention in text their main results' significance level than other covariates', this suits the examination of our hypothesis that researchers have the strongest incentive to engage in *p*-hacking for their *key* findings.

Our initial search yields 6,144 *p*-values. However, not all of these *p*-values are informative about the extent of *p*-hacking. We are particularly interested in any irregularities in the cross-sectional distribution of the *p*-values at points such 0.05, and 0.10, and thus exclude all *p*-values larger than or equal to 0.155, which leaves 4,766 *p*-values in our sample. Because we cannot uncover the exact *p*-values for all *p*-values reported after inequality operators (e.g., < or ≤), we focus on exact *p*-values in our main sample, which includes 2,404 (50.4%) *p*-values from the 4,766 *p*-values.[13] We then pool these *p*-values across articles and time to generate a distribution of *p*-value.

It is worthwhile to emphasize two advantages of our text-scraping methodology. First, it does not require us to have further knowledge about the articles in our sample, such as the topic of interest, hypotheses tested, and design of experiments, allowing us to examine all experimental articles that study any accounting topics. Second, since it is a substitute for hand collection, it minimize biases from data collectors such as research assistants, either consciously or unconsciously, to collect data that conforms to this paper's prediction.[14]

[13] Many *p*-values reported after inequality operators are placed in footnotes for robustness check or additional tests, where no information about the test statistics or degree of freedom are provided. However, if simply we pool these *p*-values with our main sample, the main results will only become stronger, because "<" and "≤" are commonly followed by threshold value 0.05. We present analyses to deal with these *p*-values in Section 4.

[14] Nonetheless, one may argue that our script can still produce some noises, for example, it may capture some irrelevant numbers, but ignore *p*-values for an article's key results. To alleviate such concerns, we randomly select 30 articles from our sample, and for each article, we identify all the hypotheses tested. We then check if the *p*-values extracted by the script from an article are the significance level for hypotheses testing in that article. Of the 260 *p*-values from the 30 articles, 251

4. Empirical results

We present two sets of results to determine whether *p*-hacking to turn insignificant results into significant exist or not. First, we provide graphical evidence in the form of histogram on the cross-sectional distributions of the *p*-values. Second, we fit the distribution of *p*-values to an exponential distribution to quantify the extent of *p*-hacking.

4.1. The cross-sectional distribution of p-values

We first plot a histogram using all *p*-values in our sample, and choose a bin width of 0.001. The result is shown in Panel A of Figure 1. One bin stands out from the distribution: there is an obvious bump in the frequency of interval right below the significance threshold, i.e., (0.049, 0.05], which we highlight using bold bar. The frequency of this interval (122) is even higher than the frequency for interval (0.039, 0.04] and (0.029, 0.03], which are 88 and 111 respectively, as if less significant *p*-values are preferred for publication. Interestingly, we observe that the frequency of *p*-values right below 0.06 is lower than that below 0.07. The two discontinuities around the conventional significance threshold cannot be explained by the publication bias, but is in line with the presence of *p*-hacking to make marginally significant results become significant.

As noted in Table 1, the average article reports 10 *p*-values in our sample, however, there is substantial variation among articles in terms of the number of *p*-values: the highest number is 64, while the lowest is 1. Furthermore, the number of *p*-values reported (or number of tests conducted) may be related to the overall significance level or robustness of the results: a researcher may conduct and report more tests when these tests have *p*-values below the significance threshold. Therefore, it is possible that certain articles that report many *p*-values are driving the main results. To alleviate this concern, in Panel B we weight each *p*-value by the inverse of the number of *p*-values reported in an article, thus giving lower weight to those articles with many *p*-values. The discontinuity around threshold becomes

(96.12%) are the significance level of the main hypotheses tested, the rest 9 *p*-values belong to 6 articles. Among the 6 articles that do not have *p*-values for the main hypotheses, 2 do not discuss the *p*-values in their text, the other 4 do not report exact *p*-values (i.e., $p < 0.05$). These results suggest that our script capture relevant *p*-values generated from hypotheses testing in experimental studies, and that our script are unlikely to generate systematic biases in our sample.

more striking after weighting: the weighted frequency of *p*-values right below 0.05 is even higher than that of *p*-values right below 0.02. Meanwhile, there is also a non-monotonic pattern in the distribution around 0.10, as if it is more desirable than 0.09 and 0.08. Overall, the weighted distribution of *p*-values shows a greater extent of *p*-hacking. This result seems to suggest that articles report *less p*-values tend to report a higher fraction of "just" significant results.

[Insert Figure 1 about here]

The Figure also shows that articles in our sample prefer reporting *p*-values with two decimal places, which accounts for 40% of the total number of *p*-values, despite there are only 16 numbers with two decimal places below or equal to 0.15.[15] This creates the indented histograms we observe in Figure 1. In Panel A of Figure 2, we use only *p*-values reported with two decimal places to plot another histogram. The resulting histogram shows that the unusual high frequency of p-values in interval (0.0475, 0.05] in Figure 1 is entirely driven by *p*-value = 0.05, which was found 122 times in our sample. The same also holds for *p*-value = 0.04 and 0.03.

In Panel B, we round all *p*-values to two decimal places, count the frequency of the 16 p-values (0.00 to 0.15), and plot another *p*-curve. Similar pattern emerges, though the difference in frequency between 0.05 and 0.04 is less salient. This is unsurprising because rounding effectively sets the bin width to 0.01, thus muting the acuteness of the histogram to detect *marginal incentive* of *p*-hacking at 0.05.

[Insert Figure 2 about here]

In sum, the results presented in this subsection provides evidence on the existence of *p*-hacking among experimental accounting studies to inflate the significant results. Since 0.05 is the maximum *p*-value for an effect to be considered as statistically significant in experimental accounting research, the relative abundance of 0.05, together with the relative scarcity of 0.06, provides strong evidence that

[15] Among the 2,404 *p*-values, the decimal places of a *p*-value range from one to five, the frequencies of the five different lengths are 2, 1013, 1257, 130, and 2, respectively.

there are too many results with "just significant" *p*-values, indicating target beating behavior. Such irregularity in the distribution of *p*-values present challenges for any other potential explanations.

4.2. Fitting the p-curve to an exponential distribution

In order to gauge the extent of *p*-hacking among our sample articles, in this subsection, we follow Masicampo and Lalande (2012), and Harvey, Liu, and Zhu (2016) to fit the histogram of *p*-values to an exponential distribution. The choice of a rapidly decreasing exponential curve aims to approximate the effects of strong publication bias, which is substantiated by the highly right-skewed distribution of *p*-value in previous section. In doing so, we are able to roughly estimate the *expected* frequency of *p*-values without *p*-hacking, and quantify its impact on the magnitude of inflation.

Because *p*-values with two decimal places are more frequently reported, in order to have a smooth histogram for fitting, we only use the two histograms in Figure 2 for this exercise. The results are presented in Figure 3. In Panel A, the sample only includes *p*-values with two decimal places. An exponential curve fits the histogram quite well except for two p-values: 0.05 and 0.06, the former lies well above, and the latter well below, the curve. Another less noticeable discontinuity appears in *p*-value 0.10. The frequency of observing 0.11 is merely half the frequency of observing 0.10. We posit that the reason for this discontinuity is that 0.10 is also a target to beat: *p*-values ϵ (0.05, 0.10] are deemed marginally significant, while *p*-values > 0.10 are insignificant. The importance of this target, however, is subordinate to 0.05. We observe similar pattern at Panel B, where all *p*-values are rounded to two decimal places.

[Insert Figure 3 about here]

We further tabulate the fitting results in Table 2. For each, *p*-value from 0.01 to 0.15 (i.e., 0.01, 0.02, 0.03… 0.15), we report its observed frequency and expected frequency implied by the fitting from Figure 3, and calculate how much the former deviates from the latter. In Panel A of Table 2, where only *p*-values reported with two decimal places are used, we observe that the residual at 0.05 is the highest. However, residuals are also sizable at *p*-values such as 0.03, 0.04, and 0.06, all of which

are negative. This suggests that compared with *p*-value = 0.05, *p*-value = 0.03 and *p*-value = 0.04 should have higher frequencies than what we observe here. We observe similar results using all *p*-values rounded to two decimal places in Panel B.

[Insert Table 2 about here]

We conduct two post hoc pairwise comparisons to determine if the residuals from fitting for 0.05 and 0.06 are significantly different from the mean residuals for the rest *p*-values in the two histograms. The comparison is based on Dunnett's Test (pairwise versus reference group).[16] Specifically, we treat residuals for *p*-values except 0.05 and 0.06 from exponential fitting as reference group, and compare if the residuals for 0.05 and 0.06 is significantly different from the mean residual of this reference group. The significance level obtained is adjusted by times of comparison. Results are reported in Table 3. Panel A and B of Table 3 are based on the fitting in Panel A and B in Figure 3, respectively. Despite the small sample size (16 residuals), we find that in both Panels, the residuals at 0.05 are always much larger than the reference group. Residual at 0.06 is much smaller than the reference group in A, though it is statistically insignificant; in B, residual at 0.06 is indistinguishable from reference group.

In terms of magnitude, the fitting shows that the expected frequency of 0.05 should be 24.88% to 26.72% lower than the observed frequency; for 0.06, the expected frequency should be 13.16% to 38.46% higher than the observed. As we stressed in section 4.1., histogram based on rounded *p*-values is not sharp enough to detect the strongest incentive to engage in *p*-hacking at 0.05. Thus the results in

[16] Dunnett's test is a many-to-one comparison procedure that compares the mean of every treatment group procedure against the mean of a single control group. Each comparison results in a test statistics that is calculated as below:

$$\frac{\bar{X}_i - \bar{X}_c}{\sqrt{2 \times MS_{error}\left(\frac{1}{n_i}+\frac{1}{n_c}\right)}},$$

where $\bar{X}$ refers to the mean of each group, *i* indexes each treatment group, and *c* the control group. MS_{error} is the mean squared error term for the whole sample. Lastly, n_i and n_c refer to the number of observations for treatment group *i* and control group. The degrees of freedom for this test, $N - k$, equal to the total number of observations in all groups minus the number of groups (including the control). The *p*-values from Dunnett's test are determined by looking up the test statistics in a table of critical values that already adjusted for family-wise error rate. Compared with other post hoc pairwise comparison methods (e.g., Bonferroni, Tukey's and Scheffé's methods), Dunnett's test is more powerful as it conducts fewer comparisons (if there are *k* groups, Dunnett's test makes *k*-1 comparison while others make $2 \times k \times (k\text{-}1)$ comparisons), and thus does not result in unnecessarily low critical *p*-values. See Dunnett (1955) for more details..

Panel B of Table 3 can be viewed as a lower bound for our results. Taken together, the results in this subsection show that there is a large proportion of just-significant *p*-values that cannot be explained by a decreasing selection process alone, suggesting that *p*-hacking has a material impact on the distribution of *p*-values.

[Insert Table 3 about here]

4.3. P-values reported after inequality operator

In our initial sample, 2,362 *p*-values scraped from research articles are reported after inequality operator (e.g., $p < 0.05$), we term these numbers "vague" *p*-values. The distribution of these *p*-values, which is reported in Panel A of Figure 4, shows that the majority of numbers reported after "<" are significance thresholds such as 0.001, 0.01, 0.05, and 0.10. The frequency of these four numbers accounts for 69.98% percent of the 2,362 *p*-values. We are particularly interested in *p*-values reported as $p < 0.05$ and $p < 0.10$, since redistributing the true values behind the two numbers may invalidate our main conclusion. For example, *p*-values reported as $p < 0.05$ can be any number between 0.00 and 0.05, excluding these *p*-values may cause an under-representation of *p*-values such as 0.03 and 0.04 in our main sample, thus causing the discontinuity around 0.05 we observed in the distribution of *p*-values.

[Insert Figure 4 about here]

To alleviate this concern, we attempt to calculate the true values for *p*-values reported as $p < 0.05$ and $p < 0.10$ (396 such cases), by using the test statistics and degree of freedom information provided in research articles. However, we notice that in many cases, such information is not provided.[17] In the end, we are able to reveal the exact *p*-values for 104 out of the 396 cases. While it is infeasible reveal all of the exact values for such *p*-values, the distribution of the 104 *p*-values may provide some insights on when researchers prefer to report "p <" instead of "p =".

[17] There are several reasons for not reporting such information. For example, many p-values are reported in footnotes for additional tests or robustness checks; multiple *p*-values are reported together as "all *p*-value < 0.05". Papers published in the beginning of our sample period are more likely to omit such information.

As shown in Panel B of Figure 4, the distribution of the 104 exact *p*-values indicates that, for a *p*-value smaller than 0.05, researchers are more likely to report it as "$p < 0.05$" when it is closer to 0.05 than to 0.00. This is reasonable, since an investigator has strong incentive to report a low *p*-value (e.g., 0.02) *exactly* to demonstrate statistical significance, whereas she may find it in her interest to hide the exact value of a *p*-value and its test statistics when it is very close to 0.05 (e.g., 0.0486). This selective reporting incentive suggests that for the rest *p*-values that we have no information to calculate exact values, the distribution of their exact values are likely to be even more concentrated near 0.05 than the distribution Figure 4. We then impose the same distribution in Figure 4 to the rest 292 p-values, and stack the estimated distribution of the 396 vague p-values (i.e., $p < 0.05$ and $p < 0.10$) with the distribution of exact *p*-values used in our main sample. The resulting distribution is reported in Panel A of Figure 5, which shows stronger evidence of *p*-hacking than using the sample of exact p-values alone. This is consistent with our expectation that a *p*-value slightly below 0.05 is more likely to be reported vaguely than one that is far below 0.05.

While our results is quite robust to including vague *p*-values reported as $p < 0.05$ and 0.10, there are also some *p*-values reported as $p < 0.04$ ($p < 0.03$), and if the exact values of these *p*-values are quite close to 0.04 (0.03), omitting them may again inflate the extent of *p*-hacking. We repeat the procedure to manually compute the exact values for $p < 0.04$ and 0.03. The data shows that exact *p*-values behind $p < 0.04$ (0.03) are approximately uniformly distributed between 0.03 and 0.04 (0.02 and 0.03). The uniform distribution is then applied to all other vague *p*-values except $p < 0.001$, $p < 0.01$, $p < 0.05$, and $p < 0.10$. For instance, the exact value of a $p < 0.09$ can be any number between 0.08 and 0.09. We further include these redistributed vague *p*-values (i.e., $p < x$, where $x = 0.02, 0.03 \ldots$ and 0.15, excluding 0.05 and 0.10) to the sample used in Panel A. By doing so, we utilize all of the *p*-values (4,766) scraped from experimental accounting articles to plot a new distribution, which we report in Panel B of Figure 5. After including all vague *p*-values, we still observe strong evidence of *p*-hacking. For example, if we fit the histogram on the right side of Panel B to an exponential

distribution, the expected frequency at 0.05 will be 24% less than what we actually observe. Overall, the main message from Figure 5 is that, after taking into account the impact of vague *p*-values on the distribution of *p*-values, our main results stand valid.

[Insert Figure 5 about Here]

5. What explain the extent of *p*-hacking?

In this section, we study the following question: *ex ante*, what help potential readers (including referees and editors) to reasonably forecast the extent of *p*-hacking? We hypothesize that when determining the extent of *p*-hacking, researchers weigh the potential benefits of *p*-hacking against the costs of doing so. Main benefits include higher chance of publication, more influence, and better career opportunities for authors; potential costs include coordination costs among authors, availability of discretion in data analyses, and reputation costs. Based on this cost-benefit framework, we propose and test 7 factors using all *p*-values rounded to two decimal places.[18]

5.1. The benefits of conducting p-hacking

Researchers are highly rewarded for top journal publications, such rewards can be of various forms. Tenure is perhaps the most important reward at the early stage of a researcher's career. By granting a permanent post for a researcher, tenure greatly mitigates career concern. Therefore, untenured researchers are under greater pressure to publish, especially in research institutes with higher tenure requirements. We posit that *p*-hacking to increase the chance of publication should be most prevalent among researchers that have not tenured yet, hence the kink around 0.05 on the *p*-value histogram will be sharper for articles written by such researchers. We classify our sample articles into two groups. The first group include articles whose authors are all non-tenured, and the second group includes the rest.[19]

[18] Subsample comparison based on tail choices of *p*-values differences in extent of *p*-hacking between two subsamples, and thus are omitted for the sake of space.

[19] We determine an author to be tenured only if he or she is tenured two years before the publication of the article of interest.

Next, we examine whether institutions' rankings ─ as measured by their numbers of top journal publications ─ are related to the extent of *p*-hacking. As Fanelli (2012) noted, research institutions and other organizations are increasingly evaluated by the number of top journal publications produced by their researchers. It is likely that high ranking institutions have higher tenure requirements, which imposes additional pressure on researchers to publish more papers in top journals, and consequently, they engage in greater extent of *p*-hacking.[20] Institution ranking is measure by the number of publication in three accounting top journals, obtained from "The UTD Top 100 Business School Research Rankings". For each article published in year *t*, we use the year *t*-2 *School Ranking* as defined in Section 3.1, and separate articles into high and low institution ranking based on its sample median. We also repeat the above procedure to split sample based on variable *PhD School Ranking*.

5.2. The costs of p-hacking

P-hacking is not costless. In particular, it can be perceived as a type of "publication risk-taking": as the extent of *p*-hacking intensifies (e.g., manipulate data to a greater extent, with the extreme case being data fabrication), researchers might be able to convince potential readers and referees with more starring results; however, once being detected, the authors may lose credibility and reputation. Consequently, the extent of *p*-hacking will be higher when researchers are less risk-averse. Moreover, *p*-hacking may require teamwork and coordination among co-authors. For example, researchers co-authoring a paper may need to concur with each other on how to identify and treat outliers, whether to collect more data, and whether to combine or transfer specific measures. If one (undisclosed) decision is controversial but yields the most significant results, the author who is in charge of data analysis might need to convince other authors why they should make that choice. The costs of *p*-hacking thus increase with the need for coordination and communication.

[20] Alternatively, researchers affiliated with high ranking schools might be able to come up with research questions that have larger true effects, or they may simply be more rigorously trained, thus reducing the need and incentive to engage in p-hacking.

We use two simple procedures to test the significance of the two type of costs outlined above. First, we separate articles by whether their author(s) contain female researcher. Prior works document that males tend to exhibit stronger risk-taking incentive under various scenarios (see, e.g., Byrnes, Miller, and Schafer, 1999; Charness and Gneezy, 2012, and Faccio, Marchica, and Mura, 2016). Hence, it is reasonable to suspect that male researchers are more willing to trade the potential uncertainties for higher chance of publication, and that articles writing by them exhibit more aggressive *p*-hacking. Second, we distinguish sole-authored paper from the rest. Compared with authors that collaborate with each other, sole authors can safely avoid the coordination costs of *p*-hacking, as he or she no longer needs to disclose the detailed research process to others. The above discussion leads to the following prediction: male researchers and sole-authors will show more aggressive *p*-hacking.

5.5. Other article characteristics

Lastly, we study whether several article characteristics are related to the extent of *p*-hacking. Specifically, we focus on the number of participants in an article's experiments, and the citations an article receives. First, as mentioned before, researchers can decide whether to conduct additional sessions after seeing their results. For example, researchers can conduct one more experiment session if the initial results are weak but not significant yet, they can also stop collecting data right after obtaining significant results. We presume that those behaviors will result in more participants in a studies' experiment(s). Put differently, higher number of participants is the consequence, rather than the cause, of *p*-hacking. Second, we check whether an article's tail choice ─ reporting one- or two-tailed *p*-values ─ is related to the extent of *p*-hacking. For a *t*-test, its one-tailed *p*-value equals to two-tailed *p*-value divided by 2, this may explain the finding that, in our sample, around 64% of the articles report one-tailed *p*-values (see Table 1), though the commonly cited reason for reporting one-tailed *p*-values is that those articles' hypotheses are directional. Because dividing the *p*-values by two will shift the distribution towards zero, without affecting the relative frequency of different *p*-values, we do not have a stance on whether and how tail choice will make the frequency of 0.05 more outstanding.

We test the cross-sectional variations in *p*-hacking by fitting the distribution of *p*-values for each pair of subsamples to an exponential distribution, and report the magnitude of their residuals (forecasted minus observed frequencies) at 0.05 in Table 4. The findings are in general consistent with the predictions derived from our cost-benefit framework. For example, the residual at 0.05 from the subsample of articles without tenured author exhibit is much higher than that from articles with tenured authors (38% versus 9%). Surprisingly, we observe stronger evidence of *p*-hacking in the subsample of articles written by authors affiliated with (or graduated as PhD from) higher ranking institutions. While it is hard to pin down the exact reason, a plausible explanation is that more pressure to publish in higher ranked institutions enhances the incentive to conduct *p*-hacking. Alternative, one could also argue that researchers from more prestigious institutions are less likely to be challenged by reviewers and other readers.

[Insert Table 4 about Here]

Turning to the costs of *p*-hacking, we document stronger evidence of *p*-hacking when it is less costly for researchers. Specifically, the residual for the subsample of articles without female author is almost twice the magnitude of the residual among articles with at least one female author, consistent with males having higher risk tolerance, and are thus more aggressive in *p*-hacking; further, researchers tend to be more aggressive in *p*-hacking when they work alone, as indicated by the much larger residual at 0.05 for sole-authored papers than articles with more than one authors (36% versus 14%). Lastly, as for other article characteristics, more participants in experiments is associated with a higher residual at 0.05, this suggests that *p*-hacking may result in a larger sample; however, after splitting articles that only report two-tailed *p*-values and those that do not, we find quite similar residuals between the two groups.

Taken together, we identify five factors that seem to be able to predict the extent of *p*-hacking, they are: authors' tenure, gender, number of author, authors' affiliated school ranking, and number of participants in the experiments of an articles. These cross-sectional variations in the extent of *p*-hacking

is consistent with the predictions from our cost-benefit framework: we will see more *p*-hacking when its benefits outweigh its costs. Consequently, the subsample analyses provide further support to our argument that the residuals around significance threshold is due to *p*-hacking. However, our results in this section is purely based on univariate analysis, and the results obtained can be relational instead of causal. It is possible that those factors are correlated with each other or other omitted variables. Our results thus need to be interpreted with these limitations in mind. Nonetheless, while some alternative stories may explain part of our results, the collection of the cross-sectional evidence is most consistent with our cost-benefit framework of *p*-hacking.

6. Additional tests

6.1. P-hacking and future citations

If *p*-hacking increases the chance of publication, an interesting question is whether it is related to the citations an article receives in the future. *A priori*, it is unclear how the two will be correlated. On one hand, to the extent that editors prefer articles with results that are more statistically significant because such articles attract more citations in the future, we may expect *p*-hacking increase the number of citations by inflating the level of statistical significance. On the other hand, if *p*-hacking increases the chance of false positive findings, these findings are less likely to hold in the future (Open Science Collaboration, 2015). Subsequent studies that build on these findings thus have lower chance to bear fruit and to publish. Hence, findings coming from *p*-hacking or data-mining may generate fewer citations than other robust findings.

We use the citation counts from Google Scholar to test how the extent of *p*-hacking is related to the number of future citations. To control for the truncation bias that articles published earlier have longer time to accumulate citations, in each year, we split all articles published during that year into two groups based on that year's median number of citations. The final *high* (*low*) *citations* group include all the *high* (*low*) *citations* groups from every year. We then check whether more cited articles

exhibit stronger or weaker evidence of *p*-hacking by plotting the distribution of *p*-values.[21] The histograms in Figure 6 deliver two findings. First, not surprisingly, more cited articles have more significant results, as indicated by the higher frequency of *p*-values near zero in the *high citation*s group than that in *low citations* group. Second, more cited articles also have a relatively higher proportion of "just" significant results: the ratio of the residual frequency to observed frequency in Panel B (24%) is almost twice the ratio in Panel A (13%). Similar results are obtained if we use the citation counts from Web of Science instead of Google Scholar (Panel I in Table 4). Taken together, these results are consistent with the argument from prior works that more significant results attract more readers and citations (e.g., Fanelli, 2012), though some of the results may be due to *p*-hacking.

6.2. Extent of p-hacking for main results

If researchers use *p*-hacking to obtain the most significant results, then they should have the strongest incentive to do so for their *key* results in the article. To test this conjecture, we slice the *p*-values in our sample as follows. For an article, we rank its *p*-values by the sequence they appeared in the text, and then group *p*-values appeared in the first half of the whole sequence as *p*-values for the key results. The rationale is that generally speaking, the key results of an article are likely to be mentioned earlier rather than later.

Results in Figure 7 demonstrate that the kink in the *p*-value histogram at 0.05 in our full sample is almost entirely driven by the subsample of *p*-values mentioned in the earlier part of each article. Moreover, in Panel A, the frequency of *p*-values at 0.01 is almost twice the frequency of *p*-value = 0.02, which is even lower than the frequency of 0.05. The potential explanation is that researchers have higher standard of significance level for their main results: many *p*-values equal to 0.02 are inflated to 0.01. In contrast, in Panel B, where the *p*-values are mentioned in the later part of each article, we

[21] Ideally, to test how *p*-hacking may affect future citations, one needs to construct an article-level measure of the extent of *p*-hacking, and use this measure to forecast future citations. However, due to the limited number of *p*-values a typical article reports (10 in our sample), article-level measure is hard to construct. That said, the method we adopt is still legitimate, because future citations are usually generated years after the writing of an article, it is implausible that the causal relation between *p*-hacking and future citations runs from the latter to the former.

observe weaker evidence of kink at 0.05. The results from this exercise thus provide strong evidence of *p*-hacking.

[Insert Figure 4 about here]

6.3. P-hacking among archival accounting research

We have documented the existence of *p*-hacking among experimental accounting studies. Potential readers of our paper will naturally concern that *p*-hacking also exists among archival accounting research, given that archival accounting researchers also have flexibility over data analyses and strong incentive to generate top journal publications. To answer this question, we randomly select 260 articles from each of the top three accounting journals TAR, JAR, and JAE (Journal of Accounting and Economics). Initially, we attempt to collect *p*-values from the text of these 780 articles using the same Python script as we used for experimental studies. However, we are able to scrape only 232 *p*-values from 76 articles, because unlike experimental studies, archival studies generally do not mention *p*-values within text. The sample is too small to conduct any meaningful analysis about the existence of *p*-hacking among those articles. We then identify 163 articles that report *p*-values in tables of regression analyses among the 260 articles. For a regression table, we record *p*-values appeared in the first three rows except those for intercepts. Arguably, *p*-values appeared in the upper part of tables are more likely to be the test statistics for key results. This method yields 4,767 *p*-values.

[Insert Figure 6 about here]

We find that the distribution of *p*-values is extremely uneven and right-skewed: there are 1,680 *p*-values reported as zero. Such a high frequency of zero would dwarf the heights of bins for other *p*-values, we therefore exclude these *p*-values. The histogram for the rest *p*-values is presented in Figure 8. Compared with Figure 1, the distribution of *p*-values from archival accounting do not show a discontinuity at 0.05. We can think of two explanations the difference. First, archival studies is less prone to *p*-hacking, probably due to the public availability of databases many archival researchers rely on, which makes replication easier and provides monitoring mechanism for *p*-hacking. Alternatively,

as archival researchers usually use much larger samples than experimentalists, they have much higher statistical power. It is possible that the archival researchers aim for $p \leq 0.01$ when doing *p*-hacking. The high frequency of *p*-values near zero is probably the result of this. In sum, though we do not observe a sharp discontinuity at 0.05 on the histogram of *p*-values from archival accounting articles, we still cannot rule out the possibility that those studies aim for the highest statistical significance level (*p*-values near zero) while doing *p*-hacking.

7. Potential remedies to guard against *p*-hacking

While we have provided evidence of *p*-hacking among experimental accounting studies, in this section, we further discuss potential remedies that may help to guard against *p*-hacking. As mentioned in the introduction, *p*-hacking can hinder academic progress because it increase the chance of false positive findings, which is largely driven by the problem of Multiple Hypotheses Testing (MHT) accompanied by insufficient disclosure on research agenda. We first introduce three widely used MHT adjustment methods developed to limit the number of false positive discoveries, with discussions on the implementation of the methods in experimental accounting research; we then illustrate the application of the methods using simulated experimental data. We also consider the Bayesian approach and provide some guidelines on improving research transparency.

7.1. Multiple hypotheses testing

Multiple Testing is common among many disciplines. It can be due to multiplicity in data, hypothesis, or analyses methods (Bender and Lange, 2001). For instance, data multiplicity can occurs when researchers choose among several samples to test the same set of hypotheses; hypothesis multiplicity arises when researchers investigate the effect of a treatment (or manipulation) on several outcome variables. Multiple Testing could be problematic because it increases Type I errors. This can be illustrated as follows. If a single hypothesis testing is conducted at a significance level α, the probability of not rejecting a true null hypothesis is by definition (1-α). Now consider conducting M independent tests, the probability of not rejecting all true null hypotheses is $(1 - \alpha)^{M}$, therefore, the

probability of rejecting at least one true null hypothesis become 1 - $(1-\alpha)^M$, which is also *called* *family-wise error rate* (*FWER*) or *experiment-wise error rate* (*EER*). Formally, it is defined as:

$$FWER=Prob(\cup_{m=1}^{M}\{\text{reject } H_{m0} \text{ when it is true}\}),$$

where *m* indexes the hypotheses tested, and H_{m0} is the null for hypothesis *m*. It is easy to see that *FWER* is an increasing function of *M*. When multiple tests are conducted, it therefore becomes necessary to control the FWER at a reasonably low level. The statistical literature has developed ways to control for FWER, we describe some of the most widely used methods and their implementations.

The first method we consider is Bonferroni method, which offers an intuitive procedure to control for FWER. Specifically, suppose we need a FWER lower than α, which is also the false positive rate for each individual test among the *M* tests. However, according to Bonferroni' inequality, we will have

$$FWER \leq \sum_{m=1}^{M} Prob(\text{reject } H_{i0} \text{ when it is true}) = \alpha \times M,$$

To guarantee a FWER lower than α, the Bonferroni method sets the false positive rate for each individual test as $\frac{\alpha}{M}$, then

$$FWER \leq \frac{\alpha}{M} \times M = \alpha,$$

consequently, under the Bonferroni method, an investigator compares the *p*-values of all tests to a single significance threshold α/M, and rejects a null hypothesis H_m if $p_m \leq \alpha$/M, where m indexes each test. By doing so, she controls for FWER at the level α.

The Bonferroni method has been used in experimental accounting studies since 1980s, but its application is quite limited.[22] However, in situations where the chances of false positives are high, Bonferroni method offers a simple adjustment that offers protection against conclusions that may be driven by chance. To be more specific, if the key hypothesis in an experimental study is examined by

[22] See Margheim (1986), Peecher (1996), and Libby and Brown (2013) for examples of the application of Bonferroni method.

means of multiple comparisons, then Bonferroni adjustment is necessary. For example, it is not uncommon for experimental studies to compare the mean of a control group with each of mean of four treatment groups, without specifying clearly the *ex ante* priorities among the comparison. In this specific case, the Bonferroni-adjusted *p*-values for significance at the 5% level should be 1.25% (5%÷4).

While widely known for its simplicity, Bonferroni method also lacks power when the number of tests are high, as for a large number of tests, the critical *p*-value will be extremely small. In addition, it treats each test as independent from each other. In practice, however, the test statistics from a family of tests are almost always correlated, and Bonferroni methods will over-correct the FWER. We introduce our second method to account for the interdependence of test statistics.

The second method we consider is bootstrapping adjusted *p*-values, which is a resample-based method designed to control for the FWER from multiple testing (Westfall and Young, 1993). This method is implemented as follows. Suppose there are *M* tests, each with an actual *p*-value of p_m, where $1 \leq m \leq M$. The first step is to bootstrap the sample data by sampling observations with replacement to create a pseudo-sample with same sample size as the original sample; second, for the *M* tests, calculate their respective *p*-values $\{p_1^b, p_2^b, \cdots p_M^b\}$ using the pseudo-sample, and get the minimum value p_{min}^b, where superscript *b* denotes that these *p*-values come from the bootstrap; third, for any test *m*, create a dummy variables, C_m^b, which equals to one if $p_{min}^b \leq p_m$ and zero otherwise (so we will have *M* counters for each bootstrapping). Lastly, repeat the above steps for *B* times, and the adjusted *p*-value for an actual *p*-value is calculated as $\frac{\sum_{b=1}^{B} C_m^b}{B}$, which is the ratio of times when the minimum *p*-value is lower than or equal to the actual *p*-value over the total times of bootstrapping. The adjusted *p*-values are compared against the 5% significance threshold. In our setting, we set *B* equals to 500.

As Westfall, Young, and Wright (1993) point out, the adjusted *p*-value measures how extreme a given *p*-value is relative to the distribution of the most extreme *p*-values from bootstrapping. Compared with the Bonferroni method, the biggest advantage of the bootstrapping adjusted *p*-values

is that it takes into account the dependence structure among individual test statistic, thereby improving the power of the multiple testing procedure. This method is thus particularly useful when the number of tests in an experiment is high, as in such cases the Bonferroni method will give extremely small critical *p*-values. For example, an investigator might examine the effect of two treatments on five outcome variables, generating a total of then hypotheses.[23]

The last method we consider to adjust for multiple comparison is controlling for False Discovery Ratio (FDR), which controls for the "proportion", rather than the "number", of false rejections. This is particularly important when we conduct a large number of tests and are thus more willing to tolerate a higher number of false rejections (Different from Bonferroni method that controls for the FWER, the probability of rejecting even one of the true null hypotheses). We adopt the method in Benjamini and Hochberg (1995) to control for the False Discovery Ratio δ at the 5% level. Specifically, we first rank a family of M *p*-values from minimum to maximum, and define: $j^* = \max\{j: p_j \leq \frac{j\times\delta}{M}\}$, j^*indexes the largest *p*-value that satisfies this inequality; Next, we reject all H_i, i = 1, 2, …, j^*. It is helpful to consider an example. If M = 5, and the five *p*-values are 0.01, 0.02, 0.03, 0.05, and 0.06. In this example, j^* = 3 ($\frac{3\times0.05}{5}$= 0.03). Therefore, despite the fourth hypothesis has a *p*-value of 0.05, it is still higher than 0.04 ($\frac{4\times0.05}{5}$), we thus only reject the first three hypotheses. Benjamini and Hochberg (1995) theoretically show that this procedure controls for FDR at δ for independent test statistics.

7.2. Simulations

To demonstrate the application of the multiple test procedures outlined above, we follow Simmons, Nelson, and Simonsohn (2011) to generate computer simulated experimental data, and show two sets of results: 1) how the rejection rates change as the number of tests increases; 2) the rejection rates after adjusting for multiplicity. Each "experiment" is a 2 × 2 design with 20 "participants" per

[23] Again, if it is possible to clearly specify the hypotheses tested ex ante and their relative priorities, then multiplicity problem is greatly mitigated.

cell, and an outcome variable drawn from standard normal distribution. Since we are interested in the magnitude of false positive rates, "participants" are randomly assigned to each of the 4 cells to construct a zero treatment effect. We label this design the baseline case.

We then add three discretionary choices on top of the baseline case to mimic the process of *p*-hacking, they are: choosing among different dependent variables, choosing the sample size, and using covariates.[24] A researcher can choose any combination of the three "degree of freedom" to search for the one that gives highest level of statistical significance. For instance, one can choose to use the first dependent variable, the second, or the average of the two, and whether to add another experiment session or not. There are 12 unique combinations of all the choices available; in each experiment, we thus run 12 Analyses of Variance (ANOVAs) to exhaust all possible combinations of the choices. To simplify the analysis, suppose we are only interested in finding statistically significant *interaction effect* of any pair of independent variables. We then perform 1,000 simulated experiments, and count ─ for each combination ─ how often the minimum *p*-value among all the interaction effects is below 0.05. The purpose of the simulation is to show that, if one looks into a given data set long enough and hard enough, it will be much easier to find statistical significant results by chance (i.e., spurious findings).

Column (1) of Table 5 reports each situation and its corresponding false positive rates. In the baseline case, the false positive rates is 3.8%, which is a bit lower than 5%. In situation 2, a researcher has the flexibility to analyze one of two dependent variable (correlation coefficient 0.5), or their mean, and choose whichever gives the highest statistical significance. This simple flexibility doubles the chance of obtaining a false positive finding. As we move to Situation 7 in the last row, where the researcher combines all choices, the rejection rate increases to 34.3%. This implies the researcher can

[24] We choose these three researchers' "degree of freedom" because these are the commonly made decisions while a researcher is collecting and analyzing data, and also because they are easily implementable. We can add more choices, but the fundamental takeaway from the exercise will be the same.

find at least one statistical significance in every three simulations. Because we only consider three choices, these estimates may still be conservative.

Column (2) to (4) of Table 5 provide the rejection rates under the multiple testing procedures introduced. The Bonferroni method is the simplest but also the most conservative one, as evidenced by the low rejection rates compared with other methods. Taking Situation 1 as an example, under the traditional single hypothesis testing, there are 76 times among the 1000 simulations with at least one significant interaction effects; under the Bonferroni method, the number drops to 27. Turning to Column (3), the rejection rates across all situations under the Bootstrapping adjusted *p*-values are in general higher than the rejection rates under Bonferroni method, especially for situations with higher number of tests, suggesting that taking into account the dependence among test statistics indeed gives less conservative critical *p*-values. The rejection rates under Bootstrap adjusted *p*-values are also close to the 5% significance level. Column (4) reports the result of BH method. By tolerating higher number of false rejections when the total number of tests is high, BH also method yields a higher rejection rate than the Bonferroni method.

To summarize, we find that flexibility in data collection and analyses allow researchers to easily report statistical significant findings, and after accounting for multiple testing, significant findings are much less frequent, and the rejection rates move towards 5%. Given that experimental accounting studies in general rely on small samples with low statistical power (higher *p*-values), we recommend the use of Bootstrap adjusted *p*-value and the BH method, which are more powerful than Bonferroni method. However, one important limitation in implementing multiple testing procedures is that, in reality, we can only observe those tests that are disclosed by researchers, and the reported tests may only be a subgroup of all performed tests. Hence, the use of multiple tests procedure cannot fully protect against the bias caused by *p*-hacking.

[Insert Table 5 about here]

7.3. The Bayesian approach

In our simulation, the treatment effect is designed to be zero and has no theoretical underpins. Consequently, the prior probability of running into a true discovery should also be low, i.e., the prior odds ratio between the null hypothesis and the alternative hypotheses will be high. However, the computation of *p*-value does not incorporate this prior information. Recall that the *p*-value indicates $p(D|H_0)$, which the probability of observing the result or more extreme results given that the null hypothesis being true, not $p(H_0|D)$, which is the probability of null being true given the data. If, however, we can incorporate this prior belief into our hypotheses testing, then we can potentially have a method to help guard against extreme findings that result from *p*-hacking. This is the basic logic of Bayesian approach.

Essentially, the Bayesian approach assess the viability of hypothesis based on a posterior odds ratio between the null hypothesis and the alternative hypotheses, given data likelihood ratio and prior odds ratio. According to the Bayesian theorem, the posterior odds ratio is defined as:

$$\underbrace{\frac{f(\theta_0|data)}{f(alternative|data)}}_{\text{Posterior odds ratio}} = \underbrace{\frac{f(data|\theta_0)}{\int (data|\theta)\pi_A(\theta)d\theta}}_{\text{Bayesian factor}} \times \underbrace{\frac{\pi_0}{\pi_1}}_{\text{prior odds ratio}},$$

where $f(x|y)$ is the probability density function of x conditional on y, $\pi_A(\theta)$ is the probability density under the alternative hypotheses, and $\int (data|\theta)\pi_A(\theta)d\theta$ is thus the data likelihood under the alternative hypotheses. The odds ratio between the data likelihood under the null and the alternative, $\frac{f(data|\theta_0)}{\int (data|\theta)\pi_A(\theta)d\theta}$, is also called the Bayesian factor. After multiplying the prior odds ratio by the Bayesian factor, we have the posterior odds ratio on the left-hand-side. Economically, the Bayesian factor measures the extent to which our prior belief are changed after observing the data, with smaller Bayesian factor provides stronger evidence *against* the null.

As can be seen from the above equation, calculating the Bayesian factor requires knowledge on the probability density of the alternative hypotheses, i.e., $\pi_A(\theta)$. To calculate $\pi_A(\theta)$, we need to specify priors on all possible alternative hypotheses, and for each of these hypothesis, we calculate the

data likelihood under the null. However, in financial and accounting studies, we usually do not have well-specified alternative hypotheses, this makes it is infeasible to calculate the Bayesian factor. Fortunately, there is one type of Bayesian factor that does not rely on the prior specification of alternatives: the minimum Bayesian factor (MBF). Harvey (2017) provides an excellent description on MBF, which we draw upon here.

Specifically, MBF is calculated for the case when a researcher focuses on a specific alternative hypothesis that turns out to be the maximum likelihood estimate of the data, i.e., the hypothesis that is most supported by the data. In other words, the MBF is calculated under the situation that what we believe (the alternative hypotheses) coincides with what we see (the data), it therefore provides the strongest evidence against the null hypothesis. In this regard, the MBF is particularly suitable for experimental accounting studies, which usually develop directional hypotheses, implying that prior density of the alternative hypotheses is likely to be concentrated in one side of the null hypothesis; therefore, in using MBF, we give an alternative hypothesis the best shot by concentrating the prior density at the alternative rather than the null. Moreover, MBF is easy to calculate under the standard normal distribution assumption:

$$\text{MBF}=e^{-\frac{z^2}{2}},$$

where e is the natural exponent, and z is the corresponding z-value from the null effect. For example, if p-value = 0.05 (two-tailed), then $z = \pm 1.96$, and MBF = 0.15, or $\frac{1}{6.8}$. As a result, the observed value supports the alternative hypothesis 6.8 times as strongly as it does the null. Once MBF is obtained, we can calculate the Bayesianized p-value as follows:

$$\text{Bayesianized } p\text{-value} = \frac{\text{posterior odds ratio}}{1+\text{posterior odds ratio}} = \frac{\text{MBF*prior odds ratio}}{1+\text{MBF*prior odds ratio}}.$$

Continuing the above numerical example, if we believe that *ex ante*, the null hypothesis and the alternative is equally likely to be true (i.e., prior odds ratio = 1), then the Bayesianized p-value is 0.13, which indicates the chance of the null being true is 13%. If instead, the prior odds ratio = 0.50

(i.e., there is modest chance in favor of the alternative), then the Bayesianized *p*-value becomes only 0.07. In both cases, the Bayesianized *p*-values are higher than the conventional *p*-value 0.05. However, the Bayesianized *p*-values carries a fundamentally different interpretation: it is a statement about the *viability* of the null hypothesis, as opposed to the likelihood of observing the data assuming null being true.

How can the Bayesian approach be helpful in experimental accounting studies? Bloomfield, Rennekamp, and Steenhoven (2018) mention that Bayesian approach can be used as a complement to the Null Hypothesis Significance Testing (NHST). We follow their line of reasoning and offer some recommendations in here. Specifically, experimental researchers can take their priors into account, and present Bayesianized *p*-values alongside the usual test statistics. Given the relative high cost (e.g., recruiting and rewarding participants) of conducting experimental studies and small sample size, experimentalists generally do not commence data collection unless they believe that their underlying theories are highly plausible. This suggests that experimental accounting researchers generally have a low prior odds ratio (i.e., the alternative is more likely to be true than the null). The Bayesianized approach can be particularly valuable because it offers a convenient way for researchers to incorporate their prior beliefs, which is unfeasible in NHST. In addition, the choice of prior odds ratio is transparent and less subject to manipulations.[25] Researchers can also choose to report Bayesianized *p*-values under multiple prior odds ratios, and thus readers with different priors can make their own assessments about the viability of the hypotheses given the test results.

7.4. Disclosure-based solution

Since *p*-hacking relies on exploiting *undisclosed* flexibility in data collection and analyses, another natural solution is to require researchers to make reasonably transparent disclosures on their research agendas and methodologies. Such disclosures are likely to curb researchers' ex ante incentives

[25] One might think the choice of prior odds ratio is subjective. However, writing his presidential address, Harvey (2017) emphasize the point from Fisher (1925): "There is no 'objective' evaluation of hypotheses". Researchers from various disciplines relies too much on *p*-values to assess the viability of hypotheses, while ignoring the fact that *p*-value was not designed for this type of interpretation.

to conduct *p*-hacking, because once a carefully planned research agenda is disclosed, researchers have limited room for cherry-picking afterwards. Consequently, disclosure-based solution has been widely discussed and adopted. For example, top journals in many disciplines usually require researchers to disclose their data or code.[26]

Of course, greater disclosure and transparency are not without costs. For instance, Bloomfield, Rennekamp, and Steenhoven (2018) point out that while JAR's a Registration-based Editorial Process can potentially enhance reproducibility by mitigating questionable practices, it may also shut down the channel for authors to learn from their data and refine research agendas.[27] In what follows, we highlight three type of disclosures that journals can require from authors to mitigate *p*-hacking while imposing minimum costs on authors and reviewers. This list of disclosures is not intended to be exhaustive; rather, it is based on some of the questionable practices that are frequently cited as enabler of *p*-hacking.

1) *Specifying sample collection stopping rule*. Simmons et al. (2011), and Button et al. (2013) cite arbitrary stopping rule as a common practice among researchers to inflate results. As such, if journals aim to improve transparency in author reporting and documentation, a critical dimension of transparency is sample selection procedure. We thus recommend that journals' disclosure policy can include guidelines that encourage authors to decide their sample selection stopping rule *ex ante*. For example, authors can state that "we decide to recruit 120 MBA participants for our experiment", or "we attempt to conduct a pilot study with 60 participants, and include the pilot sample as part of the experiment proper". The stopping rules can be study-specific, but once chosen, journals and reviewers should make sure that authors stick to the rules.

[26] For a partial list of top journals in social science that adopted similar policies, please see: https://web.stanford.edu/~cy10/public/data/Data_Availability_Policies.pdf.

[27] Journal of Accounting Research (JAR) experimented a Registration-based Editorial Process (REP) in 2017. Under REP, authors first submit their proposals containing data collection and analyses to JAR. After going through the process of peer review and revisions, editors and reviewers grant in-principal publication to proposals that meet publication standard, therefore guaranteeing publication regardless of the final results obtained under the agenda. The purpose of REP, according to Professor Leuz, is to enhance reproducibility of findings and encourage academic risk-taking (Wiley, 2018).

2) *Report all the variables collected*. A transparent disclosure of outcome variables and conditions constructed in an experiment allows readers to assess the potential number of tests conducted by authors, and interpret their *p*-values in context.[28] Consequently, journals' disclosure policy should put great emphasize on requiring authors to list all variables collected in a study. Moreover, the survey conducted by Bloomfield, Rennekamp, and Steenhoven (2018) also show that researchers believe that "choosing analyses to highlight and report" (including choosing different variables) is most likely to "overstate results without improve informativeness". This evidence suggests that journals' effort to address this issue are likely to meet researchers' cooperation.

3) *Reporting p-values*. Authors should keep consistency in the choice between reporting one-tailed or two-tailed *p*-values. The articles in our sample can generally choose to report one-tailed *p*-values if the corresponding hypotheses are directional. However, it is not uncommon to find articles that report one-tailed *p*-values in the main tests, but two-tailed *p*-values in follow-up tests. Such discretional use of tail choice is indicative of cherry-picking, especially in cases where the *p*-values would be higher than the significance threshold if two-tailed *p*-values are used.

In addition, journals should ensure that the underlying statistics of *p*-values (*F*- or *t*-stat, and degree of freedom) be available, either in tables or in the text where a *p*-value is mentioned. This is the reporting policy of *The Accounting Review*. However, in our sample, we still find articles that do not report degree of freedom in neither tables nor text, especially when the results are not reported in the form of ANOVA tables. This is potentially problematic, as authors may have incentive to misstate *p*-values when readers cannot easily verify the accuracy of *p*-values.[29]

To summarize, we discussed three type of researcher' discretions that could be used to overstate the importance of results, and suggest journals that are aiming to improve transparency emphasize

[28] For example, if authors use three measures to capture the same outcome in an experiment and only report the one that generates strongest results, readers will have no way to assess the multiplicity of the test because the true number of tests is unknown.

[29] When we attempt to calculate the exact values for *p*-values reported as "p <", we find several cases that authors overstate their *p*-values. For example, when authors report "p < 0.10", the true value is p = 0.12 or 0.14. Had the articles not provide the underlying statistics such as *t*-statistics or *F*-statistics, we would not be able to discover these overstatements.

authors' disclosure from these dimensions, which we believe can greatly enhance transparency without imposing much costs on both authors and reviewers. Notably, by no means are the three type of discretions a complete list of researchers' "degree of freedom". We hope the discussion here can further stimulate the conversation on improving transparency and reproducibility in experimental accounting research.

8. Conclusion

We uncover a discontinuity at 0.05 in the pooled cross-sectional distribution of *p*-values from experimental accounting studies, and we interpret the discontinuity as a result of researchers' practice of *p*-hacking to generate significant results. The extent of *p*-hacking is more pronounced when authors face pressure of tenure, among male researchers, and when a researcher works alone. Furthermore, researchers graduated from higher-ranking schools seem to be more aggressive in *p*-hacking. Papers that exhibit greater extent of *p*-hacking also attract more citations after publication.

The results in this paper may underestimate the true extent of *p*-hacking due to the following reasons. First, our sample could be biased because we use *p*-values mentioned in the text of articles, and the decision to mention a *p*-value in text is made by researchers. Arguably, researchers would like to discuss the most significant results in the paper, and are reluctant to emphasize a *p*-value that is just above the 0.05 threshold. Second, we only use articles published from three of the most prestigious accounting journals, it is reasonable to suspect that articles published on lower-tier journals are also subject to similar, or even higher extent of *p*-hacking than what we documented here. Lastly, we focus only on discontinuity at 0.05, it is possible that some researchers aim for higher *p*-value targets such as 0.04, 0.03 or even 0.01 to avoid being too close to the 0.05 red line. However, such practice is much harder to detect because aiming for higher significance level will cause a more right-skewed *p*-value distribution, which can also be explained by publication bias.

We also propose several methods to guard against *p*-hacking. Specifically, we introduce several multiple testing procedure to control for false positive findings due to conscious or unconscious *p*-

hacking. Moreover, we propose a Bayesian approach for researchers to account for their prior beliefs on the likelihood of their theories. Lastly, we discuss some guidelines on enhancing disclosure with minimum burden on both authors and reviewers. Given the increasing attention on the prevalence and potential consequence of *p*-hacking in experimental accounting research, we believe it is the right time to discuss potential remedies, and with the cooperation among journals, reviewers, and authors, experimental accounting research will continue to produce high-quality works that boost scientific progress in accounting.

References

Bloomfield, R., Rennekamp, K., Steenhoven, B., 2018. No System Is Perfect: Understanding How Registration-Based Editorial Processes Affect Reproducibility and Investment in Research Quality. Journal of Accounting Research 56, 313-362

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

Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience 14, 365-376

Byrnes, J.P., Miller, D.C., Schafer, W.D., 1999. Gender differences in risk taking: A meta-analysis. Psychological Bulletin 125, 367-383

Charness, G., Gneezy, U., 2012. Strong evidence for gender differences in risk taking. Journal of Economic Behavior & Organization 83, 50-58

Dunnett, C.W., 1955. A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association 50, 1096-1121

Dwan, K., Altman, D.G., Arnaiz, J.A., Bloom, J., Chan, A.-W., Cronin, E., Decullier, E., Easterbrook, P.J., Von Elm, E., Gamble, C., 2008. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PloS ONE 3

Faccio, M., Marchica, M.-T., Mura, R., 2016. CEO gender, corporate risk-taking, and the efficiency of capital allocation. Journal of Corporate Finance 39, 193-209

Fanelli, D., 2010. "Positive" results increase down the hierarchy of the sciences. PloS ONE 5, e10068

Fanelli, D., 2012. Negative results are disappearing from most disciplines and countries. Scientometrics 90, 891-904

Feise, R.J., 2002. Do multiple outcome measures require p-value adjustment? BMC medical research methodology 2

Fisher, R.A., 1925. Theory of statistical estimation. In: Mathematical Proceedings of the Cambridge Philosophical Society, pp. 700-725. Cambridge University Press

Greenwald, A.G., 1975. Consequences of prejudice against the null hypothesis. Psychological bulletin 82, 1-20

Harvey, C.R., 2017. Presidential address: the scientific outlook in financial economics. Journal of Finance 72, 1399-1440

Harvey, C.R., Liu, Y., Zhu, H., 2016. … and the cross-section of expected returns. The Review of Financial Studies 29, 5-68

Head, M.L., Holman, L., Lanfear, R., Kahn, A.T., Jennions, M.D., 2015. The extent and consequences of p-hacking in science. PLoS biology 13

Henry, E., 2009. Strategic disclosure of research results: The cost of proving your honesty. The Economic Journal 119, 1036-1064

Khan, M.J., Tronnes, P., Christen, 2018. P-hacking in Experimental Audit Research. Behavioral Research in Accounting

Krawczyk, M., 2015. The search for significance: a few peculiarities in the distribution of P values in experimental psychology literature. PloS one 10

Libby, R., Bloomfield, R., Nelson, M.W., 2002. Experimental research in financial accounting. Accounting, Organizations and Society 27, 775-810

Libby, R., Brown, T., 2013. Financial statement disaggregation decisions and auditors' tolerance for misstatement. The Accounting Review 88, 641-665

Margheim, L.L., 1986. Further evidence on external auditors' reliance on internal auditors. Journal of Accounting Research, 194-205

Masicampo, E., Lalande, D.R., 2012. A peculiar prevalence of p values just below. 05. The Quarterly Journal of Experimental Psychology 65, 2271-2279

Neyman, J., Pearson, E.S., 1933. IX. On the problem of the most efficient tests of statistical hypotheses. Phil. Trans. R. Soc. Lond. A 231, 289-337

O'Riordan, M., 2018. The Road to Reproducibility: Registered Reports and the Journal of Accounting Research. URL https://hub.wiley.com/community/exchanges/discover/blog/2018/05/07/the-road-to-reproducibility-registered-reports-and-the-journal-of-accounting-research

Pashler, H., Harris, C.R., 2012. Is the replicability crisis overblown? Three arguments examined. Perspectives on Psychological Science 7, 531-536

Peecher, M.E., 1996. The influence of auditors' justification processes on their decisions: A cognitive model and experimental evidence. Journal of Accounting Research, 125-140

Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological science 22, 1359-1366

Simonsohn, U., Nelson, L.D., Simmons, J.P., 2014. P-curve: a key to the file-drawer. Journal of Experimental Psychology: General 143, 534-547

Statzner, B., Resh, V.H., 2010. Negative changes in the scientific publication process in ecology: potential causes and consequences. Freshwater Biology 55, 2639-2653

Sterling, T.D., 1959. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. Journal of the American statistical association 54, 30-34

Vermeulen, I., Beukeboom, C.J., Batenburg, A., Avramiea, A., Stoyanov, D., van de Velde, B., Oegema, D., 2015. Blinded by the light: How a focus on statistical "significance" may cause p-value misreporting and an excess of p-values just below. 05 in communication science. Communication Methods and Measures 9, 253-279

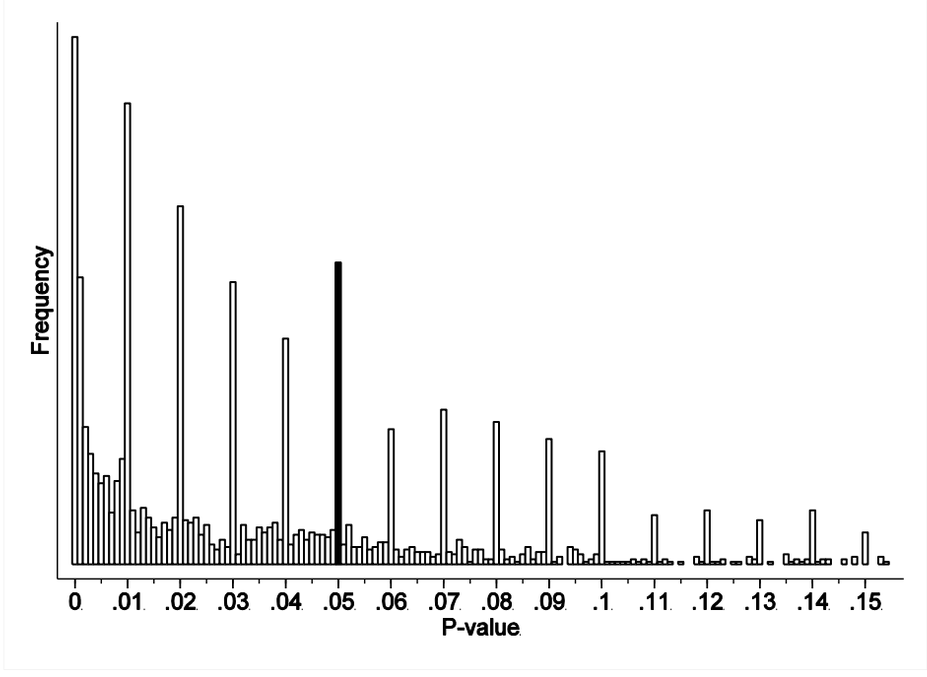
Westfall, P.H., Young, S.S., 1993. Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment (Wiley Series in Probability and Statistics).

Westfall, P.H., Young, S.S., Wright, S.P., 1993. On adjusting P-values for multiplicity. Biometrics 49, 941-945

Wicherts, J.M., Veldkamp, C.L., Augusteijn, H.E., Bakker, M., Van Aert, R., Van Assen, M.A., 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. Frontiers in Psychology 7

Young, N.S., Ioannidis, J.P., Al-Ubaydli, O., 2008. Why current publication practices may distort science. PLoS medicine 5

Panel A: Raw distribution of all *p*-values



Panel B: Raw distribution weighted by article's number of *p*-values

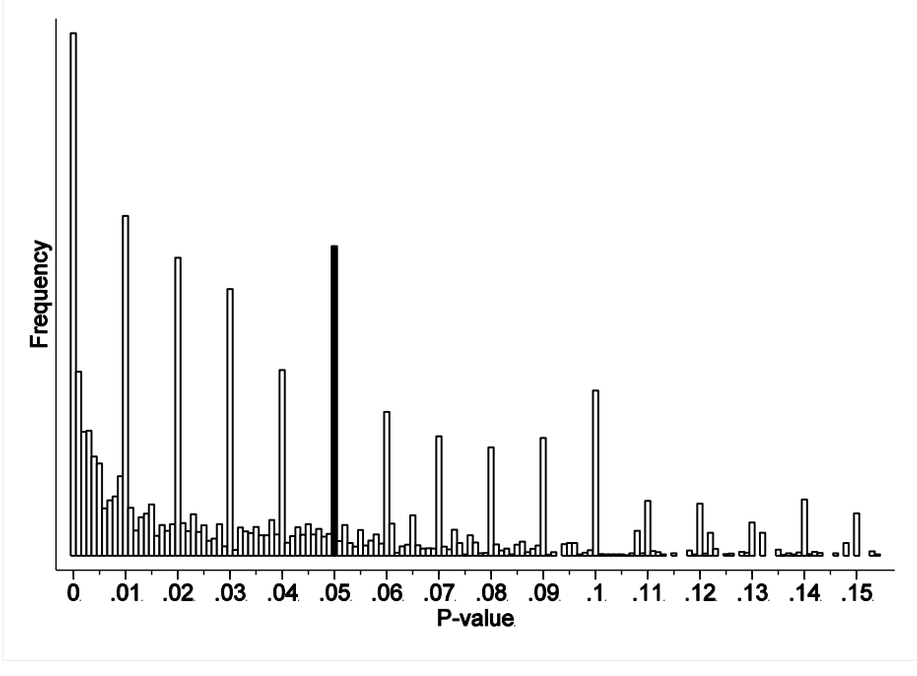
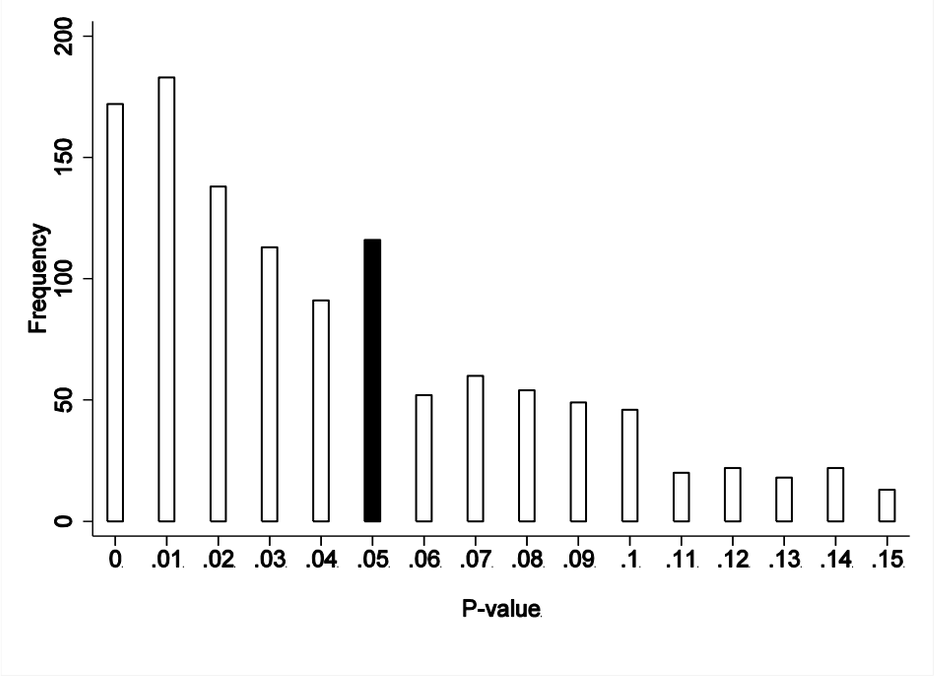


Figure 1. Empirical distribution of all *p*-values from 1990 to 2016.
Bin width is 0.001. The bold bar indicates frequency of *p*-values ϵ (0.049, 0.05]. In Panel B, the distribution uses the inverse of the number of *p*-values collected to weight observations.

Panel A: Only including *p*-values reported with two decimal places



Panel B: Rounding all *p*-values to two decimal places

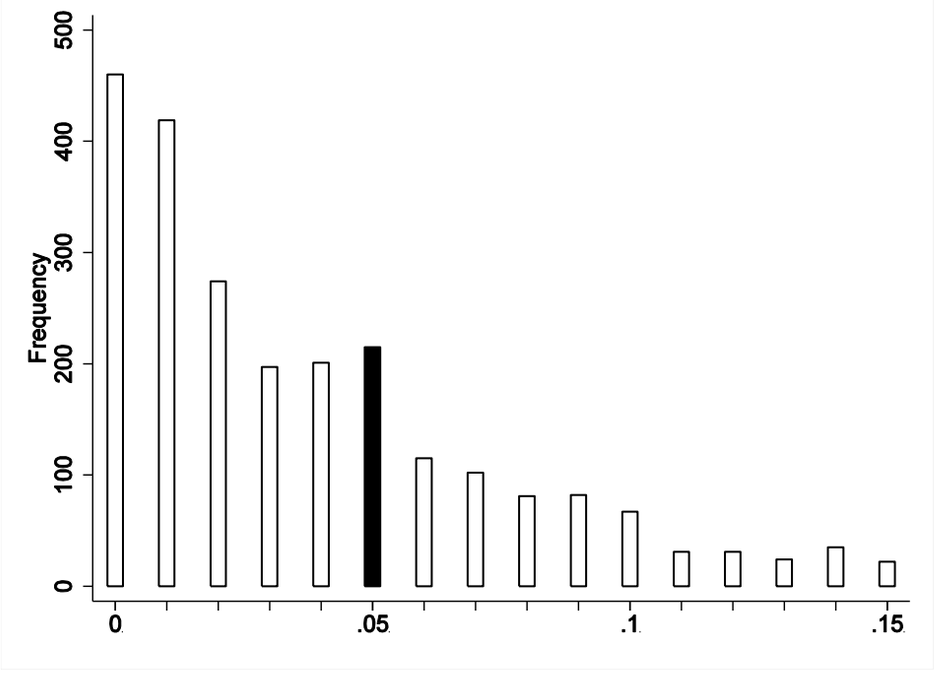
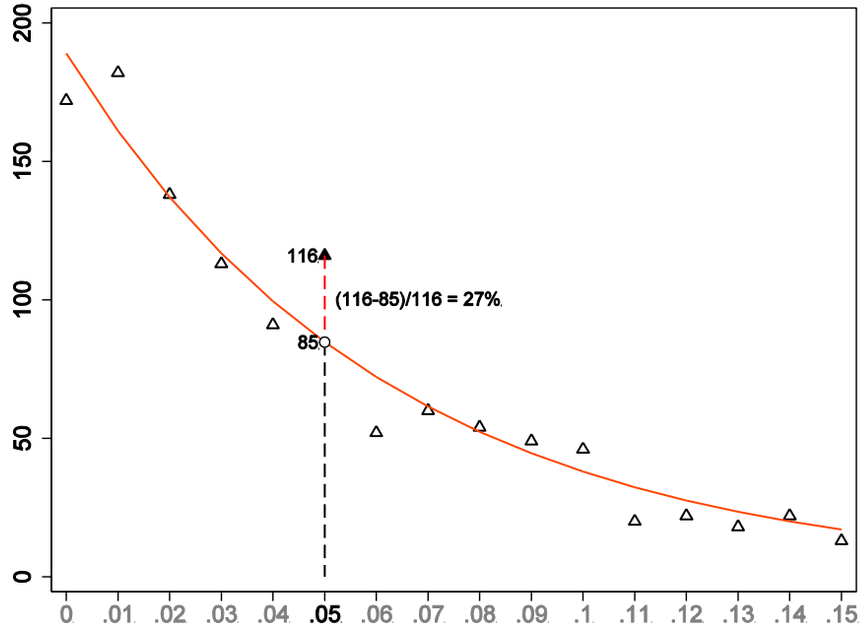


Figure 2. Isolating *p*-values reported with two decimal places.
In Panel A, we use only *p*-values reported with two decimal places; in Panel B we round all p-values and use the rounded values to generate the histogram.

Panel A: Only including *p*-values reported with two decimal places



Panel B: Rounding all *p*-values to two decimal places

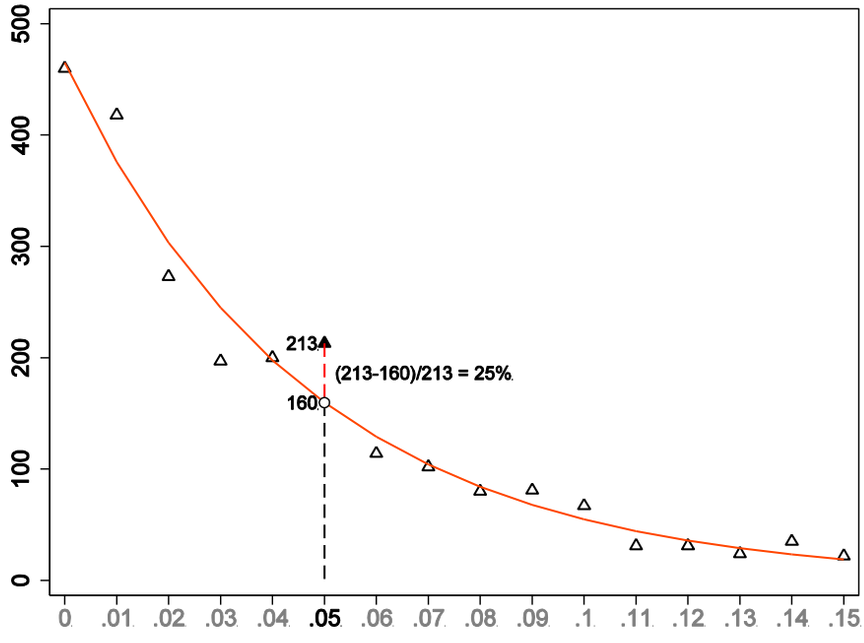
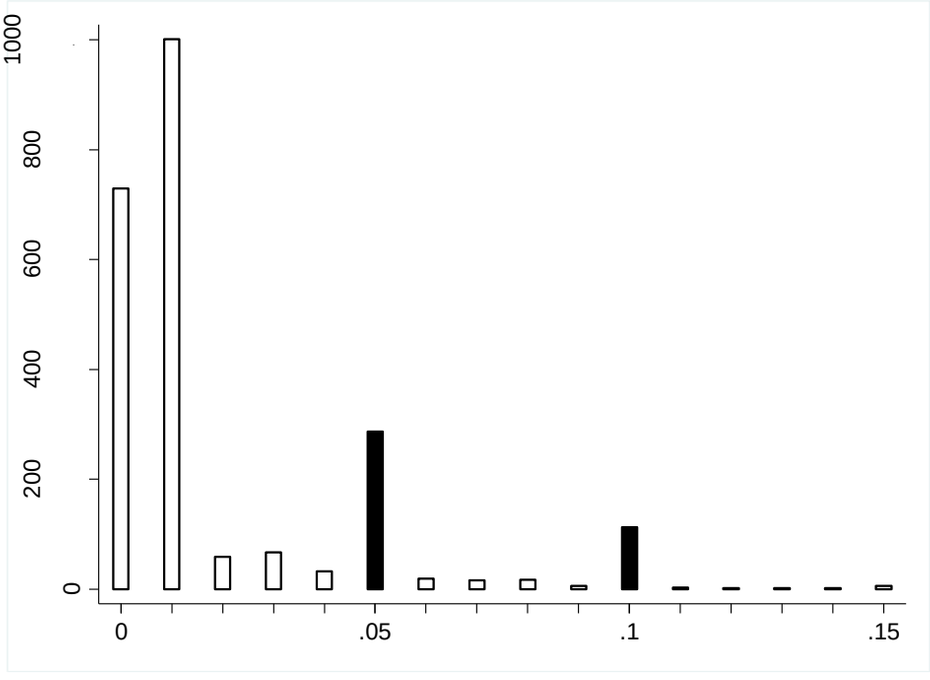


Figure 3. Fitting histogram of *p*-values to exponential distribution
Red line represents the fitted exponential curve that minimize the squared fitting errors; grey triangular represents the frequency of each observed *p*-value on x-axis. The exponential function in Panel A and B are $y = 205.49e^{-17.65x}$ and $y = 458.464e^{-21.96x}$, respectively.

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



Panel B: The distribution of 104 exact *p*-values calculated from p < 0.05 or p < 0.10

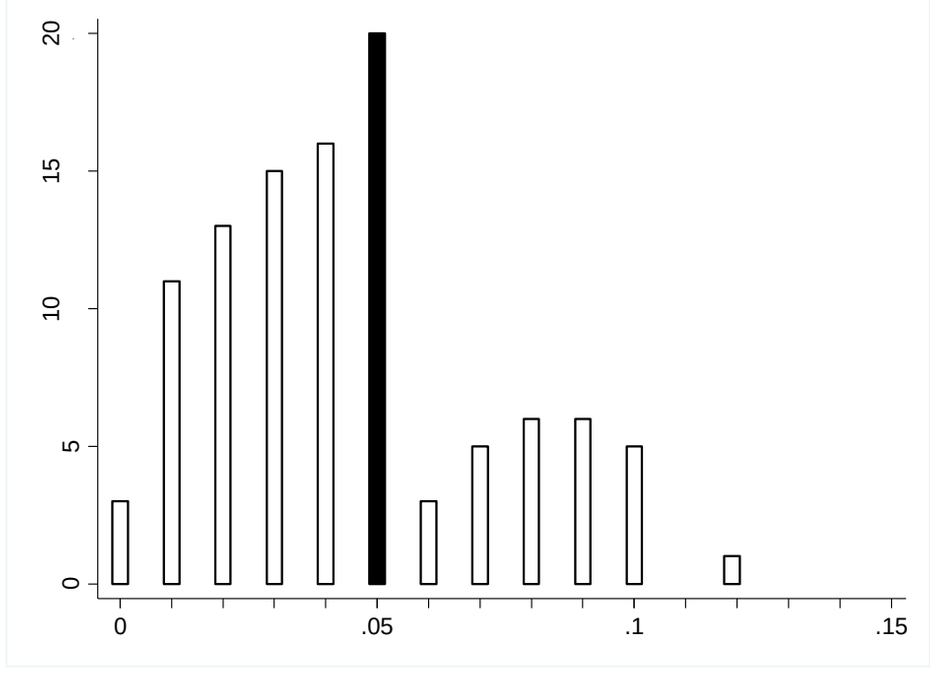
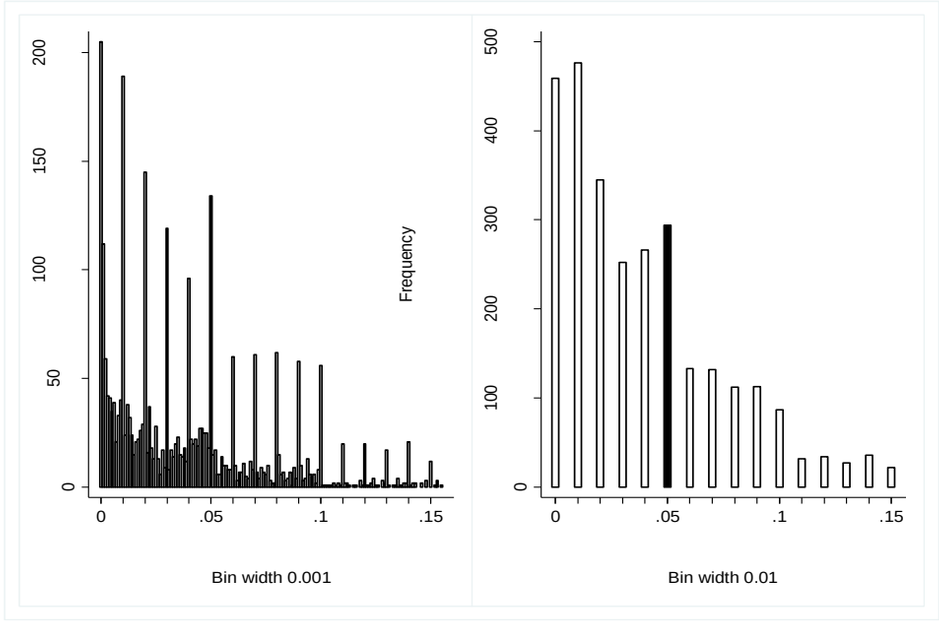


Figure 4. The distribution of vague p-values

The histogram in Panel A shows the distribution of 2,362 *p*-values reported as "p <"; the histogram in Panel B shows the distribution of 104 exact *p*-values we are able to compute for the 396 *p*-values reported as p < 0.05 or p < 0.10, which are indicated with the bold bars.

Panel A: Adding the estimated distribution of $p < 0.05$ or $p < 0.10$ to main sample



Panel B: Adding the estimated distribution of all vague *p*-value to main sample

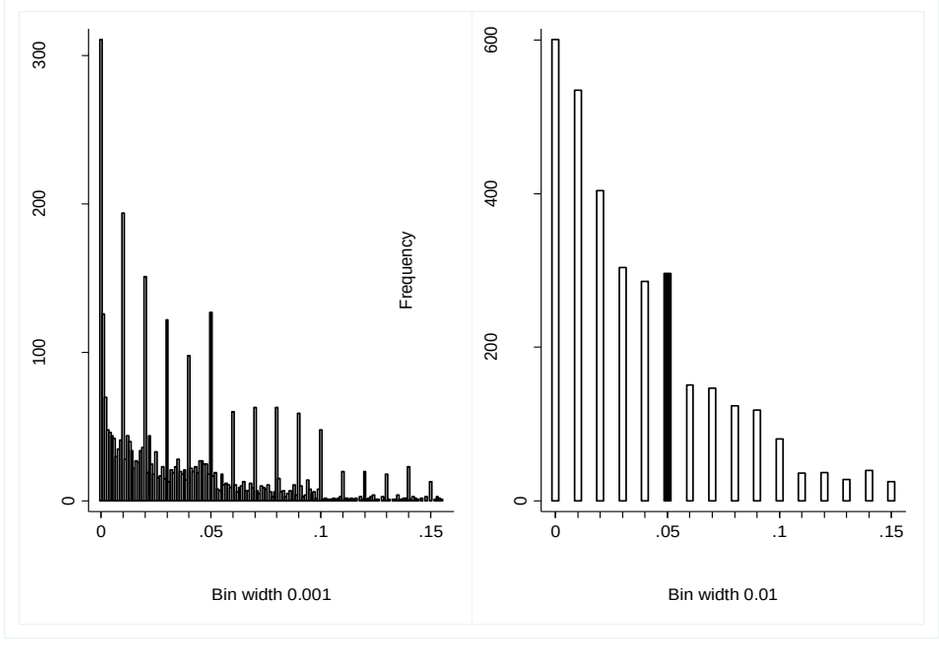
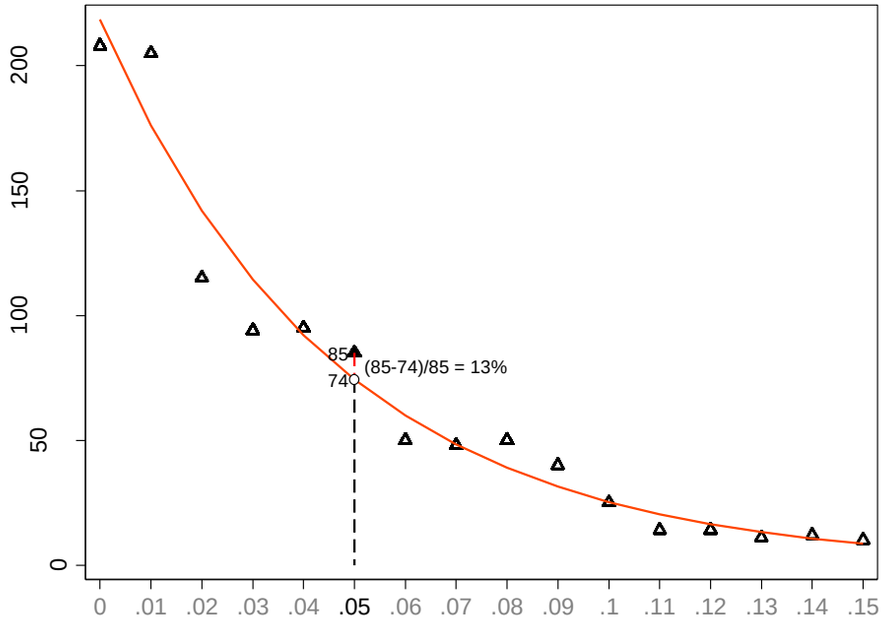


Figure 5. Including vague p-values

In Panel A, we combine the empirically estimated distribution of 396 vague *p*-values ($p < 0.05$, 0.10) with the distribution of exact *p*-values; in Panel B we further include other vague *p*-values by assuming a uniform distribution for $p < x$ between (x-0.01) and x, where x denotes 0.02,0.03, …, and 0.15 (except 0.05 and 0.10).

Panel A: Articles in *low citation* group



Panel B: Articles in *high citation* group

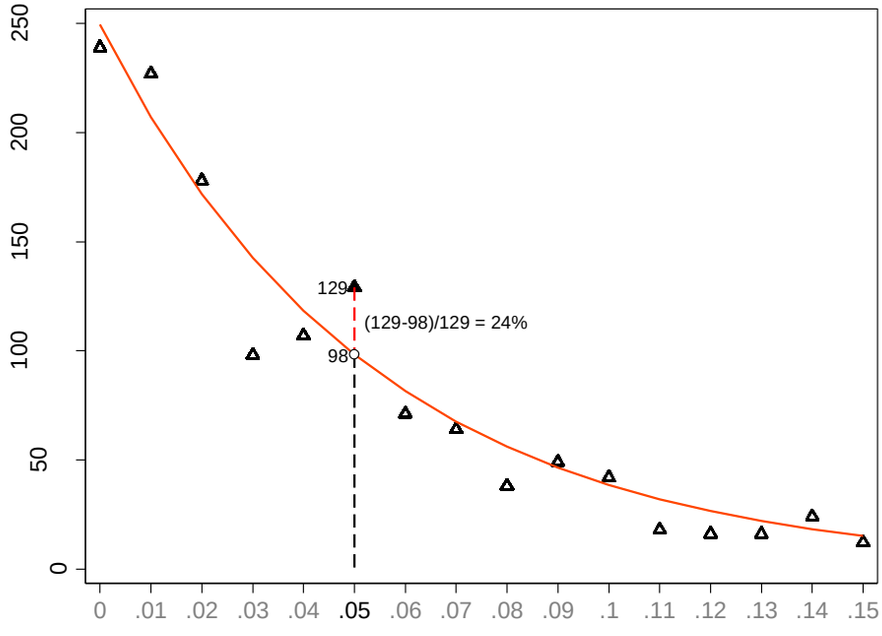
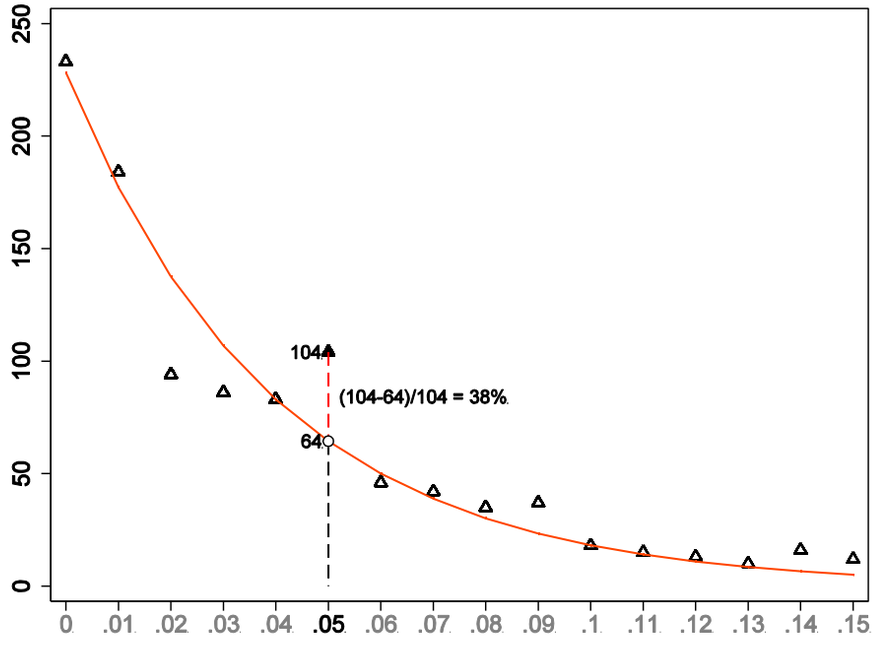


Figure 6. *P*-hacking and citations
In each year, we split articles into high and low citation groups based on the median number of citations of all articles published in that year. The final low (high) citation group consists of low (high) citation groups from every year.

Panel A: *P*-values appeared in the earlier part of an article



Panel B: *P*-values appeared in the later part of an article

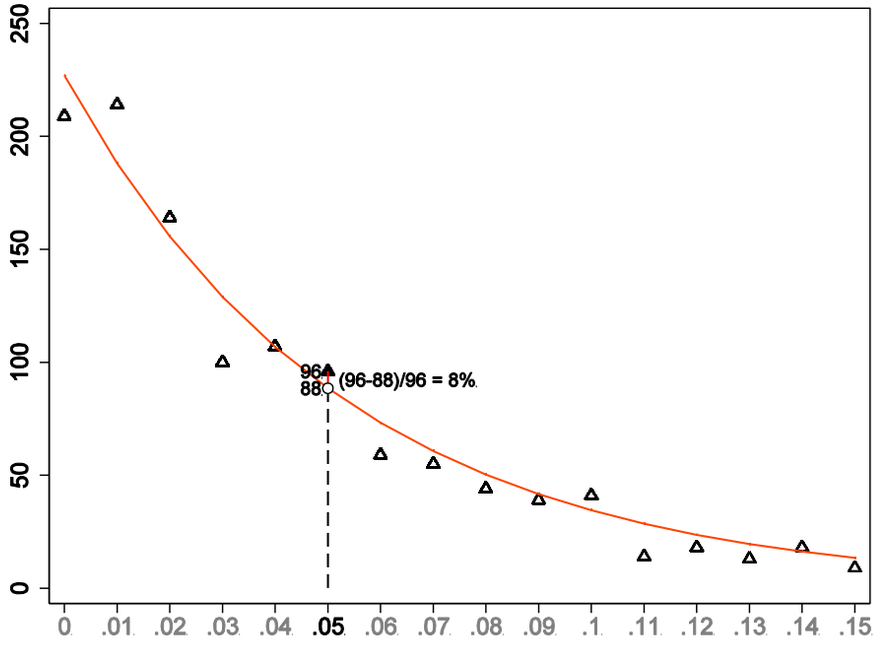


Figure 7. Do main results of an article show higher degree of *p*-hacking?
For each article, we sort all of its *p*-values by the sequences they appear in that article, then we identify the median value of the sequential number. Panel A contains those *p*-values that appear before each articles' respective median sequential number, Panel B contains the rest.

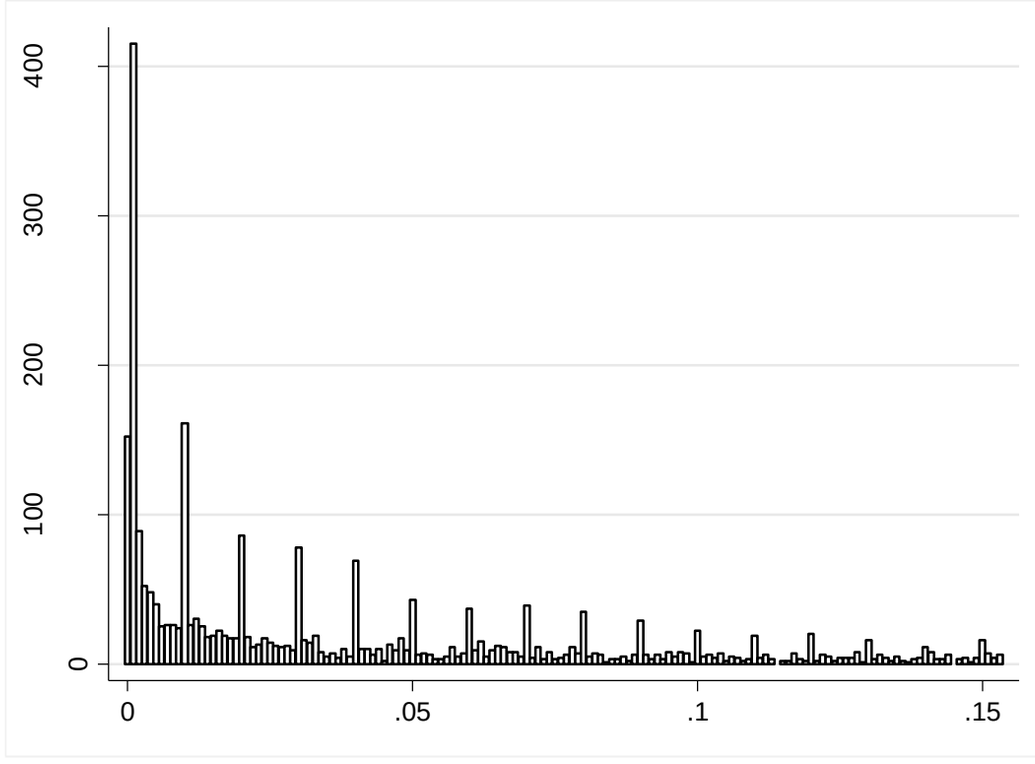


Figure 8. Histogram of *p*-values from archival accounting studies.
The figure presents the histogram of 3,087 *p*-values manually collected from 163 archival accounting studies. The bin width is 0.001.

Table 1. Summary statistics at the article-level

This table displays summary statistics for author and paper characteristics based on a sample of 240 experimental articles published in Top 3 accounting journals between 1990 and 2016. *N_Authors* is the number of listed author(s) for a paper. *Female* is a dummy variable that equals to one if the listed author(s) contain at least one female, and zero otherwise. *N_Participants* is the number of subjects participated in a paper's experiments. *N_p-values* is the total number of p-values scraped from a paper. *Two_tailed* is a dummy that equals to one if an article only report two-tailed *p*-values. *Sole* is a dummy variable that equals to one if a paper contain only one author, and zero otherwise. *Tenured* is the fraction of authors that are tenured among all authors of a paper. *School Ranking* is the mean ranking of all authors' affiliated schools for a paper. *PhD School Ranking* is the mean ranking of authors' PhD alma mater for a paper. *Experience* is the average number of years passed after an articles' authors obtained their PhD degree, calculated at the publication year of that article. The subsample for JAE are not tabulated separately because there are only three articles from JAE in our sample.

	N	Mean	SD	p25	p50	p75
Full sample						
N_Authors	240	2.171	0.937	1.000	2.000	3.000
Female	240	0.500	0.501	0.000	0.500	1.000
N_Participants	233	116.940	87.926	65.000	97.000	138.000
N_p-values	240	17.992	10.589	10.000	16.000	23.000
Two-tailed	239	0.343	0.476	0.000	0.000	1.000
Sole	240	0.287	0.454	0.000	0.000	1.000
Tenured	235	0.420	0.373	0.000	0.500	0.667
School Ranking	240	45.027	30.504	19.000	40.000	67.167
PhD School Ranking	229	29.361	21.739	15.000	22.500	39.000
Experience	230	9.173	6.000	4.000	8.000	12.500
JAR						
N_Authors	77	2.156	0.875	2.000	2.000	3.000
Female	77	0.455	0.501	0.000	0.000	1.000
N_Participants	74	119.486	57.891	78.000	106.000	145.000
N_p-values	77	18.364	11.397	10.000	17.000	24.000
Two-tailed	77	0.416	0.496	0.000	0.000	1.000
Sole	77	0.247	0.434	0.000	0.000	0.000
Tenured	75	0.441	0.356	0.000	0.500	0.667
School Ranking	77	38.315	27.507	17.500	33.000	49.000
PhD School Ranking	75	26.218	20.089	13.000	20.000	34.000
Experience	75	9.856	6.459	4.000	9.000	14.000
TAR						
N_Authors	160	2.188	0.972	1.000	2.000	3.000
Female	160	0.525	0.501	0.000	1.000	1.000
N_Participants	156	117.205	99.251	63.000	91.000	130.500
N_p-values	160	17.962	10.183	11.000	16.000	23.000
Two-tailed	159	0.308	0.463	0.000	0.000	1.000
Sole	160	0.306	0.462	0.000	0.000	1.000
Tenured	157	0.411	0.384	0.000	0.500	0.667
School Ranking	160	48.433	31.599	20.250	48.250	70.250
PhD School Ranking	151	31.016	22.552	15.500	25.000	43.000
Experience	152	8.850	5.753	4.000	7.583	12.333

Table 2. Deviations from the exponential distribution

This Table reports the deviations of the frequency of each p-values with two decimal places (0.01, 0.02... 0.15) from its forecasted frequency determined by the exponential fitting. In Panel A, we only use p-values reported with two decimal places; in Panel B we round p-values to two decimal places and use all p-values in our sample. In Column (1), *Freq* is the observed frequency for each p-value; in Column (2), *Residual Freq* is the observed frequency minuses the expected frequency of each p-value based on exponential fitting. The last column measures the ratio of forecast residual in (2) to observed frequency in (1).

p-value	(1) Freq	(2) Residual Freq	(3) (Residual Freq)/ Freq
Panel A: Using p-values reported with two-decimal places			
0.01	187	8.229	4.40%
0.02	143	-7.660	-5.36%
0.03	111	-15.969	-14.39%
0.04	88	-19.004	-21.60%
0.05	122	31.822	26.08%
0.06	60	-15.998	-26.66%
0.07	61	-3.047	-5.00%
0.08	58	4.024	6.94%
0.09	54	8.511	15.76%
0.10	44	5.664	12.87%
0.11	20	-12.308	-61.54%
0.12	20	-7.227	-36.14%
0.13	17	-5.946	-34.98%
0.14	21	1.662	7.91%
0.15	12	-4.297	-35.81%
Panel B: Rounding up all p-values to two decimal places			
0	447	-17.143	-3.84%
0.01	432	54.226	12.55%
0.02	293	-14.476	-4.94%
0.03	192	-58.260	-30.34%
0.04	202	-1.691	-0.84%
0.05	214	48.213	22.53%
0.06	121	-13.937	-11.52%
0.07	112	2.172	1.94%
0.08	88	-1.390	-1.58%
0.09	89	16.244	18.25%
0.10	67	7.782	11.62%
0.11	32	-16.198	-50.62%
0.12	30	-9.229	-30.76%
0.13	27	-4.929	-18.26%
0.14	36	10.012	27.81%
0.15	22	0.848	3.85%

Table 3. The magnitude and statistical significance of *p*-hacking.

This table reports the results of post hoc pairwise comparison for the residuals from curve fit in Figure 3. We compare the residuals at 0.05 and 0.06 to the residuals of the rest *p*-values. In Panel A, the sample only include *p*-values reported with two decimal places; in Panel B, we use the full sample of exact *p*-values. Column (1) is the observed frequencies for 0.05 and 0.06, (2) is the residual frequency from the fitted exponential function, computed as the difference between observed frequency and predicted frequency. Column (3) reports the ratio of the residual frequency in (2) to the observed frequency in (1). *P*-values in Column (4) are the *p*-values of mean differences between the residual of focal *p*-value (0.05 or 0.06) and the reference group, and *p*-values in the last column is adjusted for multiple comparison using the Dunnett's test.

	(1)	(2)	(3)	(4)	(5)
p-value	Freq	Residual Freq (Observed-predicted)	(2)/(1)	*p*-value	*p*-value adjusted for multiple comparison
Panel A: Only including p-values reported with two decimal places					
0.05	122	32	26.23%	0.002	0.005
0.06	60	-16	-26.67%	0.281	0.340
Panel B: Rounding all p-values to two decimal places					
0.05	214	48	22.43%	0.053	0.113
0.06	121	-14	-11.57%	0.594	0.856

Table 4. Comparing the extent of *p*-hacking across various subsamples.

This table provides results of post hoc pairwise comparison for each histogram in each pair of subsamples. For each histogram in a pair of subsamples, we derive the residual of exponential fit for every *p*-value from 0.00, 0.01, to 0.15, and then compare the residuals at 0.05 and 0.06 with the mean residuals of the rest 14 *p*-values. Results for 0.06 are all insignificant and thus omitted. Column (1) is the observed frequencies for 0.05, (2) is the residual frequency from the fitted exponential function, and (3) is the ratio of residual frequency to observed frequency. Column (4) reports the *p*-values from tests of difference between the residual at 0.05 and the mean residuals for *p*-values in the reference group, and *p*-values in the last column is adjusted for multiple comparison using the Dunnett's test.

Subsamples	(1) Freq	(2) Residual Freq (Observed-predicted)	(3) (2)/(1)	(4) *p*-value	(5) *p*-value adjusted for multiple comparison
Panel A: At least one tenured authors					
Yes	104	9	8.65%	0.311	0.526
No	72	27	37.50%	0.016	0.033
Panel B: Experience					
More	93	11	21.88%	0.417	0.648
Less	107	28	16.98%	0.022	0.051
Panel C: Contain female authors					
Yes	109	15	13.76%	0.122	0.238
No	105	28	26.67%	0.083	0.170
Panel D: Sole authored article					
Yes	67	24	35.82%	0.027	0.06
No	147	20	13.61%	0.157	0.298
Panel E: School Ranking					
High	100	32	32.00%	0.083	0.151
Low	114	17	14.91%	0.271	0.474
Panel F: PhD School Ranking					
High	96	21	27.38%	0.050	0.107
Low	106	18	9.64%	0.158	0.298
Panel G: N_ Participants					
High	111	28	25.23%	0.051	0.110
Low	99	14	14.14%	0.139	0.226
Panel H: Google_Citations					
High	129	31	24.03%	0.026	0.059
Low	85	11	12.94%	0.377	0.616
Panel I: Web_Citations					
High	132	34	25.76%	0.026	0.059
Low	62	1	1.61%	0.808	0.974

Table 5. Rejection rates after adjusting for multiple hypotheses testing.

The table reports the percentage of at least one significant interaction effect from analysis of variance (ANOVA) based on 1,000 random samples. The Column "Researcher's degree of flexibility" displays the flexibility available for researchers. The baseline is a 2 × 2 "experiment" with 20 observations per cell and only one dependent variable. In situation 1, a researcher collects two dependent variable and conduct 3 ANOVA tests, one on each variable and a third on the mean of the two; we report significant result if any one of the ANOVA delivers statistically significant interaction effect. In situation 2, the researcher can choose to whether collect additional 10 observations per cell or not. In situation 3, the researcher can add another condition or not, we report significant results in any two of the conditions have statistically significant interaction effects. Situation 4 to 7 are different combinations of the first 3 situations. In each situation, we get the lowest *p*-values among all interaction effects, and check if it is lower than 5%.

	% of *p*-values below significance threshold			
	(1)	(2)	(3)	(4)
Researcher's degree of flexibility	p<=.05	Bonferroni method	Bootstrap adjusted *p*-value	BH method
Situation 0: Baseline	3.80%	3.80%	3.90%	3.80%
Situation 1: Two dependent variables	7.60%	2.70%	3.90%	3.70%
Situation 2: Adding 10 obs per cell	6.30%	3.10%	4.20%	3.40%
Situation 3: Adding another IV	14.10%	4.30%	5.00%	5.80%
Situation 4: Combining 1 and 2	12.30%	2.30%	3.70%	2.90%
Situation 5: Combining 1 and 3	25.60%	2.60%	6.20%	4.80%
Situation 6: Combining 2 and 3	22.00%	4.50%	5.90%	5.30%
Situation 7: Combining 1, 2, and 3	34.30%	2.90%	6.60%	4.70%