P-hacking in Experimental Audit Research

Mohammad Jahanzeb Khan
Charles Sturt University

and

Per Christen Tronnes
UNSW Australia

Please do not cite without the authors' permission

September 2017

Corresponding Authors:

Mohammad Jahanzeb Khan (Jahanzeb)
Lecturer
School of Accounting and Finance
Charles Sturt University
Boorooma Street
North Wagga NSW 2678
Australia
Email: jkhan@csu.edu.au

Per Christen Tronnes
Lecturer
School of Accounting
UNSW Australia
High Street
Kensington NSW 2052
Australia
Email: p.tronnes@unsw.edu.au

Acknowledgements
Kevin Liang assisted on this project as a Research Assistant (UNSW Australia) in 2015.

P-hacking in Experimental Audit Research

ABSTRACT

A focus on novel, confirmatory, and statistically significant results by journals that publish experimental audit research may result in substantial bias in the literature. We explore one type of bias known as *p*-hacking: a practice where researchers, whether knowingly or unknowingly, adjust their collection, analysis, and reporting of data and results, until non-significant results become significant. Examining experimental audit literature published in eight accounting and audit journals in the last three decades, we find an overabundance of *p*-values at or just below the conventional thresholds for statistical significance. The finding of too many "just significant" results is an indication that some of the results published in the experimental audit literature are potentially a consequence of *p*-hacking. We discuss some potential remedies that, if adopted, may (to some extent) alleviate concerns regarding *p*-hacking and the publication of false positive results.

JEL Classification: M40

Keywords: Audit, Audit Experiment, Statistics, *p*-values, *p*-hacking

I. INTRODUCTION

To draw conclusions from data, experimental audit researchers (like any social scientist) usually rely on significance testing. This typically means calculating a *p*-value, which essentially denotes the probability of seeing the results observed in the data if there really was no effect of the experimental manipulations. If the *p*-value is sufficiently small, then the result is declared to be statistically significant, and the researcher(s) conclude that there is a systematic effect, rejecting their null hypothesis in favour of their alternative hypothesis. The conventional cut-off for the significant criterion in experimental audit research is usually a *p*-value that is less or equal to 0.05. However, some researchers also use a *p*-value that is less or equal to 0.10 (but higher than 0.05) as criterion to denote marginally significant results. Accounting and audit journals that publish experimental studies do not appear to be any different from other social science journals that place great importance on the significant criterion, and are therefore biased towards publishing studies that can report these apparent significant effects (Lindsay, 1994; Ioannidis, 2005).

The importance placed on *p*-values by journals, and the broader research community, is hammered in from an early stage in every researcher's career. Achieving the mythical *p*-value that is below the 0.05 threshold is the key to progress throughout their entire academic career. From first getting through graduate school and obtaining a PhD, to achieving publications in good journals that will determine their progress through the academic ranks, and to finally become a tenured professor. Accounting research is not an exception to this concerning trend in research practice, it is often the case that valuable data and information remain unpublished, due to unfavourably perceived *p*-values associated with the findings (Lindsay, 1994). The strong incentives and demands to publish may often shift the researchers focus from practicing rigorous and informative science, instead using methodology of convenience, whether

knowingly or unknowingly, in order to achieve *p*-values that fall below 0.05 (Nusso 2014; Masicamp and Lalande 2012; Head et al. 2015; Krawczyk 2015). This is a process known as *p*-hacking: manipulation or tweaking of the data collection, analysis, and reporting designed so as to achieve a desired outcome of a *p*-value less than or equal to 0.05, thus be able to claim "statistical significance" and subsequently publish the results. We examine whether there is any evidence of *p*-hacking by experimental audit researchers which would be indicated by an unusually high number of published effects which have corresponding *p*-values equal to, or just below, 0.05 and 0.10 – as these are the conventional cut-off values for statistically significant and marginally significant results, respectively. Even in accounting research, these conventions need to be acknowledged and addressed, as accounting research is still at a stage where there is still room for further discourse and attention with regards to these research and statistical practices (Lindsay, 1997). Addressing these concerns and issues is crucial to mitigate unintended consequences, in particular the risk of type-1 error inflation.

We collect and analyse 2,631 reported *p*-values from 411 published experimental audit research articles in eight of the top accounting/audit journals (Lesage and Wechtler, 2012). Our findings indicate that the number of *p*-values in the experimental auditing literature that barely meet the criterion for statistical significance (i.e., that are reported as equal to or fall just below 0.05) is unusually large and the same appears to be the case for the criterion for marginal statistical significance (i.e., that are reported as equal to or fall just below 0.10), given the number of *p*-values occurring in other ranges, and predicted based on the overall distribution of *p*. Consistent with findings across other social science disciplines, we interpret this as evidence that *p*-hacking exists within the experimental auditing literature.

A systematic overrepresentation of false positive results in the academic literature hinders scientific progress. When false positive results enter the literature they can become very persistent, because in many fields, including experimental auditing research, there is little

incentive to replicate research (Lindsay and Ehrenberg, 1993). False positives can inspire investments in fruitless research programs and even discredit entire fields (Lindsay, 1994; Trafimow and Marks, 2015). Importantly, it can mask potential implications of non-significant *p*-values (Franco et al., 2014), which are either manipulated into false positives or remain unpublished. As such, adopting practices that can discourage *p*-hacking is an important scientific endeavour. Our paper contributes to this discourse.

It is important to clarify the intention of this investigation, and importantly, the implications of the reported findings. Firstly, our results neither imply that the experimental audit research literature is unreliable as a whole, nor does it discredit the usefulness and importance of null hypothesis significance testing as a helpful tool for scientific reasoning. Quite the contrary, our results also does seem to confirm that there is "evidential value" in the literature as a whole, and that in aggregate the published statistically significant *p*-values predominately document non-zero effects. However, our results also indicate that *some* of the published result in experimental audit research may potentially be prone to *p*-hacking practices and consequently reporting false positive results. In this respect, experimental audit research is no different than other social sciences (Ioannidis, 2005; Masicampo and Lalande 2012; Head et al. 2015).[1] Second, our findings may simply indicate that experimental audit researchers are not immune to human biases – and given that the rewards of academia are strongly linked to publishing results – this may lead some of them to, whether consciously or unconsciously, tip the scales by setting up the data collection and analysis to yield false positive results. Another

[1] We would also like to stress that we do not think that experimental audit research is somehow particularly prone to arbitrary or self-serving choices in data analysis and reporting, especially in comparison to other type of research within the broader accounting discipline. We simply chose not to conduct any analysis on archival audit research because those publications are full of multivariate regressions and control variables. This would require us to distinguish between *p*-values reported on control variables versus the *p*-values of the treatment effects of interest, or accept much noise in the collected *p*-values. Experimental audit research tends to focus on the treatment effects of interest and report those *p*-values.

consequence of this behaviour is that non-significant results (Lindsay, 1994; Rosenberg, 2005; Hubbard and Lindsay, 2013) are essentially shelved, despite their potential for interesting and important implications.

How we evaluate evidence is the very foundation of all statistical work. At the same time, many of the questions we continually ask ourselves throughout the research process – such as: What conclusions are justified based on which process? ; How should we interpret the result from an experiment? ; When is data good or bad? – are fundamentally philosophical in nature, where reasonable researchers may disagree and hold different yet justifiable opinions (Hubbard and Lindsay, 2013). We believe our results are a reminder that a single study or experiment neither proves nor disproves assertions. Instead, they can only provide evidence for or against such assertions.

The remainder of our paper is organized as follows. In the following two sections, we give a brief background to *p*-values and null hypothesis testing, and the notion of *p*-hacking. In section IV, we develop our expectations as to the distribution of *p*-values and our process of collecting the evidence. We report on our results in Section V. Section VI presents our conclusions and the implications our findings have for experimental audit research. We also present some suggestions of practices that can discourage *p*-hacking.

II. P-VALUES AND NULL HYPOTHESIS TESTING

Fisher (1925) introduced null hypothesis significance testing (NHST) to objectively separate interesting findings from background noise, and is the most frequently used data analysis method in experimental audit research. The null hypothesis is a statement of no relationship between the variables, or no effect of the experimental manipulation/intervention. When using NHST, one computes the probability (i.e. the *p*-value) of finding the observed effect (or a more extreme effect) in the data given that the null hypothesis of no effect is true. If the analysis

reveals a *p*-value equal to or below the arbitrary cut-off of 0.05 (or sometimes 0.1) then the effect is considered (marginally) statistically significant. That is, it would be highly improbable to obtain such results in the data if the null hypothesis were true. The null hypothesis is therefore rejected in favour of the alternative hypothesis that a relationship or effect exists. Thus, findings with small *p*-values that are equal or fall below the arbitrary cut-off 0.05 (or sometimes 0.1) are described as "statistically significant".

The *p*-value, however, can easily be misinterpreted as it is often equated with the strength of the relationship (i.e. effect size) between the variables of interest. A small effect size can have very low *p*-values with a large enough sample size, and so a low *p*-value does not necessarily mean that a finding is of major importance. The practice of predominantly relying on *p*-values to draw conclusions from experimental data has attracted a fair share of critics over the years (e.g. Rozeboom, 1960; Bakan, 1966; Falk and Greenbaum, 1995; Cohen, 1994; Ziliak and McCloskey, 2008; Nuzzo, 2014; ASA, 2016; among others). While we do indeed share most of the sentiments raised by the critics of *p*-values (Lindsay, 1997; Hubbard and Lindsay 2013), we would like to stress that the goal of this research, in similar fashion to *p*-hacking investigations in other disciplines (e.g. Misacampo and Lalande, 2012), is not to discredit the usefulness and importance of NHST. Rather, our aim is simply to test for evidence of potential misuse, whether intentional or unintentional, of NHST in experimental audit research, particularly through *p*-hacking.

III. P-HACKING PRACTICES

As a consequence of the reliance on NHST, a prevalent issue with regard to the academic record is publication bias. Publication bias is the greater likelihood of statistically significant results being published than statistically non-significant results, holding fixed research quality (Rosenthal, 1979; Scargle, 1999; Duval and Tweedie, 2000; Franco et al., 2014). As a consequence of this publication bias, non-significant results are much more difficult to "sell"

despite their implications (Lindsay, 1994).[2] While publication bias can lead to what is known as the "file-drawer problem" (Rosenthal, 1979; Rosenberg, 2005; Franco et al., 2014), publication bias may also create incentives for more questionable practices: researchers may engage in creative analysis and reporting practices in order to tweak the results to achieve a low enough *p*-value and claim statistical significance for their findings (Lindsay, 1997), thereby increasing their chances of publication, even if their initial analysis yielded "non-significant" results.

Commonly identified practices that may potentially lead to *p*-hacking include (among others): conducting analysis midway through experiments to decide whether to continue collecting data; recording many response variables and deciding which to report post analysis; deciding whether to include or drop outliers post analysis; excluding combining or splitting treatment groups post analysis; stopping data exploration if an analysis yields a significant *p*-value; decisions to include or exclude co-variates post analysis; run through a series of different sophisticated tests (e.g. run a series of parametric and non-parametric statistical test); reporting only those effects that yield the lowest *p*-value; report on only some of the experimental sessions; and even choosing not to report on all of a study's conditions (Masicampo and Lalande, 2012; Hubbard and Lindsay, 2013; Nuzzo, 2014; Head et al., 2015; Krawczyk, 2015).[3]

[2] Franco et al (2014) documents an interesting effect with regard to the file drawer problem. It is not necessarily the case that studies that report non-significant effects is rejected from journals, but often the case that because of a disciplines strong preference for statistically significant results, researchers believe non-significant results have no publication potential (even if they themselves view the finding interesting) so that non-significant results are often not written up in the first place.

[3] Some of these practices are legitimate research design choices. For example, it is a justifiable practice to exclude observations that failed the manipulation check. However, when such practices lack transparency and are used aggressively for the sole reason to turn a non-significant result into a significant one to increase the chances of having the result published, then it becomes an issue of *p*-hacking.

Furthermore, the researcher can choose to report the *p*-value directly (e.g. "*p*=…") or by means of inequality at some conventional cut-off for statistical significance (e.g. "p≤.05"). If the researcher decides on reporting it directly, then the decision on decimal places should also be disclosed. For example, a *p*-value equal to 0.05351 can also reported as 0.0535, 0.054 or simply as 0.05. It is apparent that reporting with less precision, for example to just two decimal places, is appealing to the researcher as they can potentially claim a statistically significant effect by rounding down. The researcher may also revise initial hypothesis to being directional post analysis so as to report one-tailed *p*-values in order to achieve lower *p*-values.

If the researcher makes choices with regard to their analysis and reporting with the sole view of obtaining low *p*-values, this may very well yield a reportable *p*-value for which the researcher can claim a significant effect, and thereby increase the chances of their findings being published. It may also be that by engaging in this practice and achieving the desired outcomes, the researcher is able to convincingly justify their choices from a methodological viewpoint (Krawczyk, 2015). Nevertheless, and while the practice of *p*-hacking may not be considered unethical to the same degree as data fabrication (Stone, 2015), it is still a practice which may affect both the actual and perceived reliability of *p*-values to draw meaningful conclusions. Furthermore, potentially meaningful and authentic conclusions that could have been drawn from non-significant results either remain unpublished or are eventually published as false positives.

IV. TESTING FOR EVIDENCE

We examine whether the distribution of *p*-values is disturbed around the critical values of 0.05 (and to a lesser degree 0.10). Head et al. (2015) note that that if the true effect size for a studied phenomenon is zero, every *p*-value is equally likely to be observed. That is, the expected distribution of *p*-values under the null hypothesis is uniform because *p*-values less than 0.05 will occur 5% of the time, and p-values less than 0.04 will occur 4%, *p*-values less than 0.03

will occur 3% of the time and so on. On the other hand, when the true effect size is non-zero, the expected distribution of *p*-values is exponential with a right skew. That is, when the true effect is strong, researchers are more likely to obtain very low *p*-values than moderate *p*-values and are less likely still to obtain very *p*-values that are above conventional levels for statistical significance. So if true effect sizes are present, the distribution of *p*-values should be right skewed; and as the true effect size increases, so does the right skewness of this *p*-value distribution.

The distribution of *p*-values can reveal certain interesting characteristics regarding the published literature. A notable drop in observed p-values *above* conventional thresholds for what is considered statistically significant can be interpreted as evidence of publication bias, but it does not distinguish between whether there is a file drawer problem and *p*-hacking (Gerber and Malhotra 2008; Masicampo and Lalande, 2012: Leggett et al 2013). On the other hand, if researchers *p*-hack and turn a marginal non-significant result into a significant one, then the distribution of *p*-values will also be disturbed *below* the conventional significance thresholds (such as $p \leq 0.05$ and $p \leq 0.1$). Specifically, the distribution of *p*-values will have an overabundance of *p*-values at or just below these thresholds. That is, both *p*-hacking and selection bias such as the file drawer problem suggest that the distribution of *p*-values will have discontinuity in the *p*-values around thresholds for statistical significance, but only *p*-hacking predicts an overabundance of *p*-values just below the thresholds (Head et al. 2015). The analysis of *p*-values in other academic fields generally find that they are unable to account for the overly large number of "just significant" findings at or below 0.05 and attribute this to the possibility *p*-hacking (e.g. Masicampo and Lalande, 2012; Head et al., 2015). We apply a similar procedure to Masicampo and Lalande (2012) to investigate whether the experimental audit literature contains evidence of *p*-hacking.

We expect that if a sufficient number of researchers in the experimental audit literature engage in *p*-hacking in order to meet the NHST standards of a *p*-value less than or equal to 0.05 (and 0.1), then that may be reflected in the distribution of *p*-values across published effects, given that we obtain a sufficiently large sample of *p*-values from the literature (drawn from a sufficiently large set of studies testing a range of effect sizes). More specifically, the number of *p*-values equal to or immediately below the arbitrary cut-off (Masicampo and Lalande, 2012) of 0.05 (and 0.10) may be much higher than what would be expected based on the frequency of the *p*-values in other segments of the distribution.

We identified published experimental auditing and assurance research for inclusion into our sample in two ways. First, we used the database of Audit Research prepared by the AAA Auditing Section Research Committee (2009). This database documents all auditing articles published in eight journals (*AOS*, *AJPT*, *BRIA*, *CAR*, *JAE*, *JAPP*, *JAR*, and *TAR*) over a 33–year period up till the year 2009. It also classifies these articles by research methods, and the articles selected were the ones identified as experimental. Second, we verified this listing and then extended this database from 2009 to 2015 by reviewing articles published in these eight journals, searching by title and keywords for any relationship to audit or assurance. Since our inferences are based on mapping the distribution of *p*-values, we only used *p*-values that were reported exactly (*p*=…), and excluded *p*-values that were reported by means of inequality (e.g. $p \leq 0.05$).[4] Furthermore, we followed Masicampo and Lalande (2012) and focused on *p*-values greater than 0.01, but unlike them we extended the upper bound of the range from 0.10 to include *p*-values that were less or equal to 0.15.[5] In that way both our critical values for

[4] We note that a large proportion of *p*-values in the experimental audit literature are not reported directly and rather as by means of inequality at some conventional cut off for statistical significance (e.g. "$p \leq 0.05$").

[5] We note that there are quite a few reported *p*-values in the range $0<p\leq.0.01$ in the literature that we surveyed. This is in line with previous observed patterns in psychology that report a sharply decreasing density with most values located very close to 0 (see Krawcyk 2015). In line with Masicampo and Lalande (2012), we excluded these observations when mapping the distribution of precise *p*-values.

significant (0.05) and marginal significant (0.10) results are included in the range of *p*-values and have sufficient data points on both sides of the threshold. We identified 411 published articles that reported one or more exact *p*-values greater than 0.01, but less or equal to 0.15 in either the article's abstract, text, tables or footnotes. Our final sample encompassed 2,631 reported *p*-values falling on a continuum from 0.01 to 0.15.

V. RESULTS

We conducted two separate analyses and followed the procedure outlined in Masicampo and Lalande (2012): we divided the range of interest (0.01 to 0.15) into intervals of equal size to examine the frequency distribution of *p*-values. The only difference between the two analyses was the size of the intervals into which the range of *p*-values was divided: 0.01, and 0.005. For each analysis, we counted the number of *p*-values within the various intervals, resulting in a frequency distribution of *p*. Curve estimation procedures were used to determine the best fit for the resulting distributions. An exponential model best fitted the data points regardless of the size of the intervals (see Figure 1, Panels A and B).

[Insert Figure 1 Here]

The graphs in both Panels A and B Figure 1 shows trend lines with acceptable fit. The R^2 of the trend line in Panel A is 92.18% and the R^2 of the trend line in Panel B is 72.94%. The graph in Panel B where the divisions of the *p*-value range are 0.005 shows spikes in the number observations around the *p*-values ranges where the upper boundaries are 0.01,0.02,…, 0.15 in comparison to the *p*-value ranges where the upper boundaries are 0.015, 0.025,..., 0.145. This corresponds to the fact that many reported *p*-values within the experimental audit literature are rounded to two decimal points. This also explains the relative lower fit of the trend line in Panel B in comparison to the fit of the trend line in Panel A.

Both trend lines in Panel A and Panel B show that the distribution of published p-values is exponential with a right skew. This suggests experimental audit researchers appear to be predominately studying phenomena with non-zero effect sizes, and that there is "evidential value" in the literature as a whole. It is reassuring that that the observed distribution of *p*-values are consistent with most experimental audit researchers investigate phenomena that lead to refutation of the null hypothesis, implying that the average true effect size studied by experimental audit researchers is nonzero. However, in both graphs the frequency of observed *p*-values that falls just below the $p \leq 0.05$ and $p \leq 0.10$ thresholds deviates more from the trend line than any other place on the distribution of *p*-values and this may be suggestive of some *p*-hacking.

In Table 1, Panel A and B, we tabulated the number of actually observed *p*-values, and compare them to the expected frequencies based on the trend line for each *p*-value ranges. Panel A, Table 1, tabulate the actual and expected number of observations for divisions of 0.01 and as such corresponds to Panel A, Figure 1. Panel B, Table 1, tabulate the actual and expected number of observations for divisions of 0.005 and as such corresponds to Panel B in Figure 1.

[Insert Table 1 Here]

Panel A in Table 1 shows that the number of actually observed *p*-values that is in the range that falls just below the threshold $p \leq 0.05$ is 57.09% higher than what we would expect based on the trend line. Similarly, the number of actually observed *p*-values that is in the range that falls just below the threshold $p \leq 0.10$ is 47.43% higher than what we would expect based on the trend line. These positive deviations are, respectively, the highest and the second highest deviations from the trend line (in absolute value). It is also telling that the number of reported *p*-values in the two ranges, $0.04 < p \leq 0.05$ and $0.09 < p \leq 0.10$, is higher than the number of reported *p*-values

in the preceding ranges, given that the trend line for the *p*-value distribution is downward sloping.

Panel B in Table 1 shows a similar pattern. The number of actually observed *p*-values that is in the range that falls just below the threshold $p \leq 0.05$ is 119.53% higher than what we would expect from the trend line. Similarly, the number of actually observed *p*-values that is in the range that falls just below the threshold $p \leq 0.10$ is 121.29% higher than what we would expect from the trend line. Again, these positive deviations are the two highest deviations from the trend line (in absolute value). Panel B also shows a similar pattern to that depicted in Panel A: the number of reported *p*-values in the two ranges, $0.04 < p \leq 0.05$ and $0.09 < p \leq 0.10$, is higher than the number of reported *p*-values in the three preceding ranges.

In Table 2 we formally test whether the proportion of expected number of p-values to the actually observed number of p-values in the "bins" just below the thresholds of 0.05 and 0.10 is higher than the proportion in the adjacent "bin" immediately before.

[Insert Table 2 Here]

In all the four test conducted, and given a null hypothesis of no difference in deviation from the trend lend on observed p-values for adjacent bins below the thresholds for statistical significance, we find that there is a low probability of observing the high number of observations of *p*–values we do just below or at these threshold, if the null hypothesis was true. While the irony of using *p*-values to show that the proportion of observed to expected number of *p*-values in the bins just below conventional thresholds for statistical significance is not lost on the authors, we believe that the analysis in Table 2 clearly complement what is shown in Figure 1 and Table 1: namely, there appears to be an overabundance of *p*-values in the published experimental audit literature that is just under the conventional thresholds for what is considered statistical significance.

Our result from Figure 1 and Tables 1 and 2 indicates that the number of *p*-values in the experimental auditing literature that barely meet the criterion for statistical significance (i.e., equal to or that fall just below 0.05) is unusually large, and the same appears to be the case for the usual criterion for marginal statistical significance (i.e., equal to or that fall just below 0.10), given the number of *p*-values occurring in other ranges and predicted based on the overall distribution of *p*.

The anomaly in the *p*-value distribution is interpreted as evidence that results based on *p*-hacking may exist in the published literature for experimental auditing and assurance discipline, most likely a reflection of a researcher's decisions, and other discretionary actions that the researcher may take, in order to obtain and report favourable (small/significant) *p*-values (Simmons, Nelson, and Simonsohn, 2011). The anomaly in the *p*-value distribution is also consistent with the notion that researchers in the experimental audit discipline, similar to other experimental sciences, may be responding to the pressure of reporting statistical significance in order to improve the likelihood of publication (Sterling, 1959). It may also be that this practice is further reinforced by both reviewers and editors conforming to a standard of obtaining statistical significance in order for findings (and their contributions) to be considered meaningful (Masicampo and Lalande, 2012). Yet, there is also evidence that the literature as a whole contains "evidential value" because the distribution of *p*-values is also clearly right skewed (Head et al. 2015).

VI. CONCLUSIONS

We collect and analyse 2,631 reported *p*-values from 411 published experimental audit research articles in eight of the top accounting/audit journals. While or study find a distribution of *p*-values that are consistent with that the experimental audit literature has evidential value and document non-zero effect sizes as a whole, our study also provides some evidence that *p*-

hacking exists within the experimental audit literature. This is problematic as publication of false positives hinders scientific progress. Eliminating *p*-hacking entirely is unlikely when career advancement is assessed by publication output, and publication decisions are affected by the *p*-value or other measures of statistical support for relationships (Head et al., 2015). However, this does not necessarily imply that the experimental audit research community should be complacent to such practices. Firstly, it is important to acknowledge the potential consequences of overreliance on *p*-values. A solid understanding by researchers, reviewers, and editors of what *p*-values measure, and what they do not measure, and what other alternatives are available, could potentially do a great deal to alleviate future concerns regarding *p*-hacking. The American Statistical Association recently released statement on statistical significance and *p*-values, and this is a good starting point (along with the supporting submissions and discussions to this statement) which highlights that while *p*-values have their importance, they are not a substitute for scientific reasoning. In their own words the ASA (2016, p. 132) note the following (see also Wasserstein and Lazar, 2016).

Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning.

Second, and as noted by Stone (2015), there are recent proactive advances by accounting journals to make positive steps in the right direction. For example, the Journal of Accounting Research (JAR) recently released more stringent submission requirements, effective January 2015, that require inclusion of a description of how the raw data were obtained or generated and a complete description of the steps taken to collect and process the data used in the final analyses reported in the paper. Furthermore, Dyckman and Zeff (2014) and Salterio (2014)

both pointed out the criticality of replication to the scientific process and in this respect Behaviour Research in Accounting has recently committed to promoting and encouraging replication studies.

Lastly, a potential solution is the adoption of a pre-registration policy by accounting journals. Preregistration requires that the researcher prepare in advance a detailed research plan, including the statistical analysis to be applied to the data. By proposing and committing to the plan outlined, it provides reviewers, editors, and readers the opportunity to check the preregistered plan against what is reported, and thus provide more confidence that the analysis was conducted as originally intended. Such an approach could likely be a strong deterrent against *p*-hacking practices (Christensen and Miguel, 2016). The practice of pre-registration of studies is an idea that is currently getting more traction in other disciplines and facilitated by the fact that open research registries already exist (such as the Open Science Framework) where researchers can preregister their research plan with a date stamp.

Ultimately, if the standard for publication is predominantly based on statistical significance, then this will take on the appearance of Campbell's law (Nichols et al., 2007): When a measure becomes a target, it ceases to be a good measure, simply because researchers will attempt to manipulate it. Therefore, it is important to continually take steps to ensure that the processes of evaluating and reporting research findings (including the use of *p*-values) are transparent and rigorous to maintain the integrity and confidence in findings reported by published experimental audit studies. We hope that our paper makes a positive contribution to the important discourse on how we can advance good practices in evaluating research findings.

REFERENCES

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66, 1-29.

Christensen, G. S., & Miguel, E. (2016). *Transparency, Reproducibility, and the Credibility of Economics Research* (No. w22989). National Bureau of Economic Research.

Cohen, J (1994). The earth is round (p<.05). *American Psychologist*. 49(12):997-1003

Duval, S., & Tweedie, R. (2000). Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455-463.

Dyckman, T. R., & Zeff, S. A. (2014). Some methodological deficiencies in empirical research articles in accounting. *Accounting Horizons*, *28*(3), 695-712.

Fisher, R. A. (1925). *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biol*, *13*(3), e1002106.

Falk, R. & Greenbaum, C.W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology*. 5(1): 75-98.

Franco, A., Malhotra, N. & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345, 1502–1505.

Hubbard, R., & Lindsay, R. M. (2013). From significant difference to significant sameness: Proposing a paradigm shift in business research. *Journal of Business Research*, *66*(9), 1377-1388.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLos med*, *2*(8), e124.

Killeen, P. R. (2005). Replicability, confidence, and priors. *Psychological Science*, *16*(12), 1009-1012.

Kline, R. B. (2009). Becoming a behavioral science researcher.

Krawczyk, M. (2015). The search for significance: a few peculiarities in the distribution of p values in experimental psychology literature. *PloS one*, *10*(6), e0127872.

Leek, J. T., & Peng, R. D. (2015). Statistics: P values are just the tip of the iceberg. Nature, 520(7549), 612.

Lesage, C., & Wechtler, H. (2012). An inductive typology of auditing research. *Contemporary Accounting Research*, *29*(2), 487-504.

Lindsay, R. M. (1994). Publication system biases associated with the statistical testing paradigm. *Contemporary Accounting Research*, *11*(1), 33-57.

Lindsay, R. M. (1997). Lies, damned lies and more statistics: the neglected issue of multiplicity in accounting research. *Accounting and Business Research*, *27*(3), 243-258.

Lindsay, R. M., & Ehrenberg, A. S. (1993). The design of replicated studies. *The American Statistician*, *47*(3), 217-228.

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below. 05. *The Quarterly Journal of Experimental Psychology*, *65*(11), 2271-2279.

Nichols, S. L., Berliner, D. C., & Noddings, N. (2007). Collateral damage: How high-stakes testing corrupts America's schools.

Nuzzo, R. (2014). Statistical errors: P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. *Nature*, *506*(7487), 150-153.

Rosenberg, M. S. (2005). The file-drawer problem revisited: a general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution*, *59*(2), 464-468.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, *86*(3), 638.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological bulletin*, *57*(5), 416.

Salterio, S. E. (2014). We Don't Replicate Accounting Research—Or Do We?. *Contemporary Accounting Research*, *31*(4), 1134-1142.

Scargle, J. D. (1999). Publication Bias (The" File-Drawer Problem") in Scientific Inference. *arXiv preprint physics/9909033*.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological methods*, *1*(2), 115.

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*, 0956797611417632.

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*(285), 30-34.

Stone, D. N. (2015). Post-hunton: Reclaiming our integrity and literature. *Journal of Information Systems*, *29*(2), 211-227.

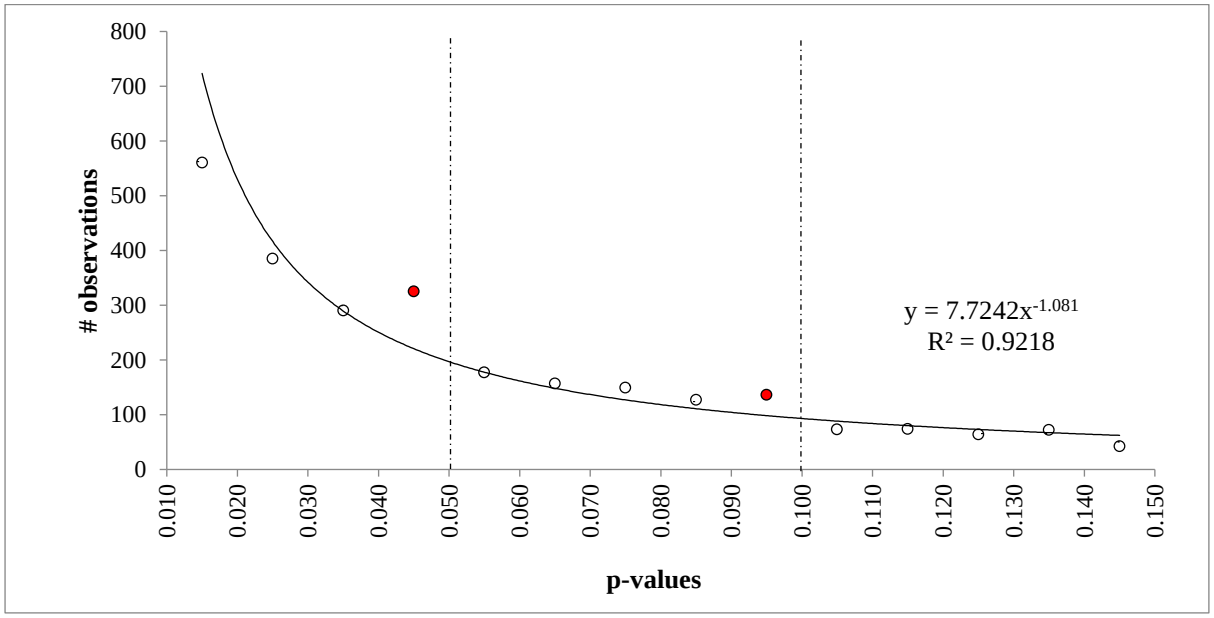
Trafimow, D. & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*, 1–2.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*.

Ziliak, S. T. & McCloskey, D.N. (2008). *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives.* University of Michigan Press

Figure 1: The graphs show the distribution of 2,631 p –values from experimental audit research

Panel A: Frequencies at divisions of 0.01



Panel B: Frequencies at divisions of 0.005

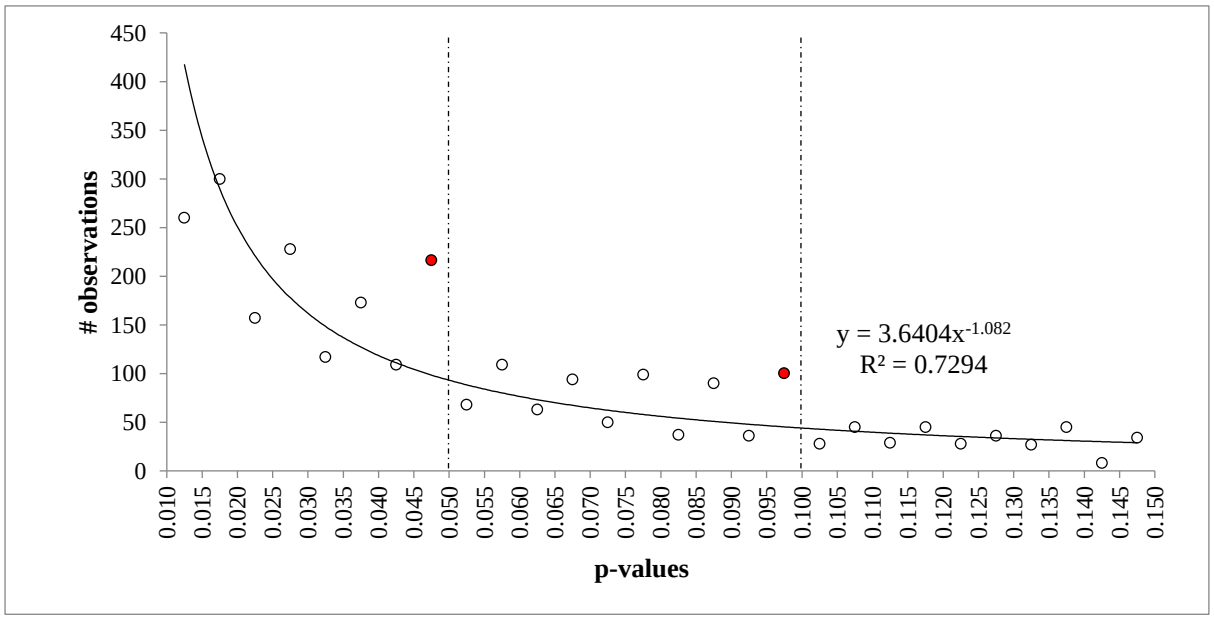


Table 1: Deviations from trend in each range of the various p-value distributions.

Panel A: Divisions of .01

P-value Range	# observed	# predicted	% Deviation
0.01<p≤0.02	560	678.43	-17.46%
0.02<p≤0.03	385	390.56	-1.42%
0.03<p≤0.04	290	271.47	6.83%
0.04<p≤0.05	325	206.89	57.09%
0.05<p≤0.06	177	166.54	6.28%
0.06<p≤0.07	157	139.03	12.93%
0.07<p≤0.08	149	119.10	25.10%
0.08<p≤0.09	127	104.03	22.08%
0.09<p≤0.10	136	92.24	47.43%
0.10<p≤0.11	73	82.79	-11.82%
0.11<p≤0.12	74	75.03	-1.37%
0.12<p≤0.13	64	68.56	-6.66%
0.13<p≤0.14	72	63.09	14.12%
0.14<p≤0.15	42	58.40	-28.08%

Panel B: Divisions of .005

P-value Range	# observed	# predicted	% Deviation
0.010<p≤0.015	260	417.15	-37.67%
0.015<p≤0.020	300	289.86	3.50%
0.020<p≤0.025	157	220.85	-28.91%
0.025<p≤0.030	228	177.74	28.28%
0.030<p≤0.035	117	148.35	-21.13%
0.035<p≤0.040	173	127.07	36.14%
0.040<p≤0.045	109	110.98	-1.78%
0.045<p≤0.050	216	98.39	119.53%
0.050<p≤0.055	68	88.30	-22.99%
0.055<p≤0.060	109	80.02	36.22%
0.060<p≤0.065	63	73.12	-13.83%
0.065<p≤0.070	94	67.27	39.73%
0.070<p≤0.075	50	62.27	-19.70%
0.075<p≤0.080	99	57.93	70.89%
0.080<p≤0.085	37	54.14	-31.66%
0.085<p≤0.090	90	50.80	77.15%
0.090<p≤0.095	36	47.84	-24.75%
0.095<p≤0.100	100	45.19	121.29%
0.100<p≤0.105	28	42.81	-34.59%
0.105<p≤0.110	45	40.66	10.67%
0.110<p≤0.115	29	38.71	-25.08%
0.115<p≤0.120	45	36.93	21.86%
0.120<p≤0.125	28	35.30	-20.68%
0.125<p≤0.130	36	33.81	6.49%
0.130<p≤0.135	27	32.43	-16.74%
0.135<p≤0.140	45	31.15	44.45%
0.140<p≤0.145	8	29.97	-73.31%
0.145<p≤0.150	34	28.87	17.75%

Table 2: Statistcal test of proportion of observed to expected number of p-values in "bins" just below thresholds for "statistical significance

P-value Bin: 0.03<p≤0.04	**P-value Bin: 0.04<p≤0.05**
$n_1 = 290$ *observed p-values*	$n_2 = 325$ *observed p-values*
$y_1 = 271.49$ *expected p-values*	$y_2 = 206.89$ *expected p-values*
$\hat{p}_1 = \frac{271.49}{290} = 0.9362$	$\hat{p}_2 = \frac{206.89}{325} = 0.6366$

H_0: $p_1 = p_2$; H_A: $p_1 \neq p_2$
Z statistic = 8.9223 ; p-value <0.00001
Conclsuion: Null hypothesis rejected

P-value Bin: 0.08<p≤0.09	**P-value Bin: 0.09<p≤0.10**
$n_1 = 127$ *observed p-values*	$n_2 = 136$ *observed p-values*
$y_1 = 104.03$ *expected p-values*	$y_2 = 92.23$ *expected p-values*
$\hat{p}_1 = \frac{104.03}{127} = 0.8191$	$\hat{p}_2 = \frac{92.23}{136} = 0.6782$

H_0: $p_1 = p_2$; H_A: $p_1 \neq p_2$
Z statistic = 2.6239 ; p-value = 0.008693
Conclsuion: Null hypothesis rejected

P-value Bin: 0.040<p≤0.045	**P-value Bin: 0.045<p≤0.050**
$n_1 = 109$ *observed p-values*	$n_2 = 216$ *observed p-values*
$y_1 = 110.89$ *expected p-values*	$y_2 = 98.39$ *expected p-values*
$\hat{p}_1 = \frac{110.89}{109} = 1.0173$	$\hat{p}_2 = \frac{98.39}{216} = 0.4555$

H_0: $p_1 = p_2$; H_A: $p_1 \neq p_2$
Z statistic = 9.9831 ; p-value <0.00001
Conclsuion: Null hypothesis rejected

P-value Bin: 0.090<p≤0.095	P-value Bin: **.0.095<p≤0.100**
$n_1 = 36$ *observed p-values*	$n_2 = 100$ *observed p-values*
$y_1 = 47.84$ *expected p-values*	$y_2 = 45.19$ *expected p-values*
$\hat{p}_1 = \frac{47.84}{36} = 1.3289$	$\hat{p}_2 = \frac{45.19}{100} = 0.4519$

H_0: $p_1 = p_2$; H_A: $p_1 \neq p_2$
Z statistic = 9.7057 ; p-value <0.00001
Conclsuion: Null hypothesis rejected

Notes: The test statistic for testing the difference in two population proportions, that is, for testing the null hypothesis

$$H_0: p_1 - p_2 = 0$$

is:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where:

$$\hat{p} = \frac{y_1 - y_2}{n_1 + n_2}$$

is the proportion of "expected observations" in the two samples combined.