Ian D. Gow*

The Elephant in the Room: p-hacking and Accounting Research

https://doi.org/10.1515/ael-2022-0111
Received December 14, 2022; accepted September 1, 2023; published online September 21, 2023

Abstract: Ohlson (2025. Empirical accounting seminars: Elephants in the room. *Accounting, Economics, and Law: A Convivium 15*, 1–8) draws on his experience in empirical accounting seminars to identify five "elephants in the room". I interpret each of these elephants as either a variant or a symptom of p-hacking. I provide evidence of the prevalence of p-hacking in accounting research that complements the observations made by Ohlson (2025. Empirical accounting seminars: Elephants in the room. *Accounting, Economics, and Law: A Convivium 15*, 1–8). In this paper, I identify a number of steps that could be taken to reduce p-hacking in accounting research. I conjecture that facilitating and encouraging replication alone could have profound effects on the quality and quantity of empirical accounting research.

Keywords: empirical accounting research; p-hacking; empirical methods

JEL Classification: M; C1

Table of Contents

***Corresponding author: Ian D. Gow**, University of Melbourne, Melbourne, VIC, Australia,
E-mail: ian.gow@unimelb.edu.au. https://orcid.org/0000-0002-6243-8409

Empirical Research in Accounting and Social Sciences: Elephants in the Room

1. Empirical Accounting Seminars: Elephants in the Room, by James A. Ohlson, https://doi.org/10.1515/ael-2021-0067.
2. Limits of Empirical Studies in Accounting and Social Sciences: A Constructive Critique from Accounting, Economics and the Law, by Yuri Biondi, https://doi.org/10.1515/ael-2021-0089.
3. Accounting Research's "Flat Earth" Problem, by William M. Cready, https://doi.org/10.1515/ael-2021-0045.
4. Accounting Research as Bayesian Inference to the Best Explanation, by Sanjay Kallapur, https://doi.org/10.1515/ael-2021-0083.
5. The Elephant in the Room: p-hacking and Accounting Research, by Ian D. Gow, https://doi.org/10.1515/ael-2022-0111.
6. De-emphasizing Statistical Significance, by Todd Mitton, https://doi.org/10.1515/ael-2022-0100.
7. Statistical versus Economic Significance in Accounting: A Reality Check, by Jeremy Bertomeu, https://doi.org/10.1515/ael-2023-0002.
8. Another Way Forward: Comments on Ohlson's Critique of Empirical Accounting Research, by Matthias Breuer, https://doi.org/10.1515/ael-2022-0093.
9. Setting Statistical Hurdles for Publishing in Accounting, by Siew Hong Teoh and Yinglei Zhang, https://doi.org/10.1515/ael-2022-0104.

1 Five Elephants or One?

Ohlson (2025) identifies five of what he calls "elephants in the room" (or topics considered taboo in seminars). I read these not so much as five elephants, but as five alternative descriptions of the one elephant, much like the elephant in the parable of the blind men and an elephant (Wikipedia, 2023).

What exactly is that elephant in the room? I argue that Ohlson's five elephants are simply alternative perspectives on the same elephant, which is p-hacking, a term for a set of practices engaged in by researchers searching for "significant" and "positive" results.[1] To be sure, this is a very big elephant: I conjecture that p-hacking is the dominant mode of research in academic accounting in 2023, and below I provide (admittedly circumstantial) evidence consistent with this conjecture.

That the basic concern of Ohlson (2025) is with p-hacking is clearest with the last of his five elephants: "Issues Related to 'Screen-Picking' and 'Data-Snooping' ", as terms like "data-snooping" are simply synonyms of p-hacking.[2] The key insight of Ohlson

1 Here "significant" refers to statistical significance and "positive" refers to results that reject so-called "null hypotheses" and thereby (purportedly) push human knowledge forward. As pointed out by Simmons (2018), it is very easy for researchers to engage in p-hacking without being conscious that they are doing so.

2 I discuss in an appendix below (Section 6) how the remaining four elephants relate to p-hacking.

(2025) may be in highlighting how merely suggesting the possibility of p-hacking is taboo (Ohlson, 2025 uses terms such as "unacceptable," "a personal assault," "too sordid," "testing ethical boundaries," and "a more or less painful private matter").

Many researchers appear not to understand how p-hacking vitiates the whole research endeavor. So if even suggesting the possibility p-hacking is taboo, it will be much more difficult to address and accounting research will continue to be a largely pointless exercise.[3] I agree with Ohlson (2025) that we need to confront this "elephant in the room" and make a number of proposals for how we might do so.

In the rest of this paper, I first describe the practice of p-hacking. I then offer some circumstantial evidence of its prevalence in accounting research. Finally, I offer some ideas on how to address the "elephant(s) in the room" of accounting research.

2 The Practice of p-hacking

According to Wigglesworth (2021), Campbell Harvey, professor of finance at Duke University, suggests that "at least half of the 400 supposedly market-beating strategies identified in top financial journals over the years are bogus." Harvey (2017) cites research suggesting that 90 % of published studies report the "significant" and "positive" results. Reporting "positive" results is important not only for getting published, but also for attracting citations, which drive behavior for both researchers and journals.

Simmons et al. (2011, p. 1359) provide analyses that "demonstrate how unacceptably easy it is to accumulate (and report) statistically significant evidence for a false hypothesis … [how] flexibility in data collection, analysis, and reporting dramatically increases actual false-positive rates." They attribute this flexibility to **researcher degrees of freedom**: "In the course of collecting and analyzing data, researchers have many decisions to make: Should more data be collected? Should some observations be excluded? Which conditions should be combined and which ones compared? Which control variables should be considered? Should specific measures be combined or transformed or both?" (Simmons et al., 2011, p. 1359).

It is important to note that p-hacking does not require academic misconduct or unethical behaviour. Simmons et al. (2011, p. 1359) suggest that ambiguity in how to make research design choices along with a "researcher's desire to find a statistically significant result" are sufficient conditions for p-hacking to exist.[4]

3 Some researchers agree with the very limited value of accounting research with regard to expanding human knowledge, but argue that the real value of research is in deciding who gets tenure at top universities. But this merely raises the question of the merits of making these decisions based on skills related to conducting and packaging p-hacked research, which seem unclear to say the least.
4 Journal preferences for "significant" and "positive" results could lead to "results" that are effectively p-hacked even if researcher's are not individually seeking "results".

An excessive focus on academic misconduct—such as fraudulent manipulation of data—may in fact be a distraction. Simmons (2018) suggests that "fraud is out there … but it is not very common." Instead, p-hacking is "the main culprit" behind the failure of many studies to replicate (suggesting they are not correct). By focusing on academic misconduct, we risk spending a lot of effort on a smaller problem and missing the elephant in the room.[5] Additionally, there is a risk that academic misconduct and p-hacking get unnecessarily conflated, leading to undue harm to the reputation of researchers flagged as engaging in p-hacking, even though its practice seems widespread.

Bloomfield et al. (2018, p. 317) suggest that "almost all peer-reviewed articles in social science are published under" what they call … the Traditional Editorial Process (or TEP). Under the TEP, "authors gather their data, analyze it, and write and revise their manuscripts repeatedly before sending them to editors." As such authors have access to many researcher degrees of freedom.

An alternative to the TEP is what Bloomfield et al. (2018) call the Registration-based Editorial Process (REP). According to Bloomfield et al. (2018, p. 317), "under REP, authors propose a plan to gather and analyze data to test their predictions. Journals send promising proposals to one or more reviewers and recommend revisions. Authors are given the opportunity to review their proposal in response, often multiple times, before the proposal is either rejected or granted in-principle acceptance … regardless of whether [subsequent] results support their predictions." The REP is intended to eliminate research degrees of freedom and the questionable research practices that these permit.[6]

The *Journal of Accounting Research* (JAR) conducted a trial of the REP for its annual conference held in May 2017. Unfortunately, it is unlikely that the REP will replace the TEP to any great extent in the foreseeable future. The REP is feasible when data are generated in randomized controlled trials (RCTs), as the data simply do not exist when the report is registered.[7] In contrast, most empirical accounting

5 Inevitably and understandably, the burden of proof is much higher for accusations of academic misconduct, so the costs of showing it are much higher.

6 There are two important elements of the REP that affect p-hacking. First, the requirement to specify analytical procedures in advance of having the data is intended to eliminate p-hacking. However, in practice, it can be difficult to specify every detail of data analysis and some researcher degrees of freedom can remain. For example, researchers might choose to conduct and include supplementary analyses if the pre-specified analyses do not yield statistically significant results. Second, a journal will typically commit to publishing the resulting study whether there are statistically significant results or not. This is intended to limit incentives for p-hacking (and also to produce a more faithful research record). However, if authors are concerned about citations of their papers, they may still have incentives to p-hack if there are enough researcher degrees of freedom to do so.

7 See Chambers et al. (2014) for more on the historical antecendents of the REP.

research uses existing archival data, which often makes it impossible to register a report before being able to look at the data.[8]

3 Evidence of p-hacking

3.1 Conversational Evidence

The contention of Ohlson (2025) that raising issues related to p-hacking is taboo in empirical accounting seminars seems very plausible. Outside of papers like Simmons et al. (2011) that aim to demonstrate the "power" of p-hacking, we generally only see circumstantial evidence of p-hacking in the papers themselves.[9] But sometimes (outside of seminars!) researchers can be fairly candid about their research process.

I must have had countless conversations where a colleague or student is examining the effect of X on y_1 and my natural response has been to ask whether some other variables would be the more natural things to examine instead of y_1 and the response is something like "we looked at those other variables and they didn't work". This practice of "reporting only experiments that 'work' " while discarding results that "don't work" is another well-known researcher degree of freedom discussed by Simmons et al. (2011, p. 1364), and is known as the **file-drawer problem** (because experiments that don't "work" are put in a file-drawer).

A more brazen form of p-hacking is trawling through a data set until correlations are found, at which point the challenge is to devise an "interesting" causal story to go with it. I have had conversations suggesting that research for some involves searching for a "significant" correlation and *then* developing a hypothesis to "predict" it. This form of p-hacking is known as **HARKing** (from "Hypothesizing After Results are Known").

To illustrate, consider the spurious correlations website provided by Tyler Vigen.[10] This site lists a number of evidently spurious correlations, such as the 99.26 % correlation between the divorce rate in Maine and margarine consumption or the 99.79 % correlation between US spending on science, space, and technology and suicides by hanging, strangulation and suffocation. The correlations are deemed spurious because normal human beings have strong prior beliefs that there is no underlying causal relation explaining these correlations. Instead, these are regarded as mere coincidence.

---

8 For example, if I am testing a hypothesis using data from CRSP and Compustat, I cannot credibly promise that I have not looked at the data before submitting my report. In principle, one could propose a study that only uses future data from CRSP and Compustat, but this seems unlikely to be popular in a discipline accustomed to hundreds of thousands of observations.

9 I discuss such evidence below.

10 Available at http://tylervigen.com/spurious-correlations.

However, a creative academic can probably craft a story to "predict" any correlation: Perhaps increasing spending on science raises its perceived importance to society. But drawing attention to science only serves to highlight how the US has inevitably declined in relative stature in many fields, including science. While many Americans can carry on notwithstanding this decline, others are less sanguine about it and may go to extreme lengths as a result … This is a clearly silly line of reasoning, but if one added some references to published studies and fancy terminology, it would probably read a lot like the hypothesis development sections of academic papers presented in the empirical accounting seminars discussed by Ohlson (2025).

Sherlock Holmes claims "it is a capital mistake to theorize before you have all the evidence." (Doyle, 2001, p. 27). The modern equivalent might be that "it is a capital mistake to theorize before you have a statistically significant association to 'predict' ", as there seems to be little value in devoting effort to predict a relation that is not supported by the data set you have.

3.2 Evidence From the 2017 JAR REP Trial

The 2017 JAR REP trial itself provides circumstantial evidence of p-hacking in papers produced using the TEP (i.e., almost all papers in accounting research). Bloomfield et al. (2018, p. 326) examine the results reported in the conference papers and conclude that "of the 30 predictions made in the … seven proposals, we count 10 as being supported at $p \leq 0.05$ by at least one of the 134 statistical tests the authors reported." But this is very close to the level of support expected if the null hypotheses for all 30 predictions were true.[11]

This is particularly concerning in that it seems reasonable to expect that the alternative hypotheses considered in the 2017 JAR conference papers were deemed by the authors and reviewers to be worth pursuing before knowing their results, which is a higher bar than applied to hypotheses tested using the TEP. In other words, the results of the 2017 JAR conference raise the uncomfortable prospect that many results produced by the TEP (i.e., almost all research in accounting) arise from p-hacking and are simply false rejections of true null hypotheses.

3.3 Circumstantial Evidence From Replications

Another source of evidence on the prevalence of p-hacking is replications. We expect that p-hacked papers will have results that are very fragile. By definition, p-hacked

11 See Gow and Ding (2023f) for details of the calculation in support of this claim.

results are not expected to be **reproducible**. That is, we would not expect the results to hold if the same analytical procedures were applied to a new data set.[12]

But p-hacked results should pass the test of **replicability**, which requires that the results can be produced by other authors using the same data sets and analytical procedures. Given that much of the published research in accounting uses data sets that are available to most researchers and papers typically include descriptions of the analytical procedures used, other researchers should be able to replicate results independently even without access to the code and data files used by the original authors.

In practice, it seems that many papers cannot be replicated in this way, even approximately. Ask another researcher whether she has tried to replicate results of a published paper and you are likely to hear that attempts have been made, but without success.[13] One explanation for this difficulty is that small departures from the choices made by the authors along the dimensions described in Simmons et al. (2011) can lead to apparent results disappearing (i.e., becoming statistically insignificant) and few papers describe these choices sufficiently clearly to allow precise replication.

I have extensive experience with attempted replication of papers. My ill-fated PhD dissertation attempted to identify a causal mechanism underlying the numerous results in the literature suggesting a contracting value for firms' voluntary adoption of higher levels of conditional conservatism. However, explaining results documented in research is difficult when those results cannot be replicated, and almost all replications I tried failed. Since then I have undertaken many attempted replications in various areas and most have failed.

One explanation might be that I simply do not know how to analyze data properly and that a more skilled researcher would be able to reproduce published results more readily. While it is difficult to rule out this explanation completely, I believe a significant recent project suggests that this is not a complete explanation.

In 2021, a University of Melbourne colleague (Tony Ding) and I started to pull together a course book (Gow & Ding, 2023c) aimed at helping research students to develop the portfolio of skills needed to be good researchers in accounting. As discussed in the book, a core element is material focused on data analysis skills and we have included many replication analyses to support this (Gow & Ding, 2023e). It seems these replication efforts can be organized into two eras.

---

12 Here I follow Hail et al. (2020) in distinguishing replicability from reproducibility. This is clearly related to Elephant #4, which is "Referring to the possibility of using a holdout sample".

13 Some researchers' replication experiences are limited to exercises assigned during PhD coursework, but there is a natural selection bias with these, as many instructors would look to assign exercises where results can be reproduced.

The first era covers 1968 through to about 1996. The striking thing about this era is how robust the results appear to be. Like many before us, we find that the key results of the seminal Ball and Brown (1968) are easily replicated (Gow & Ding, 2023a), and Ball and Brown (2019) show this is true in different markets and periods. We find that key results of Beaver (1968) hold in any year we look at (Gow & Ding, 2023b).[14] Not only can we generate the core results of Bernard and Thomas (1989), but we broadly replicate Foster (1977) along the way (Gow & Ding, 2023h). Replications of Dechow et al. (1995) and Sloan (1996) are also successful.

The second era covers papers from the current century and reveals a different story. Our book provides replications of numerous papers from this era, including Zhang (2007), Fang et al. (2016), Li et al. (2018), and Bloomfield (2021).

The first observation is that we can replicate all of the papers to some degree. But it is important to note that our replications often benefit from access to code and data provided by the authors. Fang et al. (2016) posted code and data starting from original sources and continuing through the production of (some) key results in their paper. Bloomfield (2021) provided code under the journal's data policy.[15]

The main purpose in selecting papers for replication in the book was pedagogical and being able to replicate results was an important criterion for inclusion. In some cases we were not able to replicate papers—and authors were not responsive to requests for assistance—that had been considered for inclusion. If authors do not share their code and data, replication is often difficult. Most of the authors of the papers we replicate went above and beyond the norms of accounting research in sharing their code and data and should receive credit for doing so. It is also important to note that there is no reason to believe that the papers we replicated are in any way unusual in terms of the fragility of their results. Most papers might be similarly fragile, but without the original code and data, there is no cost-effective way to check this.

The second observation is that the results can be very fragile. The results in Fang et al. (2016) on earnings management are robust to some alternative choices (see Fang et al., 2019), but less so to others. For example, the main measure of earnings management used in Fang et al. (2016) is one proposed by Kothari et al. (2005) that matches firms with controls based on performance. But Kothari et al. (2005) use contemporary performance, while Fang et al. (2016) use lagged performance; use contemporary performance and results vanish.[16] Additionally, strong arguments can be made for not using difference-in-difference estimators and for

14 Bamber et al. (2000) raise concerns about the reproducibility of the results in Beaver (1968), but these concerns that do not appear to hold in years after Beaver (1968) was published.

15 We did not have access to code for Zhang (2007), but that paper is unusually straightforward and based on a standard data set (CRSP).

16 See Black et al. (2022) for an extensive analysis of the results of Fang et al. (2016).

using measures of accruals that do not condition on post-treatment outcomes, such as total accruals or even simply income, but all of these changes makes results disappear.[17]

Li et al. (2018) present evidence of firms being less forthcoming with disclosure of customer identities after adoption of the inevitable disclosure doctrine in the states in which they are headquartered. But these results rely on dubious research design choices. As discussed in Gow and Ding (2023g), almost any deviation from these choices causes results to disappear.

Bloomfield (2021, p. 869) claims to "use the IV approach from Iliev (2010)" in implementing a regression discontinuity design (RDD), but actually does not.[18] Replacing the analysis of Bloomfield (2021) with a conventional IV-based RDD analysis causes results to vanish.[19]

In short, none of the papers replicated in our book in the second era is anything but extremely fragile, just as we would expect p-hacked results to be.[20] If other papers where authors do not provide detailed code and data are similarly fragile, then we would not expect to be able to replicate their results, as minor deviations from the data and analytical procedures of the original authors are likely to lead to null results. Thus p-hacking provides an explanation for the difficulty many researchers have in replicating results in published papers.

Combining the evidence above with the concerns raised by Ohlson (2025) and it seems difficult to distinguish much of contemporaneous accounting research from what you would see if p-hacking were the modus operandi of most researchers.

---

17 A post-treatment outcome is a variable observed only after treatment. As such one cannot be sure that it is not affected by the treatment itself or by the outcome of interest. In either case, inferences can be adversely affected [see Rosenbaum 1984]. See Gow and Ding (2023f) for discussion of related design issues.

18 The setting of Iliev (2010) and Bloomfield (2021) is a "fuzzy RDD" setting requiring use of instrumental variable (IV) to obtain consistent estimates of the causal effect of the treatment of interest. Accordingly, Iliev (2010, p. 1179) implements RDD using instrumental variable (IV) regressions including "linear, quadratic, and cubic terms" of the running variable. In contrast, Bloomfield (2021) does not present IV regression results at all, replacing Iliev (2010)'s approach with OLS difference-in-difference regressions using the instrument and firm fixed effects. Iliev (2010, p. 1179) reports first-stage regression results in support of his instrument, while Bloomfield (2021) does not present such analyses. Differences in sample periods mean that the instrument is plausibly weaker in Bloomfield (2021) than in Iliev (2010), and the analysis in Gow and Ding (2023i) suggests that this is indeed true.

19 See Gow and Ding (2023i) for details. It is not clear whether more closely implementing the approach of "the IV approach from Iliev (2010)" would yield statistically significant results in support of Bloomfield (2021)'s hypotheses, but given developments since Iliev (2010) such as Gelman and Imbens (2019), it is not clear that Iliev (2010) is consistent with current standard approaches to RDD.

20 It is important to caveat that I do not claim to have proven that any one of these papers *is* a p-hacked paper. Direct evidence of p-hacking is general impossible to come by and p-hacking is something more easily inferred for an area of research than for a single paper (Ioannidis, 2005).

4 What to Do?

If p-hacking is as prevalent as it seems to be, the natural question is what, if anything, can be done about it. Before addressing this, it is important to note how pernicious p-hacking is to the value of research. If all research is p-hacked, then we should simply ignore research, as p-hacking does not produce information of value other that insights into the p-hacking skills of the authors.[21]

Some researchers appear to recognize the prevalence of p-hacking, yet remain sanguine about the research enterprise. For example, one senior researcher broadly agrees with my assessment about p-hacking, but argues "there are some solid researchers doing some interesting papers." But it is important to understand that if, say, 80 % of research is p-hacked, that one cannot simply read the 20 % that is not p-hacked and ignore the rest. If it were easy to detect the p-hacked papers, we could simply avoid publishing them.

That said, I argue there are steps that could be taken to reduce p-hacking to an extent that research in aggregate might again have some value. In this section, I discuss four ideas for addressing concerns about p-hacking.

4.1 Reject Papers That Ask Silly Questions

Accounting academics appear to adore "novelty", where novelty often means asking questions that no-one has even dreamed of asking before. This is problematic for two reasons. First, if questions are so novel that no-one has asked them, how can they be important? Second, the ability to simply make up "interesting" research questions is a p-hacker's dream.[22]

Too often the bar seems to be "has someone [in prior research] asked this question before?" and if the answer is "no" then the novelty bar has been cleared. But, after more than 50 years of modern empirical accounting research, the fact that no-one has asked a question should in most cases be a strike *against* a paper, not for it. If no-one has addressed the question, then it is perhaps because no-one cares what the answer is. If editors adopted a policy of desk-rejecting papers that ask silly questions, the pay-off to p-hacking would decline significantly.

Of course, one reason for researchers needing to seek out smaller and smaller questions is simply the weight of prior research. Going first, researchers such as

21 Of course, information about the p-hacking skills of the authors is arguably relevant if the ability to produce published papers is the sole research-related criterion for evaluating a researcher, as it is at many institutions.

22 Think "What is effect of the decline in US science on suicide rates?" as a paper title based on the correlation from Tyler Vigen's website that I discussed above.

Ball and Brown (1968) and Beaver (1968) could pick the "low-hanging fruit" of more fundamental questions of the discipline, and test alternative hypotheses that were more likely to be true and hence easier to demonstrate empirically. Later researchers were left to explore the questions that remain. As such, it is perhaps "unfair" to compare research in the period from, say, 2000 to today with that in the period 1968–1999.

But this argument would suggest that, as research progresses, we should be seeing more and more papers with null results, either because the alternative hypotheses tested are less likely to be true or because the empirical challenges faced in demonstrating them are greater. This seems inconsistent with the reality that almost all published papers have "results", either ruling out the unfairness or suggesting that researchers compensate for the smallness of the (apparent) phenomena they study by simply looking harder (i.e., p-hacking).

4.2 Increase Emphasis on Replication

We saw above that being able to replicate papers such as Fang et al. (2016) and Li et al. (2018) makes it easy to see just how fragile their results are. If it were easy to replicate papers, then the incentives for p-hacking might be dramatically reduced, as it would be easy to raise doubts about papers with the very fragile results that p-hacking usually produces. One would hope that papers shown to be extremely fragile would be regarded as less reliable, and thus less likely to cited or to be evaluated positively by peers after publication, thereby reducing incentives for their production.[23] Unfortunately, the salutary effects that replications can have on incentives for p-hacking are much diminished for a number of reasons.

First, replication is a costly exercise. Most empirical researchers already spend a large portion of their research time in the critical pre-tenure phase of their careers writing code to analyze data. Independently replicating others' papers is likely to be considered a poor use of very limited time. And, as discussed above, a typical replication conducted without some cooperation from the original authors is likely to yield differences in results that are difficult to explain, requiring exhaustive checks and iterations to understand them.

While one solution to this issue is for authors to supply the data and code needed to replicate their results, very few authors do so. Authors have essentially no incentive to provide data and code voluntarily. Once a paper has been published, there is really only downside from sharing code and data for an author

23 Note that this reduction in incentives for producing p-hacked papers would be much reduced for researchers whose incentives focus on the number of papers produced, whether those papers are cited or regarded highly. Such incentives are created by many institutions around the world, including my current one.

focused on publishing papers, as the results might be due to coding errors or fragile. Thus authors have a natural incentive to make replication difficult.

On top of this lack of positive incentives is the reality that most researchers appear to be poor in organizing their code and that significant costs would need to be incurred to prepare code and data for sharing.[24]

In the absence of incentives for voluntary disclosure of code, some kind of requirement for sharing seems necessary. However, only one of the top three accounting journals (*Journal of Accounting Research*) imposes requirements for data and code, but even then these requirements rarely yield files that permit easy replication of tables found in papers.[25]

There are two steps that journals could take to enhance the credibility of results. First, journals could step up the data and code requirements for published papers. While the *Journal of Accounting Research* is a clear leader in this regard among top accounting-focused journals, there is plenty of room for improvement. Journals in other disciplines have gone further and there is some hope that these could influence best practices in accounting research. For example, *Management Science*, a broader management journal that has an accounting department, has a much more rigorous policy for code and data disclosure policy than any of the specialist accounting journals (Management Science, 2019). Under the *Management Science* policy, "all papers using code or data … must provide replication materials which need to be approved by the Code and Data Editor" (see Simchi-Levi, 2023). The policies adopted by *Management Science* are closely modeled on those of the American Economic Association and the *Journal of Finance*.

Second, journals could publish replications of papers when these provide insights on the questions in the original papers. For example, Guest (2021) identified "six discrepancies in … reporting, coding, and data" in replicating a previously published paper. The *Journal of Finance* published the replication and retracted the original paper. Because journals generally do not publish replications or corrections, there is essentially no incentive to produce these in a world where counting published papers is a dominant (often the only) research-related performance measure. More recently, the *American Economic Review* published a comment on a paper that was then retracted (Bach et al., 2023).

4.3 Decrease Emphasis on "Identification Strategies"

It is widely understood that accounting research has become increasingly concerned about "identification strategies" in recent years. Identification strategies—to use the

24 This is likely to be especially true when results are derived from the often messy process of p-hacking.
25 The *Journal of Accounting and Economics* merely "encourages" authors to share replication files. And there is nothing on this issue in the editorial policy of *The Accounting Review*, which does not appear to provide any support for such sharing.

term in common use—seek to enhance the credibility of causal inferences in empirical research by exploiting features of the research setting and purportedly appropriate statistical techniques.[26] By focusing on identification strategies, accounting research may have reduced its immunity to p-hacking. The apparent obsession with papers with "clever" identification strategies seems to have led to a new kind of p-hacking in which a researcher starts with the identification strategy (often drawn from finance and economics) and then seeks statistically significant results using outcomes popular in accounting research, such as earnings management or voluntary disclosure.

The extremely fragile results of Fang et al. (2016) and Li et al. (2018) seem to be plausible candidate illustrations of this phenomenon. Fang et al. (2016) exploits the random assignment of elimination of short-selling restrictions, but so do 60 other papers exploring "indirect effects" of these restrictions in a setting where little or no evidence of direct effects was found.

Apart from the potential inducement of p-hacking, concerns about the obsession with identification strategies in accounting research are increased when one considers the credibility of these strategies in practice. Many papers simply use difference-in-difference regressions—perhaps including "fixed effects"—which rely on the "assume a can-opener" assumption of "parallel trends".[27] Papers use instrumental variables, even though it is doubtful that any valid instruments exist in accounting research.[28] Papers in accounting research that claim to use RDD generally do not.[29]

4.4 Incorporate Discussion of p-hacking Into Research Training

One hopes that accounting research training has not "evolved" to the point that PhD students are being taught how to do p-hacking. Instead, students learn about p-hacking "on the job" in a sense. Understanding the importance of "results", students learn to exercise researcher degrees of freedom in ways that eventually yield the "stars" denoting "statistically significant" coefficients. Given these incentives, it

26 For example, researchers might seek to use complex fixed-effect structures, natural experiments, instrumental variables, or RDD.

27 This assumption maintains that, in the absence of treatment, the difference in outcome between treatment and control observations is constant over time. I label this an "assume a can-opener" assumption because its justification comes from the benefits of making it for causal inference and not at all from any underlying economic rationale for it. See Gow and Ding (2023f) for discussion of the general implausibility of the "parallel trends" assumption.

28 See Gow and Ding (2023d) for more on this point.

29 See Gow and Ding (2023i) for a recent survey of the use of RDD in accounting research.

seems important that the problems with p-hacking are addressed more forthrightly in training PhD students.

In this regard, the recent explosion of high-quality material on research methods is somewhat disappointing. Recent years have seen the emergence of high-quality resources for students looking to understand causal inference using observational data, including Angrist and Pischke (2008, 2014), Cunningham (2021), and Huntington-Klein (2021). While these are excellent resources for helping researchers to understand subtle issues not explicitly addressed by more traditional texts, none of them even touches on the topic of p-hacking. We offer an initial attempt to incorporate this topic into a PhD curriculum in our course book and hope that others find ways to build content on this topic into their PhD curricula so as to raise awareness of these issues (Gow & Ding, 2023g).

Beyond helping aspiring researchers to understand the issues with p-hacking, there are a number of complementary skills that could benefit from more attention in research training, especially skills related to management of code and data and research collaboration.

One opportunity would be educating students on approaches to retention of data and code. For example, as "the authors were unable to provide the original data and code requested by the publisher that reproduce [their] findings" of the paper, Bird and Karolyi (2019) was retracted by *The Accounting Review*. Loss of data and code can be addressed in a number of such ways, including the use of cloud services such as Dropbox or Google Drive.

Another issue is coding errors. Simchi-Levi (2023) notes that "errors in code" were discovered in 22 % of the replication packages submitted to *Management Science*. This likely represents an under-estimate given the limited scope of the replication effort and need to maintain quick turnaround times.[30] Yet it is also perhaps unsurprising. As discussed in Gow and Ding (2023e), most doctoral programs provide very little training related to coding practices despite the importance of such skills for modern empirical researchers. Incorporation of better training on coding practices may help to reduce coding errors, thereby yielding research that is both more correct and easier to replicate.[31]

30 There is no breakdown of these statistics by department, so it is not clear whether this is higher or lower with accounting papers.

31 That the research community is relatively understanding about coding errors perhaps explains efforts by researchers to attribute issues in their papers to "coding errors". For example, the code repository for Boissel and Matray (2022) included the line replace B = B/1.8 if t > −3 & t < 0. This code modifies two coefficients in a way enhances a plot used to support a claim of "parallel trends". While the responding author attributes this to a "coding error" it is difficult to imagine what the correct version of this line of code would be. The "coding error" in Bao et al. (2020) differs from that in Boissel and Matray (2022) in a number of respects. First, no code containing the claimed error was provided.

A modern empirical research paper often resembles a software development project, with small collaborative teams, a need for version control, and the demonstrated potential for coding errors. As such, accounting researchers might benefit from drawing on tools and approaches used by software developers. In some cases coding errors have been made by one author whose code has not been seen by others. In other cases, the source of the error is difficult to identify (a "coding error" can have different valence if made by a research assistant rather than a motivated co-author). Collaborative tools such as Git can be adapted to social science research. With a shared Git repository it becomes easier to do code reviews and to ensure that data has not been manipulated.[32] There appears to be some irony in a research field that often examines control systems itself having control systems without audit trails, separation of duties, or robust review processes.

4.5 Encourage More Descriptive Research

Gow et al. (2016, p. 499) point out that "there are very few studies published in top accounting journals that focus on providing detailed descriptions of institutions in accounting research settings" and document that the vast majority of empirical accounting papers focus on providing causal inference, notwithstanding all the difficulties widely understood to be faced by that endeavor. This focus on causal inference is accompanied by a desire to find "positive" results, which provides the incentives for p-hacking I have discussed.

In arguing that "accounting research can benefit substantially from more in-depth descriptive research", Gow et al. (2016, p. 499) suggest that "this type of research is essential to improve our understanding of causal mechanisms and [to] develop structural models." An additional benefit of such descriptive research is that it would enhance the understanding of the research community of how the real world works, making it more difficult to pass off p-hacked results that are not consistent with actual business practices.

Second, producing code with this issue accidentally seems even less plausible than with the line above from Boissel and Matray (2022; see Gow, 2022 for an attempt to replicate the "coding error"). Third, the authors' efforts to attribute the issue to a "coding error" is belied by earlier efforts to suggest it was an appropriate research design choice (Bao et al., 2021).

32 For example, raw data files from MTurk might be committed to a repository by an RA before any analysis is undertaken.

5 Concluding Comments

Ohlson (2025) draws on his experience in empirical accounting seminars to identify five "elephants in the room". I interpret each of these elephants as either a variant or a symptom of p-hacking. I provide evidence of the prevalence of p-hacking in accounting research that complements the observations made by Ohlson (2025).

While I identify a number of steps that could be taken to reduce p-hacking in accounting research, I conjecture that facilitating and encouraging replication alone could have profound effects on the quality and quantity of empirical accounting research.

6 Appendix: The Other Elephants

Above I explained that the basic Elephant #5 of Ohlson (2025) ("Issues Related to 'Screen-Picking' and 'Data-Snooping' ") is synonymous with p-hacking. For completeness, I close this paper with a brief discussion of how the other four elephants also reflect concerns with p-hacking.

Elephant #1 ("Referring to the Absence of a Fama-MacBeth Analysis") is likely to be seen when researchers are reluctant to adjust their standard errors in ways that make results disappear (see Gow et al., 2010 for discussion of approaches to calculating standard errors in accounting research).

Elephant #2 ("Asking whether a Key Right-Hand-Side (RHS) Variable Contributes to Explaining the Dependent Variable") and #3 ("It Takes More than Stars to Settle the Matter") both raise uncomfortable questions when the results are p-hacked, as the explanatory value of the independent variable is likely to be low and the economic significance of any apparent relation is likely to be small when the sample is large.

Elephant #4 ("Referring to the Possibility of Using a Holdout Sample") is also an awkward idea when a paper is based on p-hacked results, as we do not expect those results to hold in a new sample, pretty much by definition. Here Ohlson (2025) is effectively discussing the idea of replication by the authors of the paper themselves.

References

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Angrist, J. D., & Pischke, J.-S. (2014). *Mastering 'metrics: The path from cause to effect*. Princeton University Press.

Bach, L., Bozio, A., Guillouzouic, A., & Malgouyres, C. (2023). *Dividend taxes and the allocation of capital: Comment*. American Economic Review Forthcoming.

Ball, R., & Brown, P. (1968). An empirical evaluation of accounting income numbers. *Journal of Accounting Research*, *6*, 159–178.

Ball, R., & Brown, P. (2019). Ball and Brown (1968) after fifty years. *Pacific-Basin Finance Journal*, *53*, 410–431.

Bamber, L. S., Christensen, T. E., & Gaver, K. M. (2000). Do we really "know" what we think we know? A case study of seminal research and its subsequent overgeneralization. *Accounting, Organizations and Society*, *25*, 103–129.

Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2020). Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach. *Journal of Accounting Research*, *58*, 199–235.

Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2021). A response to "critique of an article on machine learning in the detection of accounting fraud". *Econ Journal Watch*, *18*, 71–78.

Beaver, W. H. (1968). The information content of annual earnings announcements. *Journal of Accounting Research*, *6*, 67–92.

Bernard, V. L., & Thomas, J. K. (1989). Post-earnings-announcement drift: Delayed price response or risk premium? *Journal of Accounting Research*, *27*, 1–36.

Bird, A., & Karolyi, S. A. (2019). Retraction: Governance and taxes: Evidence from regression discontinuity. *The Accounting Review*. https://doi.org/10.2308/1558-7967-92.1.000

Black, B. S., Desai, H., Litvak, K., Yoo, W., & Yu, J. J. (2022). *The SEC's short-sale experiment: Evidence on causal channels and on the importance of specification choice in randomized and natural experiments*. SSRN. https://doi.org/10.2139/ssrn.3657196.

Bloomfield, M. J. (2021). The asymmetric effect of reporting flexibility on priced risk. *Journal of Accounting Research*, *59*, 867–910.

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.

Boissel, C., & Matray, A. (2022). Dividend taxes and the allocation of capital. *American Economic Review*, *112*, 2884–2920.

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. J. (2014). Instead of "playing the game" it is time to change the rules: Registered reports at AIMS neuroscience and beyond. *AIMS Neuroscience*, *1*, 4–17.

Cunningham, S. (2021). *Causal inference: The mixtape*. Yale University Press.

Dechow, P. M., Sloan, R. G., & Sweeney, A. P. (1995). Detecting earnings management. *The Accounting Review*, *70*, 193–225.

Doyle, A. C. (2001). *A study in scarlet, classics series*. Penguin Publishing Group.

Fang, V. W., Huang, A. H., & Karpoff, J. M. (2016). Short selling and earnings management: A controlled experiment. *The Journal of Finance*, *71*, 1251–1294.

Fang, V. W., Huang, A., & Karpoff, J. M. (2019). *Reply to 'the Reg SHO reanalysis project: Reconsidering Fang, Huang and Karpoff (2016) on Reg SHO and earnings management' by Black et al. (2019)*. SSRN. https://doi.org/10.2139/ssrn.3507033.

Foster, G. (1977). Quarterly accounting data: Time-series properties and predictive-ability results. *The Accounting Review*, *52*, 1–21.

Gelman, A., & Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, *37*, 447–456.

Gow, I. D. (2022). Should Bao et al. (2020) be retracted? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4246151

Gow, I. D., & Ding, T. T. (2023a). *Ball and Brown (1968), in: Empirical research in accounting: Tools and methods*.

Gow, I. D., & Ding, T. T. (2023b). *Beaver (1968), in: Empirical research in accounting: Tools and methods*.

Gow, I. D., & Ding, T. T. (2023c). *Empirical research in accounting: Tools and methods*.

Gow, I. D., & Ding, T. T. (2023d). *Instrumental variables, in: Empirical research in accounting: Tools and methods*.
Gow, I. D., & Ding, T. T. (2023e). *Introduction, in: Empirical research in accounting: Tools and methods*.
Gow, I. D., & Ding, T. T. (2023f). *Natural experiments revisited, in: Empirical research in accounting: Tools and methods*.
Gow, I. D., & Ding, T. T. (2023g). *Panel data, in: Empirical research in accounting: Tools and methods*.
Gow, I. D., & Ding, T. T. (2023h). *Post-earnings announcement drift, in: Empirical research in accounting: Tools and methods*.
Gow, I. D., & Ding, T. T. (2023i). *Regression discontinuity designs, in: Empirical research in accounting: Tools and methods*.
Gow, I. D., Larcker, D. F., & Reiss, P. C. (2016). Causal inference in accounting research. *Journal of Accounting Research*, *54*, 477–523.
Gow, I. D., Ormazabal, G., & Taylor, D. J. (2010). Correcting for cross-sectional and time-series dependence in accounting research. *The Accounting Review*, *85*, 483–512.
Guest, P. M. (2021). Risk management in financial institutions: A replication. *The Journal of Finance*, *76*, 2689–2707.
Hail, L., Lang, M., & Leuz, C. (2020). Reproducibility in accounting research: Views of the research community. *Journal of Accounting Research*, *58*, 519–543.
Harvey, C. R. (2017). Presidential address: The scientific outlook in financial economics. *The Journal of Finance*, *72*, 1399–1440.
Huntington-Klein, N. (2021). *The effect: An introduction to research design and causality*. CRC Press.
Iliev, P. (2010). The effect of SOX Section 404: Costs, earnings quality, and stock prices. *The Journal of Finance*, *65*, 1163–1196.
Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124.
Kothari, S. P., Leone, A. J., & Wasley, C. E. (2005). Performance matched discretionary accrual measures. *Journal of Accounting and Economics*, *39*, 163–197.
Li, Y., Lin, Y., & Zhang, L. (2018). Trade secrets law and corporate disclosure: Causal evidence on the proprietary cost hypothesis. *Journal of Accounting Research*, *56*, 265–308.
Management Science. (2019). *Policy for data and code disclosure*.
Ohlson, J. A. (2025). Empirical accounting seminars: Elephants in the room. *Accounting, Economics, and Law: A Convivium 15*, 1–8.
Rosenbaum, P. R. (1984). The consquences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society. Series A (General)*, *147*, 656.
Simchi-Levi, D. (2023). From the editor. *Managemenet Science*.
Simmons, J. P. (2018). *Life after p-hacking*.
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.
Sloan, R. G. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? *The Accounting Review*, *71*, 289–315.
Wigglesworth, R. (2021). *The hidden "replication crisis" of finance*. Financial Times.
Wikipedia (2023). *Blind men and an elephant*.
Zhang, I. X. (2007). Economic consequences of the Sarbanes-Oxley act of 2002. *Journal of Accounting and Economics*, *44*, 74–115.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/ael-2022-0111).