The elephant in the room: Response to reviewers

Ian D. Gow*

25 May 2023

1 Reviewer 1

The manuscript comments on Ohlson's "Elephant" critique. It provides a useful recharacterization of the various points made in the critique in terms of an overarching concern about p-hacking; and offers possible solutions to the problem.

I find the comment interesting and useful. It provides a relevant perspective. I find myself agreeing with many of the points made in the comment.

I only have one minor comment / suggestion which I offer for perusal: The comment suggests to "decrease the emphasis on identification strategies." I personally view identification strategies or design-based approaches to inference as, in principle, useful ways to constrain researchers' degrees of freedom (and, hence, p-hacking). If properly used, design-based approaches provide researchers with a clearer understanding of what specifications are, ex ante, superior than others, given the institutional details of the setting at hand and the relevant theoretical insights (see, e.g., the discussion of design-based research of Leuz, 2022). This understanding limits the number of "possible" specifications that researchers can run and cherry-pick results from to support their predictions.

I understand (and agree) that, in practice, the emphasis on identification strategies has, at times, led to undesirable outcomes. Those outcomes result from researchers focusing on statistical tools (e.g., difference-in-differences tests with parallel trend assumption) instead of institutional features (and setting-specific theory); and researchers searching for "shocks" instead of interesting questions or institutions. (The recycling of shocks is a result of the latter which comes with various undesirable aspects. One of the undesirable aspects is that the use of shocks that have been shown to affect the economics and financing of firms can hardly be useful to isolate exogenous variation in firms' accounting-related incentives (holding all else fixed); i.e., the one-channel assumption is, by design, violated.)

*University of Melbourne, ian.gow@unimelb.edu.au

It would seem useful to acknowledge the above nuance. The issue with a focus on identification seems to be rooted more in our current practical application / use of identification strategies rather than the lacking conceptual merits of design-based accounting research. If properly applied, design-based empirical research is one possible remedy to p-hacking; not its cause.

I see conceptual merit in this comment. But I think a good analogy is the "selective attention test". By focusing on "addressing endogeneity" (counting passes by players in white), we're neglecting the gorilla walking into the scene (p-hacking).

The phrases "properly applied", "in principle", and "properly used" are doing a lot of heaving lifting in this comment. I think "design-based strategies" fall into two classes.

The first class involves statistical techniques that *when coupled with settings with particular properties* allow researchers to obtain unbiased estimates of causal effects.[1]
These include regression discontinuity designs (RDD) and instrumental variable regression (IV). But (as discussed in the manuscript) the reality is that these techniques are definitely not properly applied in accounting research (RDD is generally not used with much care at all; I cannot think of a valid use of IV in accounting research). These approaches make strong demands on the settings in which they are used, and convincing readers that these demands are met relies on deep understanding of "institutional features". Yet this seems rarely to be a constraint in practice.

The second class of approaches are those that do not address endogeneity so much as assume that it can be ignored. In this class I would put "design-based strategies" such as difference-in-differences and fixed-effect regressions. In some ways, these methods (mostly just dressed up OLS) make even more stringent demands on their settings, but seem to applied with even less care than RDD and IV.

I think fully addressing the idea of "design-based approaches" as a solution would be a paper in its own right (I am thinking to write a detailed response to Leuz, 2022). My view is that in practice the poor research training of most accounting researchers mean that these end up being mostly "Jedi mind tricks" used by "clever" researchers to get a larger share of top journal space, rather than a real solution to the problems of p-hacking.

Even when one has found a special setting where one can justifiably claim to have a good "identification strategy" and one is not simply jumping on a bandwagon of a popular "shock", I think that extant "theories" provide more than enough researcher degrees of freedom to find results (and find results one must to get published).

All the above aside, I have gone over the paper with this comment in mind and tweaked as appropriate.

[1] There are a lot of details suppressed in this characterization, including asymptotic consistency and local average treatment effects, but these are not important for my point.

2 Reviewer 2

The paper provides evidence of the prevalence of p-hacking and provides proposals for how the problem could be tackled. It thereby nicely complements Ohlson's (2022) observations and is therefore a useful contribution to the literature.

Specific comments

1. On p. 3 it would help to explain why archival data makes it impossible to register a report before being able to look at the data. The reason has to do with the fact that other published studies would have used the same data, and the researcher cannot forget the observations/findings from their own previous research. For example, the researcher might know from own experience or published research whether directional accruals or unsigned accruals (i.e., their absolute value) "work well." Second, when multiple people register their reports for data analysis on a similar topic, one might very well click—the data are correlated, while in RCTs they are by definition independent. Third, it is much harder to detect whether the researcher peeked at the data before registering the report.

I suspect it is the last one that is more relevant. Most research questions in accounting nowadays are "one and done": there is demand for one paper on a question, but no more. Knowing what "works" is less of an issue than it might be.

I added brief footnote elaborating on this issue.

2. P. 4 – it may be useful to point out that archival research is not the only kind to suffer from the problem of researchers being able to craft a story to "predict" any correlation. It happens with RCTs too–see https://rolfzwaan.blogspot.com/2013/09/how-to-cook-up-your-own-social-priming.html for an entertaining description of the process.

Absolutely. Simmons et al. (2011) demonstrate this nicely (and that blog describes a pretty similar process).

3. On p. 6 line 29, please clarify whether you mean "even if authors do make code and data available" or "if authors do not make code and data available."

I have added the missing "not" to this sentence (and revised a little bit for clarity).

4. One of the methods proposed to tackle the p-hacking problem is the use of the "specification curve" (Simonsohn et al., 2020) and it has been used in Berchicci and King (2022). It would be useful to comment on this method.

It seems that this method is applied in those two papers by people other than the authors of the original studies to demonstrate something about the robustness (or lack thereof) of those studies' results.

When done by the original authors, this is not very different from the "robustness checks" that reviewers often request and that authors no doubt apply with some discretion.

When done by other authors, the discussion of replication code and data becomes relevant. Berchicci and King (2022) appear to have encountered non-trivial challenges in replicating Khan et al. (2016).

5. On Point 4.1 rejecting papers that ask silly questions may not be as easy as it seems. Ohlson (1995) faced a lot of problem in publication, possibly because it was considered silly to say that stock prices would react to earnings levels rather than earnings changes.

I agree that this suggestion requires editors to exercise judgment about what is an important research question and there is little evidence that they are willing and able to do so. But it is difficult to see how research can proceed if the discipline has some sense of what are sensible directions for research.

6. On the same point, it was hard for me to make out what you are driving at in the two paragraphs on lines 23-38 of p. 8. It is quite possible that because the low-hanging fruit has already been picked, only the smaller problems remain. Since journals do not publish null results, the ones that are significant are likely to have been p-hacked—this is essentially what Ioannidis (2005) argues. Is it your point that journals should publish null results? On this issue, Kallapur's (2022, 81–82) suggestion is to increase the signal to noise ratio in the research setting, e.g., by limiting the sample to a specific industry as in Guo et al. (2004). His argument (Kallapur 2022, 80–81) is that p-hacking is like an option that is more valuable when variance is high, that is, the signal to noise ratio is low.

In an ideal world, publication of a well-executed study would not be a function of whether it found "results" or not. So ideally journals would publish null results. Of course, we are far from that ideal world.

I was not able to locate Kallapur (2022), but I see merit in that argument. That said, for many hypotheses it wil be unclear where the signal-to-noise ratio will be higher. Given the opportunity to publish studies focused on one industry, many researchers would be inclined to choose the one with strongest "results" ex post.

7. In Point 4.2, you seem to be recognizing both sides—the arguments against code disclosure are the points that it is difficult to prepare code for disclosure (which in my mind is a lesser problem) and more importantly, not mentioned, it would make it too easy to "question" the results and the original authors would have to spend an inordinate amount of time defending themselves. The argument in favor of code disclosure is of course that it facilitates replication. It would help if you could explain which side you are on, and why. And can you justify your assertion (line 29 of p. 9) that the requirements for data and code rarely yield files that permit easy replication? I am familiar with the Journal of Money Credit and Banking project (Dewald, Thursby, and Anderson 1986) but not with more recent evidence and anything specific to accounting.

8. On p. 10, line 29, it should be "difference-in-difference" instead of "difference-indifference."

Yes. "Difference-indifference" may have been a Freudian slip.

9. On the issue of incorporating discussions of p-hacking in research training, I agree and applaud the effort. My experience, however, is that the incentives for p-hacking are just too strong. The publication *of Guest (2021) and the retraction of the original paper helps. However, I understand that Black et al. (2022) are having a hard time publishing their paper. Anyway, this is just an aside.*

Absolutely. I think it highly unlikely that accounting research recovers from its descent into p-hacking. Note that three of my four suggestions can only really be implemented in the editorial process and if implemented (a huge "if") would lead to a change in incentives.

10. While I agree with encouraging more descriptive research, and Roberts and Whited (2013) make a similar point in their concluding paragraph, it would help to explain what kind of descriptive research increases our understanding, and provide examples.

I now direct readers to Gow et al. (2016) for further discussion of this issue and examples.

References

Berchicci, L., King, A., 2022. Corporate sustainability: A model uncertainty analysis of materiality. Journal of Financial Reporting. https://doi.org/10.2308/JFR-2021-022

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. https://doi.org/10.2139/ssrn.3657196

Gow, I.D., Larcker, D.F., Reiss, P.C., 2016. Causal inference in accounting research. Journal of Accounting Research 54, 477–523.

Guest, P.M., 2021. Risk management in financial institutions: A replication. The Journal of Finance 76, 2689–2707. https://doi.org/10.1111/jofi.13063

Guo, R.-J., Lev, B., Zhou, N., 2004. Competitive costs of disclosure by biotech IPOs. Journal of Accounting Research 42, 319–355. https://doi.org/10.1111/j.1475-679X.2004.00140.x

Khan, M., Serafeim, G., Yoon, A., 2016. Corporate sustainability: First evidence on materiality. The Accounting Review 91, 1697–1724. https://doi.org/10.2308/accr-51383

Leuz, C., 2022. Towards a design-based approach to accounting research. https://doi.org/10.2139/ssrn.4216885

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. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Simmons, J.P., Nelson, L.D., 2020. Specification curve analysis. Nature Human Behaviour 4, 1208–1214. https://doi.org/10.1038/s41562-020-0912-z