Matthias Breuer*

Another Way Forward: Comments on Ohlson's Critique of Empirical Accounting Research

https://doi.org/10.1515/ael-2022-0093
Received November 15, 2022; accepted July 24, 2023; published online August 8, 2023

Abstract: Ohlson (2025. Empirical accounting seminars: Elephants in the room. *Accounting, Economics, and Law: A Convivium* 15, 1–8) laments that the evidentiary quality of empirical accounting research is low due to inappropriate methods and practices, leaving seminar attendees and readers unpersuaded by presented or published articles. He suggests that the norms of the profession prevent a public recognition and discussion of those issues, thereby sustaining the poor state of empirical accounting research. I agree that some current empirical approaches and norms seem to hamper progress toward more convincing research. I provide a practical suggestion to possibly improve the state of empirical accounting research. I caution though that even with better methods and more honest research practices, we should not expect that any individual research article can provide conclusive answers to important accounting questions. Such progress in knowledge requires a body of high-quality and independent research.

Keywords: accounting research; Bayesian inference; hypothesis development

JEL Classification: M4

Table of Contents

***Corresponding author: Matthias Breuer**, Associate Professor, Columbia University, 665 W 130th St, New York, NY 10027, USA, E-mail: mb4468@gsb.columbia.edu. https://orcid.org/0000-0002-1754-6865

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 Summary of the Elephants Critique

Ohlson (2025) (titled "Empirical Accounting Seminars: Elephants in the Room") suggests that empirical research articles in the field of accounting are often "non-persuasive" because they rely on questionable research practices. Some of those practices are exemplified in the article. They relate to issues with researchers' degrees of freedom (e.g., regarding specification choice) and improper inferences (e.g., focus on *p*-values). Similar and further issues have been extensively discussed in prior work of Ohlson (2022) and others, in accounting and elsewhere (e.g., Bloomfield et al., 2018; Gelman & Loken, 2014; Harvey, 2017; Johnstone, 2021; Martinson et al., 2005; Simmons et al., 2011).

Ohlson (2025) argues that the questionable practices are sustained, in part, because we do not question them in public. The argument goes as follows: We are all (or almost all) aware of the shortcomings of inferential methods and our approach to empirical work, thanks to our own research practices and experiences. Accordingly, we collectively coordinate not to raise questions about those practices in seminars, as those questions would not only offend the presenter but also implicate everyone else in the room including the questioner. After all, who should throw the first stone? This coordination can help sustain questionable practices by avoiding public recognition and discussion of the problems. As a result, the collective quality of empirical research and trust in such research could be hurt.

Ohlson (2025) suggests that more critical discussions, in public (e.g., seminars and conferences), could improve research practices and trust in empirical research

articles. To incentivize such questioning, the article argues for a change in incentive systems. Reducing the high-powered publication incentives could be one way to reduce the tacit collusion possibly practiced in seminar rooms.

2 My Take on Elephants in Accounting Seminars

I am sympathetic to the assertions that empirical (accounting) research sometimes relies on questionable practices, that those practices are not always challenged in seminar rooms, and that the broader incentive system that governs our profession may partially explain this situation.[1]

I think critical questioning in seminars can be very useful. I do not think that one needs to be "free of sin" to ask questions. Likewise, the questions should not aim at exposing the presenter as a "sinner" (to stay with the biblical analogy). The questions can be critical, but should solely focus on the idea being presented, not the person presenting it. Then, seminars are a great institution that embodies the idea that research is a collective endeavor. Notably, our role in academia is not confined to creating our own research and published articles. To a large part, our role encompasses vetting and improving other people's research. This broader role of academics reflects the ultimate purpose to collectively come to more informed ways of viewing the world; in our case: to better understand accounting and its institutions.

I am not sure that more critical questioning in seminars (alone) would greatly improve the quality of our empirical research. Notably, the lack of critical questioning in public does not mean that questionable practices need to remain unchallenged. The review process, for example, can help weed out questionable practices by allowing private (confidential) questioning. In this process, the questioner does not need to fear personal costs from raising questions. Despite this private critiquing option, questionable practices appear to persist though. Hence, broader reasons for the persistence of questionable practices and other remedies than "more questions in seminars" may need to be explored.

I submit that the quality of empirical research is likely affected by both the particular (educational) challenges faced by an applied field like accounting research and resulting norms that limit both public *and* private questioning of our research

1 I do not fully agree with the characterization of some of the statistical issues raised in Ohlson (2025). Ohlson (2025), for example, points out that, due to high power afforded by large sample sizes (N), many statistically significant results (with comparably small t-statistics) are not economically important. To detect those cases, Ohlson (2025) advocates for using the "t over square root of N" heuristic or, similarly, reporting the incremental R^2 for the variable of interest. I agree that assessing the economic magnitude, not just statistical significance, is important. I also agree that the proposed heuristics can be useful. I caution though that those heuristics do not always provide the most relevant benchmark and can, in some cases, be misleading. I discuss the limitations of the proposed heuristics in an Annex (5) to this comment.

practices. To conduct empirical research in an applied field such as accounting research, we need to be trained in both accounting institutions and research methods, including theoretical modelling and empirical testing. This is a tall order. Accordingly, we tend to know a bit of everything but, understandably, must often rely on methods that are deemed acceptable instead of those that, derived from first principles, are the most appropriate methods for our research. What is and is not acceptable is typically illustrated in and passed on through seminar courses in our PhD programs, where prior published work is taken as a role model (e.g., in terms of paper structure, empirical tests, etc.).[2] The resulting norms act as heuristics, helping us to manage the educational challenges inherent in applied empirical accounting research. At the same time, they contribute to a sluggish response to new developments, resulting in the short- or even medium-run survival of questionable practices, in line with the broader argument in Ohlson (2025). In a small field like accounting, such norms can be easily sustained through coordination (e.g., "this is the way *we* do it!"). The positive aspect about this easy coordination in accounting, however, is that we can also hope to transition to a better equilibrium through coordination; especially if the gatekeepers (e.g., editors) coordinate on new norms and practices.[3]

In the next section, I provide a practical suggestion for new practices which entail abandoning the field's norm to (almost) exclusively focus on hypothesis testing. This suggestion would seem helpful in addressing many of the issues raised in Ohlson (2025). At the same time, it might be simpler to implement than other remedies (e.g., changing the incentive structure of academia) proposed in Ohlson (2025). I detail my suggestion below.

3 My Suggestion for Progress

My suggestion for progress in empirical accounting research is to abandon hypothesis testing as a norm. To be clear, this is not a new idea. It is one proposed by prominent statisticians and implemented in some leading academic journals in other fields (e.g., McShane et al., 2019; Wasserstein & Lazar, 2016). Also, my suggestion does not mean that research that explicitly tests hypotheses (e.g., in pre-registered controlled experiments) should abandon hypothesis testing or be abandoned. All I

2 It appears to me that our field is becoming more and more responsive to developments in neighboring and foundational disciplines (e.g., finance, economics, econometrics, and psychology) because of a broader education and collaboration across the boundaries of the various disciplines. As a case in point, see, for example, the speed with which methodological developments in the context of staggered difference-in-differences approaches found their way into accounting research (Baker et al., 2022; Barrios, 2021; de Chaisemartin & D'Haultfœuille, 2020).

3 E.g., policies on code disclosures and data repositories for journals.

am suggesting is that most empirical research in accounting examines associations or causal effects (e.g., of regulations) in complex data. Those studies attempt to loosely motivate their tests or explain their findings using abstract theories. For those studies, hypothesis testing, which entails spelling out pre-specified hypotheses and strictly relies on significance testing (a *p*-value cutoff) *against* those hypotheses, does not fit well to how the research is conducted. As a result, it leads to questionable practices and inferences.

Two common features of empirical accounting research render hypothesis testing a particularly poor fit for purpose (in most but clearly not all cases). First, empirical accounting research is often interested in how one accounting-related factor (e.g., disclosure quality) affects various broad outcomes (e.g., firm value). The broad outcomes are the result of many complex interactions and drivers of which the accounting-related factor is often not the most important one (hence, the low R^2) (Zimmerman, 2013). Accordingly, we need to worry about the proper specification of our empirical model (e.g., which variables to include and which to safely exclude). There are often multiple plausible models one could use. Differentiating between those models based on *a priori* reasoning is often not easy.[4] Second, and relatedly, empirical accounting research often lacks a credible theory as a null to test against. Hypothesis testing, by its very nature, is limited to informing us whether we can reject a theory or not (Popper, 1959). To learn something useful from this binary outcome, we need a theory that *could* plausibly explain the data. Otherwise, we are testing against a strawman. A prominent example in the world of physics is Einstein's theory of relativity (Einstein, 1916).[5] Finding data that allows to test and possibly reject this theory is very useful (e.g., Lea, 2022). In accounting research, however, I struggle to find a sufficient number of theories that aim to provide predictions of first-order relevance in the data that one can productively test against them.

As a result of those two related features of accounting research—complex data and lack of specific/descriptive theories—, we often have substantial degrees of freedom in choosing our empirical specification and do not test against a specific null. Instead of proper hypothesis testing, we tend to motivate potential deviations from an (often implausible) statistical null via disparate theories that *could* explain deviations in one or the other direction.[6] Those questionable practices rightly evoke

4 The relative credibility of different designs is in part a matter of statistical reasoning (e.g., about distributions and correlations) but even more so a matter of institutional knowledge and theory. For a related discussion on design-based empirical approaches in accounting research, see Leuz (2022).

5 Closer to home, a prominent example in the world of finance is the capital asset pricing model (Lintner, 1965; Sharpe, 1964).

6 Our focus on qualitative (i.e., binary or directional) inferences and "surprising" effects limits our field's ability to inform practitioners and regulators, beyond providing them with a sense for extant trade-offs. For an applied and maturing field like accounting research, it hence would seem useful to transition from solely focusing on qualitative inferences and "surprising" effects to working on quantifying the costs and benefits of various accounting properties and regulations (e.g., the elasticity

unease about the extent to which we can learn from empirical accounting research. They open the door for data dredging, invalidate the meaning of statistical significance tests, and result in unclear inference (i.e., what do we learn from rejecting an implausible null?).[7]

As an alternative to the hypothesis-testing norm, I would recommend adopting a more exploratory approach to (presenting) empirical research, which (a) includes an inversion of our standard paper structure and (b) relies on a more Bayesian approach to inference:

3.1 Paper Structure

The typical empirical accounting paper develops its hypotheses before its empirical tests. As discussed above, however, those hypotheses are often quite vague and more about reasons for deviations from a statistical null than about the null itself (i.e., the main part we learn about when we do statistical significance testing against a null). Even worse, the hypotheses (and the broader write-up) of the typical accounting paper tends to be informed by the results; that is, formulated ex post. In the paper structure and hypothesis-testing logic, however, those hypotheses are presented as ex-ante predictions. Through post-hoc revisions and adjustments of hypotheses, the entire hypothesis testing regime (including the statistical significance testing) loses its credibility and meaning.[8]

An alternative, more honest approach to structuring most empirical accounting papers would seem to be the following: We start with a motivating question (e.g., assessing the impacts of a regulatory change), briefly provide intuition for what one may expect given prior knowledge and theory (as guidance for the readers), and then dive into the data. After the data analysis, we can then discuss the theory or theories

---

of capital access with respect to disclosure). A better quantification will typically *not* provide surprising results (i.e., show the opposite sign as predicted or a particularly large or small magnitude relative to our prior expectation). Still, given the relevance for decisions made in the field of accounting (e.g., on whether and how to report or regulate), even small updates regarding the plausible range of a cost or benefit estimate can be of great importance. More broadly, the decision-making context in which the estimate is used is important for judging the "economic significance" of a statistical estimate, not simply its statistical significance or magnitude (e.g., McShane et al., 2019; McShane & Gelman, 2022).

7 For a broader discussion of issues with null-hypothesis testing in accounting, see, e.g., Cready (2022).

8 As a result, it seems *as if* we almost always find support for our hypotheses. This impression is, to a large part, owed to ex-post adjustments to our hypotheses and selection (e.g., the file-drawer problem) rather than our powerful theoretical understanding of the world. Our actual ability to predict how the world works via ex-ante reasoning appears quite limited, as revealed by the few studies that strictly followed a pre-registration approach in accounting research (Bloomfield et al., 2018).

that most likely explain the data.[9] This way, we do not pretend to know beforehand which two theories are the most relevant ones. Also, this way we can inform our theories (or toy models) by the data. That is, we then know what aspects appear unimportant vis-à-vis important to rationalize the data. This structure, with specific models as ex-post explanations after the data analysis, is common in modern applied microeconomic research (Mahoney, 2022).[10] For accounting, it offers an opportunity to bring empiricists and theorists together in a productive way (e.g., Chen et al., 2016).[11]

Importantly, this approach does not demote theory, by reducing ex-ante hypothesizing. To the contrary, it takes theory more seriously. Before the data analysis, it relies on broad theoretical concepts (e.g., supply and demand intuition) to motivate and design tests (e.g., guiding the selection of relevant forces to control for)—our priors. After the data analysis, it allows advancing more targeted, setting- and result-specific theoretical models—our posteriors.[12] It, in essence, uses theory twice—without dressing up ex-post reasoning as ex-ante hypotheses.

3.2 Bayesian Inference

Data analysis outside of hypothesis testing and related frequentist statistical significance testing can be done following a Bayesian approach to inference.[13] Just as the typical agents in accounting models, Bayesian researchers look at data to update their prior beliefs about certain quantities or relations (e.g., between disclosure quality and firm value). Notably, as Bayesians, we learn from data irrespective of any

9 This structure should not simply become a new norm. Rather, it should be permissible and used whenever it fits the underlying research process best. Notably, I expect this structure to work well for descriptive studies as well as most studies interested in identifying causal effects (e.g., regulatory effects). Studies focused on identifying causal effects attempt to isolate a specific cause (e.g., a regulation) for observed data patterns. The resulting causal effects, however, cannot easily be interpreted without theory (Armstrong et al., 2022). To aid with their interpretation, hence, we may want to propose (ex-post) context-specific theories consistent with the data and institutional details. Those theories can provide further predictions, which can lead to additional data analysis, resulting in an iterative process aimed at narrowing the set of plausible theories. This process is the basis of scientific work. It should not, however, after the fact, be portrayed as the result of ex-ante theorizing and ex-post hypothesis testing.

10 For examples in the accounting and finance literature, see, e.g., Bernard and Thomas (1989), Clinch (1991), and Kandel and Pearson (1995).

11 An alternative approach for combing theory and data is to make the theory more descriptive and take it to the data in its entirety. For a primer on structural estimation in accounting, refer, for example, to Bertomeu et al. (2023).

12 For a description of accounting theory and research in Bayesian terms, see Johnstone (2018).

13 See also Breuer and Schütt (2023), Glaeser and Guay (2017), Schütt (2022) for related suggestions for accounting research.

(arbitrary) significance thresholds. Bayesian learning is more continuous. It allows learning about the expected magnitude of the relation of interest as well as our (un) certainty about it. This approach to inference prevents inferential mistakes made when we neglect statistically insignificant coefficients ("no relationship!") and focus only on significant coefficients (Abadie, 2020). Importantly, this approach to inference explicitly incorporates that we may have other sources of information (e.g., prior studies, prior data, prior regressions) that inform our view of the world. This strength of Bayesian inference, which allows for truly cumulative science and circumvents the circularity issues inherent in our practical use of significance testing, is often cause for concern for people unfamiliar with Bayesian inference. Fortunately, in accounting research, we often deal with a lot of data. In those cases, the priors really do not matter much. Notably, this feature implies that we do not even have to go "full Bayesian." Instead, we can often simply use the same regression estimation tools we are already used to. But we can focus more on coefficient magnitudes and (un)certainty (e.g., standard errors) than on arbitrary *p*-value cutoffs (e.g., Imbens, 2021; Johnstone et al., 1986).[14]

4 Concluding Remarks

Most readers of empirical accounting research do not take the research and its purported inferences at face value. This skepticism is warranted, in accounting as well as other fields. First and foremost, it reflects that empirical (accounting) research is hard to do credibly. We tend to be interested in very specific effects on broad outcomes that are affected by a host of factors. Accordingly, without specific theory and truly exogenous variation, we need to be aware of the limitations of our research.

We can increase the "honesty" of the reporting of our research though; a matter that should be close to heart for accountants. While critical questioning in seminars may be one way to achieve this goal, I submit that there are two additional, practical avenues we can take. First, we may want to reconsider the standard format of an empirical accounting paper. I advocate for an inversion of the structure for most empirical studies (i.e., those interested in descriptive, exploratory, or regulatory research, for example). We may want to put data first and then discuss, informed by the data, the model(s) that are most compatible with the data. Second, and relatedly, we may want to use a Bayesian approach to inference. This approach takes the feedback loop from (prior) data into account and allows for learning about magnitudes and uncertainty of relations of interest. Those two approaches together can

14 Other barriers to using Bayesian inference are its required statistical knowledge and computational power. Those barriers are quickly diminishing given various primers and courses on Bayesian statistics (Schütt, 2022; van de Schoot et al., 2021) and the implementation of Bayesian estimation tools in statistical software packages (e.g., STATA).

substitute for the prevailing norm of presenting (pseudo) ex-ante null hypotheses and testing *against* them in the data. This prevailing norm does not appear to provide a good fit for most empirical accounting papers. It only appears to fit in the rare case where we have a relevant theory providing an important null prediction which we can productively test against using credible research designs that allow us to abstract from all other (un-modelled) factors (e.g., with lab or field experiments).

More honest and credible inferences can plausibly be achieved by migrating to the alternate paper-structure and inference approach proposed in this comment (and adopted in various other fields by now). Such improved inferences, however, would still not warrant placing excessive faith in any individual empirical research paper. We all make mistakes all the time (e.g., in choosing empirical designs, coding tests, sampling data, or generalizing results) given the uncertainty and cognitive limitations we face. Accordingly, we may want to continue to take the content of presented and published papers "with a grain of salt" (Ohlson, 2025, p. 1). This healthy skepticism is not an indication of a sorry state-of-affairs though. It rather reminds us that we may want to aim at creating a convincing body of research, collectively, as a profession, by promoting good practices (not simply accepted ones) through our research, teaching, and seminar attendance.

5 Annex

Ohlson (2025) advocates caution in cases characterized by "large N" and "small t-statistics." As a heuristic to identify when caution is warranted, the paper suggests using the ratio of the t-statistic over the square root of N. I agree that this heuristic can be useful. It, however, can also be misleading. In addition, it is not a sufficient (or necessary) statistic to judge the "importance" (i.e., relevance) and "existence" (i.e., credibility or identification) of a statistical association.

The heuristic reflects the fact that, all else equal (i.e., holding the economic magnitude of the association fixed), t-statistics increase in the sample size. As a result, statistically significant t-statistics can be obtained in large samples even for economically small associations. The proposed $t/\sqrt{N}$ heuristic attempts to account or correct for this relation. This simple heuristic works best for (unadjusted) t-statistics. It does not necessarily work well for t-statistics which were adjusted to account for dependence among observations of the large N sample. Many accounting databases, for example, contain thousands of observations. But those observations tend to be dependent, because they arise from a smaller number of firms (e.g., panel data with several observations per firm). To account for such dependence, accounting researchers frequently cluster standard errors, which tends to result in lower

t-statistics (e.g., Conley et al., 2018; Petersen, 2008). Notably, the adjusted *t*-statistic does not directly relate to the sample size *N* anymore. Rather, the adjusted *t*-statistic (i.e., the numerator of the heuristic) relates to the effective sample (i.e., the number of independent observations). The effective sample is smaller than *N*. This reduced effective sample size should be used in the denominator of the proposed heuristic whenever the numerator (i.e., the *t*-statistic) is adjusted for dependence.

The effect of dependence adjustments such as clustering on the proposed heuristic can best be illustrated by the extreme case of perfect dependence. Consider, for example, a case where the large *N* sample consists of several perfectly duplicated (i.e., dependent) observations per firm. In this case, the number of independent observations corresponds to the number of firms. Adjusting *t*-statistics through clustering at the firm level takes care of the dependence issue. The effective sample size (*M*) is then the number of firms or, equivalently, clusters, which is lower than *N*. The relevant heuristic in this case, accordingly, would be the ratio of the clustered *t*-statistic over the square root of *M*. In cases where the dependence is less than perfect (within cluster), the effective sample size will lie somewhere between the number of clusters and *N*. As this number is rarely reported, the proposed heuristic cannot readily be inferred by readers and seminar participants. Using a short-cut and ignoring the denominator effect (i.e., using *N*) would understate the size of the (adjusted) *t*-statistic relative to its underlying effective sample size.

An alternative heuristic, which is unaffected by dependence adjustments of *t*-statistics, is the use of (incremental) R-square values; that is, the change in explanatory power observed when including/dropping the variable of interest (e.g., Johannesson et al., 2023). While this measure is also not often reported, I agree with Ohlson (2025) that it would often merit reporting. It is important to acknowledge though that this heuristic (i.e., the size of the incremental explanatory power) is not a sufficient (or necessary) statistic for judging the existence (i.e., credibility) or importance (i.e., relevance) of a statistical association either.

The explanatory power of a variable of interest captures its comovement with the outcome *relative* to the total variation in the outcome. The total variation in the outcome can be a useful benchmark. It, however, does not have to be. In many cases, accounting research deals with noisy outcomes that are affected by various forces (e.g., firm value). Accordingly, the outcome exhibits a lot of (unexplained) variation. By contrast, our variables of interest often exhibit limited variation. A regulatory change, for example, tends to occur only once during our sample period. Accordingly, its variation is quite small in relation to the total outcome variation. It will explain little of the total outcome variation, even if it has a clear and economically sizable effect on the outcome. Accordingly, it may be at least as relevant (if not substantially more relevant) to examine the implied magnitude of a typical change in our variable

of interest (e.g., its coefficient multiplied by the standard deviation of the variable of interest) as it is to examine the incremental explanatory power of the variable of interest. (I caution though that neither the statistical significance of a statistical estimate nor its magnitude are sufficient statistics for the importance or "economic significance" of the estimate. For related discussions, see McShane et al. (2019) and McShane and Gelman (2022).)

It is naturally a subjective matter to choose the most relevant benchmark to judge the importance of a given association. Irrespective of the choice of benchmark, however, the importance of the association (e.g., size of correlation or explanatory power) does not establish its existence, where existence refers to the identification of an underlying causal relationship as specified by the empirical model. Even relatively large associations or explanatory power changes do not allow researchers to claim that they have identified a causal link. Identification rests on an assumption. In the usual case of linear regressions, this assumption intuitively requires that other (un-modelled) factors determining the outcome are not correlated with the variable of interest. This assumption cannot be tested. It must be argued for based on theoretical and institutional grounds (e.g., Breuer & deHaan, 2023; Leuz, 2022). Notably, research designs which provide more credible arguments for the validity of the identifying assumption often explicitly focus on a subset of the variation in the variable of interest, which is plausibly exogenous.[15] This approach improves identification; that is, researchers' confidence that the relation exists. At the same time, however, this approach often mechanically results in a lower explanatory power as the variation in the numerator of the R-square formula (i.e., the variation of the variable of interest) is reduced, while the denominator (i.e., the variation of the outcome) tends to remain unchanged. Hence, it is important to realize that, in those cases, the explanatory-power benchmark should only be used cautiously, as it tends to "bias" against "true" research findings.

Below, I briefly illustrate the above conceptual arguments with a stylized econometric model and a corresponding simulation:

5.1 Econometric Model

I am interested in estimating the causal relation between an outcome (y) and a variable of interest (x). The variable of interest contains variation that is independent of other factors determining the outcome (i.e., variation which satisfies the

15 The extent to which variation in the variable of interest is *plausibly* exogenous depends on the credibility (or "plausibility") of the theoretical and institutional arguments supporting the claim that the variation can be expected to be uncorrelated with other omitted factors.

identifying assumption) ($\tilde{x}$ with $\tilde{x} \perp \varepsilon$). It, however, also contains variation correlated with the other factors (ε):

$$x = \tilde{x} + \rho\varepsilon. \tag{1}$$

Equipped with data on the outcome and variable of interest, I run the following regression:

$$y = \widehat{\beta}x + \widehat{\varepsilon}. \tag{2}$$

The resulting regression coefficient is given by:

$$\widehat{\beta} = \beta\left(\frac{\sigma_{\tilde{x}}{}^2}{\sigma_{\tilde{x}}{}^2 + \rho^2\sigma_\varepsilon^2}\right) + \frac{1}{\rho}\left(1 - \frac{\sigma_{\tilde{x}}{}^2}{\sigma_{\tilde{x}}{}^2 + \rho^2\sigma_\varepsilon^2}\right), \tag{3}$$

where β is the true coefficient (relating the exogenous part of x to the outcome y) and $\left(\frac{\sigma_{\tilde{x}}{}^2}{\sigma_{\tilde{x}}{}^2+\rho^2\sigma_\varepsilon^2}\right)$ is the signal-to-noise ratio of the variable of interest (where the variation of the exogenous part is the signal and the variation correlated with the error part is the noise). The estimated coefficient is biased if the total variation in the variable of interest contains variation that is correlated with the other determinants of the outcome (i.e., $\rho \neq 0$). In case of a positive correlation ($\rho > 0$), for example, the estimated coefficient would be upward biased ($\widehat{\beta} > \beta$).

To obtain an unbiased estimate, I should focus on the exogenous part of the variable of interest, not the full variation:

$$y = \beta\tilde{x} + \varepsilon. \tag{4}$$

By focusing on a subset of the variation in the variable of interest, the explanatory power of the regression could decrease though. The comparison of the two R-square values for the full-variation and the exogenous-variation regressions illustrates this point. The R-square value for the full-variation regression (Equation (2)) is given by:

$$\widehat{R}^2 = \frac{\widehat{\beta}^2\left(\sigma_{\tilde{x}}{}^2 + \rho^2\sigma_\varepsilon^2\right)}{\sigma_y^2}. \tag{5}$$

The R-square value for the exogenous-variation regression (Equation (4)), by contrast, is given by:

$$R^2 = \frac{\beta^2\sigma_{\tilde{x}}{}^2}{\sigma_y^2}. \tag{6}$$

We observe that the non-exogenous part of the variation of the variable of interest affects the numerator of the R-square measure in two ways. It changes the (squared) coefficient and increases the variation in the variable of interest. In the case where

other factors are positively correlated with the variable of interest ($\rho > 0$), both forces increase the R-square of the biased regression relative to the R-square value of the unbiased regression. Bias, however, does not always favor the biased regression's R-square value. If the estimated coefficient value is biased toward zero (due to negatively correlated other factors), for example, the R-square value of the biased regression can be lower than the R-square value of the unbiased regression, even though the biased regression uses more variation (in the numerator).[16] This ambiguity illustrates that a focus on R-square values and other heuristics can be misleading. The importance and existence of causal effects should, accordingly, not be exclusively or even primarily be judged by reference to those heuristics. The heuristics may be more promising for other applications (e.g., prediction or description), though several limitations of the heuristics carry over to those other applications too.

5.2 Simulation

I illustrate the impact of dependence and corresponding adjustments (e.g., clustering) on the $t/\sqrt{N}$ heuristic and the impact of confounded variation on coefficient estimates and R-square values using simulated data. The setup of the simulation closely follows the econometric model described above. The parameterization of the simulation is as follows:

Parameter	Value
$\tilde{x}$	Normal (0, $\sigma_{\tilde{x}}{}^2$)
ε	Normal (0, σ_{ε}^2)
$\sigma_{\tilde{x}}{}^2$	1
σ_{ε}^2	1
β	0.5
P	0.2

Based on this parameterization, I simulate two samples. The first independent sample is based on 1000 observations ($N = M$). The second dependent sample adds 99 additional duplicates for each observation of the first sample. This sample exhibits an

16 The regression uses the full variation in x (i.e., $\sigma_x^2 = \sigma_{\tilde{x}}{}^2 + \rho^2\sigma_{\varepsilon}^2$) instead of only the variation in its exogenous subcomponent $\tilde{x}$ (i.e., $\sigma_{\tilde{x}}{}^2$).

effective sample size (i.e., number of firms) of 1000 (M), but a total sample size of 100,000 (N).

I obtain the following regression estimates for the full- and exogenous-variation regressions (using ordinary least squares regressions without constants):

Regression estimates						
Sample & specification:	**Independent**		**Dependent & unadjusted**		**Dependent & adjusted**	
Outcome:	***y***		***y***		***y***	
x	0.648		0.648		0.648	
	(26.05)		(260.59)		(26.05)	
$\tilde{x}$		0.493		0.493		0.493
		(17.54)		(175.46)		(17.54)
Observations (N)	1000	1,000	100,000	100,000	100,000	100,000
Clusters (M)	–	–	–	–	1000	1000

The table shows that the full-variation regression results in an upward biased coefficient estimate (estimate: 0.648; true value: 0.5). By contrast, the exogenous-variation regression almost perfectly recovers the true coefficient value (estimate: 0.493; true value: 0.5). The point estimates are not affected by the 99-times duplication of the data. The unadjusted (robust) t-statistics, however, increase by a factor of 10 (i.e., the square root of 100). This inflation in the t-statistics is perfectly reversed by adjusting for clustering of duplicate observations within firms.

The regression estimates map into the following heuristic values:

Heuristics						
Sample & specification:	**Independent**		**Dependent & unadjusted**		**Dependent & adjusted**	
Variable of interest:	***x***	***x***	***x***	$\tilde{x}$	***X***	$\tilde{x}$
$t/\sqrt{N}$	0.824	0.555	0.824	0.555	0.082	0.055
$t/\sqrt{M}$	–	–	–	–	0.824	0.555
R^2	0.384	0.215	0.384	0.215	0.384	0.215

The $t/\sqrt{N}$ heuristic provides relative rankings consistent with the ranking of R-square values ($t/\sqrt{N}$: 0.824 > 0.555; R^2: 0.384 > 0.215). The absolute values of the $t/\sqrt{N}$ ratio are unaffected by the duplication of data (i.e., dependence) if the t-statics remain unadjusted. If the t-statistics are adjusted for the duplication, however, the absolute values of the $t/\sqrt{N}$ heuristic are downward biased (by a factor of 10;

i.e., square root of 100) if the denominator is not correspondingly adjusted (i.e., $t/\sqrt{N}$: 0.082 for x and 0.055 for $\tilde{x}$). When the denominator is also adjusted for the duplication, by using the effective sample size M instead of the total sample size N in the denominator, the heuristic again recovers the same absolute values as before (i.e., $t/\sqrt{M}$: 0.824 for x and 0.555 for $\tilde{x}$).

The heuristics table shows that, for an apples-to-apples comparison, both the numerator and denominator need to remain unadjusted or need to be adjusted. Putting adjusted t-statistics in relation to (unadjusted) total sample sizes is not particularly useful, especially in case of high dependence. The heuristics table also shows that, in the case of other positively confounding factors, the heuristics can favor the biased instead of the unbiased specification. Those biases and ambiguities corroborate and illustrate the conceptual insights discussed above. They document that the proposed heuristics are questionable as sole or even primary metrics for judging the quality of specifications and, especially, the importance and existence of causal relationships.

Acknowledgment: I acknowledge helpful comments from Yuri Biondi (editor), an anonymous reviewer, Jonathan Glover, Anthony Le, Christian Leuz, Rongchen Li, Maximilian Müller, James Ohlson, Stephen Penman, and Harm Schütt. All errors are my own.

References

Abadie, A. (2020). Statistical nonsignificance in empirical economics. *The American Economic Review: Insights*, *2*, 193–208.

Armstrong, C., Kepler, J. D., Samuels, D., & Taylor, D. (2022). Causality redux: The evolution of empirical methods in accounting research and the growth of quasi-experiments. *Journal of Accounting and Economics*, *74*, 101521.

Baker, A. C., Larcker, D. F., & Wang, C. C. Y. (2022). How much should we trust staggered difference-in-differences estimates? *Journal of Financial Economics*, *144*, 370–395.

Barrios, J. M. (2021). Staggeringly problematic: A primer on staggered DiD for accounting researchers. Working Paper. www.ssrn.com/abstract_id=3794859

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

Bertomeu, J., Liang, Y., & Marinovic, I. (2023). A primer on structural estimation in accounting research. *Foundations and Trends in Accounting*, *18*, 1–137.

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.

Breuer, M., & deHaan, E. (2023). Using and interpreting fixed effects models. Working Paper.

Breuer, M., & Schütt, H. (2023). Accounting for uncertainty: An application of bayesian methods to accruals models. *Review of Accounting Studies*, *28*, 726–768.

Chen, Q., Gerakos, J., Glode, V., & Taylor, D. J. (2016). Thoughts on the divide between theoretical and empirical research in accounting. *Journal of Financial Reporting*, *1*, 47–58.

Clinch, G. (1991). Employee compensation and firms' research and development activity. *Journal of Accounting Research*, *29*, 59–78.

Conley, T., Goncalves, S., & Hansen, C. (2018). Inference with dependent data in accounting and finance applications. *Journal of Accounting Research*, *56*, 1139–1203.

Cready, W. M., 2022. Accounting research's "flat earth" problem. *Accounting, Economics, and Law: A Convivium*.

de Chaisemartin, C., & D'Haultfœuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *The American Economic Review*, *110*, 2964–2996.

Einstein, A. (1916). Die Grundlage der allgemeinen Relativitätstheorie. *Annalen der Physik*, *354*, 769–822.

Gelman, A., & Loken, E. (2014). The statistical crisis in science data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *American Scientist*, *102*, 460–465.

Glaeser, S., & Guay, W. R. (2017). Identification and generalizability in accounting research: A discussion of Christensen, Floyd, Liu, and Maffett (2017). *Journal of Accounting and Economics*, *64*, 305–312.

Harvey, C. R. (2017). Presidential address: The scientific outlook in financial economics. *The Journal of Finance*, *72*, 1399–1440.

Imbens, G. W. (2021). Statistical significance, p-values, and the reporting of uncertainty. *The Journal of Economic Perspectives*, *35*, 157–174.

Johannesson, E., Ohlson, J. A., & Zhai, S. W. (2023). The explanatory power of explanatory variables. *Review of Accounting Studies*. https://doi.org/10.1007/s11142-023-09781-w.

Johnstone, D. (2018). Accounting theory as a bayesian discipline. *Foundations and Trends® in Accounting*, *13*, 1–266.

Johnstone, D. (2021). Accounting research and the significance test crisis. *Critical Perspectives on Accounting*, *89*, 102296.

Johnstone, D. J., Barnard, G. A., & Lindley, D. V. (1986). Tests of significance in theory and practice. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *35*, 491–504.

Kandel, E., & Pearson, N. D. (1995). Differential interpretation of public signals and trade in speculative markets. *Journal of Political Economy*, *103*, 831–872.

Lea, R. (2022). *Einstein's greatest theory just passed its most rigorous test yet*. Scientific American.

Leuz, C. (2022). Towards a design-based approach to accounting research. *Journal of Accounting and Economics*, *74*, 101550.

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics*, *47*, 13–37.

Mahoney, N. (2022). Principles for combining descriptive and model-based analysis in applied microeconomics research. *The Journal of Economic Perspectives*, *36*, 211–222.

Martinson, B. C., Anderson, M. S., & de Vries, R. (2005). Scientists behaving badly. *Nature*, *435*, 737–738.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*, 235–245.

McShane, B. B., & Gelman, A. (2022). Selecting on statistical significance and practical importance is wrong. *Journal of Information Technology*, *37*, 312–315.

Ohlson, J. A. (2022). Researchers' data analysis choices: An excess of false positives? *Review of Accounting Studies*, *27*, 649–667.

Ohlson, J. A. (2025). Empirical accounting seminars: Elephants in the room. *Accounting, Economics, and Law: A Convivium 15*: 1–8.

Petersen, M. A. (2008). Estimating standard errors in finance panel data sets: Comparing approaches. *Review of Financial Studies*, *22*, 435–480.

Popper, K. R. (1959). *The logic of scientific discovery*. Hutchinson.

Schütt, H. (2022). What can bayesian inference do for accounting research? *Journal of Financial Reporting*. 1–8. https://doi.org/10.2308/JFR-2021-002.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, *19*, 425–442.

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.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*, 1.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, *70*, 129–133.

Zimmerman, J. L. (2013). Myth: External financial reporting quality has a first-order effect on firm value. *Accounting Horizons*, *27*, 887–894.