Critical Perspectives on Accounting xxx (xxxx) xxx

Contents lists available at ScienceDirect

Critical Perspectives on Accounting

journal homepage: www.elsevier.com/locate/cpa

Accounting research and the significance test crisis

David Johnstone

University of Wollongong, NSW 2522, Australia

ARTICLE INFO

Article history:
Received 28 November 2019
Revised 9 February 2021
Accepted 20 February 2021
Available online xxxx

Keywords:
Empirical accounting research
Significance tests
Replication crisis
p-levels
p-hacking
Registered reports

ABSTRACT

The emerging or at least threatening "significance test crisis" in accounting has been prompted by a chorus across multiple physical and social sciences of dissatisfaction with conventional frequentist statistical research methods and behaviors, particularly the use and abuse of *p*-levels. There are now hundreds of published papers and statements, echoing what has been said behind closed doors for decades, namely that much if not most empirical research is unreliable, simply wrong or at worst fabricated. The problems are a mixture of flawed statistical logic (as Bayesians have claimed for decades), "*p*-hacking" by way of fishing for significant results and publications, selective reporting or "the file drawer problem", and ultimately the "agency problem" that researchers charged by funding bodies (their Universities, governments and taxpayers) with conducting disinterested "objective science" are motivated more by the personal need to publish and please other researchers. Expanding on that theme, the supply of empirical research in the "market for statistical significance" is described in terms of "market failure" and "the market for lemons".

1. Introduction

In October 2011, Joseph Simmons, a psychologist at the Wharton School of the University of Pennsylvania, published a clearly preposterous result in the respectable journal *Psychological Science*. Together with Uri Simonsohn, also at Wharton, and Leif Nelson of the University of California, Berkeley, Simmons showed that people who listened to the Beatles song "When I'm Sixty-Four" grew younger, by nearly 18 months. But if the result was laughable, the point of the paper was serious: to show how standard scientific methods could generate scientific support for just about anything. ... The paper came at a critical time for psychologists. Earlier that year, another paper using standard methods had shown that extrasensory perception was a real phenomenon – a result the authors meant seriously, to the dismay of other psychologists. "If you use the techniques that everyone is using in their normal research ... and it supports the existence of bullshit, then there is good reason to think that that method is wrong more generally and you shouldn't be using it,"... . (Kupferschmidt, 2018)

This paper is about the statistical logic and positivist research ethos that appeared in accounting theory and research, starting in the 1970s. There are two interrelated "crises". The first is the slower boiling "significance test crisis" (Morrison and Henkel, 1970) which is partly about significance tests being flawed in logic and open to abuse. The second is the "replication crisis" which erupted in psychology in about 2010 (Simmons, Nelson, & Simonsohn, 2018a) and is mainly about significance tests being purposely or naively used in ways that produce "false positives" and unrepeatable conclusions.

E-mail address: djohnsto@uow.edu.au

https://doi.org/10.1016/j.cpa.2021.102296
1045-2354/

The replication crisis has arisen and been widely discussed across the sciences (Aarts et al., 2015; Begley, & Ellis, 2012; Ioannidis, 2005a, 2005b, 2019; Simmons, Nelson, & Simonsohn, 2011; Simmons, Nelson & Simonsohn, 2018a) has begun to be discussed in accounting (e.g. Dyckman, 2016; Dyckman & Zeff, 2014; Stone, 2018; Kim, Ahmed, & Ji, 2018,; Ohlson, 2015, 2019) and has roused concerns in the upper echelons of positivist accounting research (Bloomfield, Rennekamp, & Steenhoven, 2018; Hail, Lang, & Leuz, 2020). The celebrated Bayesian paper by medical researcher Ioannidis (2005b) titled "*Why most published research findings are false.*" has been cited more than 10,000 times.

There is no published paper in the accounting literature describing the many different problems with significance tests and their use, despite the hundreds of published papers on that topic in statistics proper and applied disciplines. There is much therefore to say. To impose some order on this discussion, I have broken this paper into three main sections. Section 1 explains the statistical critique that Bayesians have directed at significance tests since the 1950s. Section 2 explains the replication problem and its starting point in researchers' use of significance tests. Section 3 discusses the culture of the empiricist academy, which for decades has not acknowledged the significance test crisis or its potential to affect accounting research quality and reliability.

2. Bayesian logic and significance tests

The Ioannidis (2005b) critique and most criticism of significance tests and testing is Bayesian in its logic and philosophy. The difference between orthodox "frequentist" and Bayesian statistical method is that frequentists start with a null hypothesis H and data x (and an underlying model) and ask "what is the probability of observing data *as or more discrepant* with H as the actual observed data x, if in fact H is true?", whereas Bayesians ask much more directly and to the point "what is the probability of H being true once having observed exactly x?".

Given observed sample evidence x, the Bayesian probability of H is written as $p(H|x)$ – meaning "the probability of H given x" – and is found using the probability law called Bayes theorem, i.e., $p(H|x) = p(H)p(x|H)/p(x)$. The anti-Bayesian frequentist school agrees completely with Bayes theorem as a mathematical law, but does not use it. It is set aside ostensibly because the prior probability of H, $p(H)$, is "too subjective". Bayesians regard this argument as humbug (see e.g. Gelman, 2008) because the entire analysis, including the model itself, summarized by $p(x|H)$, is subjective. Their position is that all statistical analyses are innately subjective and the best approach is to admit and embrace that fact of life. Any analysis that attempts to fence out subjective inputs will necessarily introduce bias by suppressing whatever is already believed or suspected based on related research and theory. By using only what is deemed "admissible" in "objective" analysis, the range of possible conclusions from empirical research is artificially constrained and open to arbitrary manipulation.

Worse still, the frequentists' attempt to avoid explicit subjectivity violated the probability logic of Bayes theorem by imposing a new "logic" (e.g. see Lindley, 1987). That logic of significance tests is due primarily to the revered R.A. Fisher (Johnstone, 1987a; Seidenfeld, 1979). It is a roundabout logic and goes like this: you observe a value x, call it x_{obs}. You then find the probability of observing any other x deemed "as or more discrepant" with H than the x_{obs} that you actually observed. That new probability is called a p-level, and the smaller it is the more "discredit", in some unstated and necessarily subjective sense, you attach to H. Moreover, you don't worry if the probability of H given your observed x_{obs} is high under Bayes theorem, because you have already rejected Bayes theorem despite that you fully agree with it as a valid law of reason.

Note that the mathematical definition of the p-level is $p(x \geq x_{obs}|H)$ but the common mistake is to think of the p-level as the Bayesian probability $p(H|x_{obs})$, which is a natural inclination and fudge given that the Bayesian measure is what researchers ultimately want.

A simple example of the clash of Bayesian and frequentist logics goes as follows. Let there be two feasible hypotheses H_0 and H_1 (where H_1 represents not-H_0), and three possible sample observations, x_1, x_2 and x_3. The "null hypotheses" being tested is H_0, and H_1 is called the "alternative hypothesis". The agreed probabilities, all of the mutually acceptable form $p(x|H)$, are shown in Table 1 (these constitute the so-called "likelihood function", which both Bayesian and Fisherian methods make use of – in their distinctly different ways).

Note from the Table that x_1 is the most probable of the three possible observations if in fact H_0 is true, and x_3 is the least probable, so x_1 is regarded as the least discrepant with H_0 and x_3 as the most discrepant with H_0.

Now imagine that the experiment is run and the observation that arises is x_2. How do we interpret that observation with regard to the null hypothesis tested, H_0? The rival answers are as follows:

Fisher's approach (p-levels). Having observed x_2, the probability of a result as or more discrepant with H_0 is

$$p - \text{level} = prob(x_2 \text{ or } x_3|H_0) = 0.045 + 0.0045 = 0.0495.$$

Since this observed p-level is less than the conventional threshold of 5% – which was institutionalised by accident in statistical history and has no more status as a threshold than say 0.049 – we regard the null hypothesis as having been discredited and publish our result.

Bayesian approach. We find the probability of H_0 given the actual observation x_2. It will depend on the prior probability of H_0, which we can introduce if required, but is not necessary to grasp the problem. The fact is that the observed data x_2 is nine times more probable under H_0 than it is under H_1, so, by the logic that is implicit and intuitive within Bayes theorem, the observed result x_2 supports H_0 and clearly does not discredit it. More such evidence would ultimately lead to very strong Bayesian belief in H_0, regardless of the starting prior belief (i.e., "data swamps prior").

Table 1
$p(x|H)$.

	H_0	H_1
x_1	0.45	0.0005
x_2	0.045	0.005
x_3	0.0045	0.05

We could find the posterior probability of H_0, i.e., $p(H_0|x_2)$, but that would only accentuate the fact that the actual evidence x_2 is far more consistent with the null hypothesis than the alternative, despite the null being "rejected at 5%" by a low p-level.

Note the basic difference between the two approaches regarding how data is summarized. In the Bayesian way, the test data or signal is the point observation x, whereas in significance testing x is lost or blurred by treating it not as x precisely, but as data "as or more discrepant with" H_0. That obfuscation of observation x into something more like "x greater than or equal to x" is what underlies the famous *reductio ad absurdum* held against frequentist methods by the founding Bayesian Harold Jeffreys:

> An hypothesis that may be true is rejected because it has failed to predict observable results that have not occurred. (Jeffreys, 1939, p. 316)

Quips like this had personal bite and gave rise to the malice that characterized relations between conservative frequentists and the more critical and liberal "subjectivist" Bayesian statistical theorists, and remained an unhealed sore in statistical theory until at least the 1980s when Bayesian theoretical contributions became increasingly mainstream and began to appear routinely in the *Journal of the American Statistical Association* and similarly authoritative theory journals. Now that the statistics war is largely over, and with the boost from modern computing power and Bayesian software, Bayesian methods are de rigueur, especially with their wide application in computer science and machine learning.

There is a list of logical defects that Bayesians find in conventional significance tests (see e.g., Berger, 1985) and also, just as much an issue, there is a further list of ways in which results of significance tests can be either misinterpreted or twisted so as to misrepresent what was found (now commonly called "p-hacking"). Bayesian alternatives and quasi-Bayesian modifications to significance tests have been developed to counter intrinsic problems in the logic of significance tests, and there is much being written on codes of research practice that could lead to clearer and more justifiable interpretations of evidence. The current initiative by the journal editors of JAR (see authors Hail et al., 2020) falls into the second category of intended remedies, but does not go so far as to question the underlying logic of significance testing or raise the issue of conclusions that might not follow logically from what was observed (see the example above).

The most widely known allegation against significance tests is that "*statistical significance*" is just not significant, meaning that "statistically significant" results, if properly examined, often carry only weak evidence of any material effect.

To illustrate, suppose that the "real world" contains a relationship of the simple form

$$Y = a + bX + \epsilon,$$

where $a = 0$, $b = 1$, X is a normally distributed variable N(1,1) with mean of one and standard deviation one, and lastly ϵ is normally distributed N(0,10) with mean zero and standard deviation 10. The random influence on Y is thus ten times as large as the effect of the independent variable X.

Having constructed an argument or "story" about why X is important in driving Y, the researcher draws a random sample of $n = 100$ pairs (X,Y) and examines that data via the scatter plot shown in Fig. 1.

Looking at the plot in Fig. 1, the relationship between X and Y is visibly weak and appears to offer little or no practical significance in terms of producing better Y outcomes by making changes to the causal factor X (there is too much extraneous random influence over Y).[1]

However, a regression is run on the observed sample and it is found that the (one-sided) p-level of the data with respect to the usual null hypothesis $H_0 : b = 0$is 0.0053, which is low given the wide spread of data points and the relatively small sample size. Despite that wide scatter, a p-level of 0.0053 is easily sufficient to satisfy the journal editor of there being a "statistically significant" or simply "significant" relationship between the variables X and Y. The real significance of this relationship is best seen by a visual inspection of the plotted data.[2] The effect of explanatory variable X is there but is swamped by the "noise" of other unknown or random effects. Also, showing the danger of a small sample of just 100 observations, and no replication, the estimated slope coefficient b is 2.603, or 2.6 times its true value.

This simple simulation is evidence of the paucity of "significant at 5%" as an indicator of a substantive and reliable result. It reveals the widely flouted distinction between "statistically significant" and practically significant results. One accounting

[1] The statistician Tukey (1977) put the case that descriptive data analysis by plots should precede any fancy statistical methods, because statistical significance is no substitute for a relationship that is clearly visible or invisible in the data. See the modern statement of this approach in Behrens (1997).

[2] It was correctly pointed out to me by a referee that the Bayesian and frequentist confidence intervals in this mini-case are equally wide. That is correct, provided that the Bayesian uses a uniform prior distribution. The point however is that the Bayesian interprets this result as too weak to mean much, whereas the frequentist interprets it as "highly statistically significant".

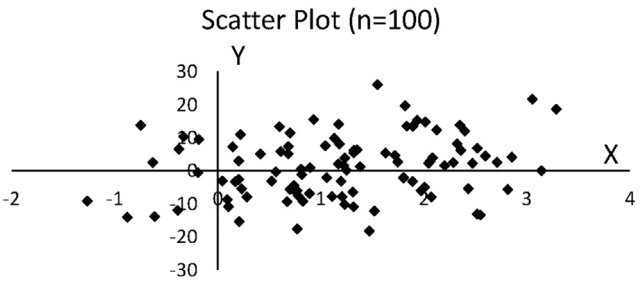

Fig. 1.

research paper in which this distinction is made explicitly is (Basu & Markov, 2004) who showed that their virtually zero observed effect is nonetheless "statistically significant". By relying on the convenient publication criterion of "significant at 5%", unplotted data patterns resembling the one in Fig. 1 are portrayed throughout the literature as "statistically significant". This is the issue of "false positives", whereby a positive or actual relationship is implied when in fact there is either no relationship at all, or when that relationship is too small and possibly too fleeting and sample specific to be of substance or practical significance.

Remarkably, especially given that superb graphical methods are now automated in software, data plots are rarely reported in the empirical literature. A recent paper by Basu, Vitanza and Wang (2020) makes this point tellingly. Fig. 1 may well explain why data plots that discredit findings of statistical significance, which of themselves are held out as evidence of a publishable relationship between variables, are not an automatic inclusion in empirical research papers. A good example of how a plot can inform is found in the famous Ball and Brown (1968) graph of stock price reactions to positive and negative surprises. This plot shows such a clear empirical correlation that significance tests are redundant. That is the hallmark of a strong empirical result – it is so strong, and in this case so replicable, that little statistical finery is needed to support it.

Heightening the problem of practically insignificant results being marketed in journals as "statistically significant" is the "Lindley Paradox" (Johnstone & Lindley, 1995; Lindley, 1957). In essence, Lindley showed that as the sample size in the test gets larger, a result which has *p*-level of 5% (or 1% or any given fixed hurdle) represents stronger and stronger evidence *in favour of the null hypothesis* (i.e., against the hypothesis that is supposedly "rejected at 5%"). Against intuition, a result that seems stronger because of its higher sample size is actually weaker. Details and simple illustration are provided in Johnstone (2018, pp.37–44).

2.1. Bayesian versus frequentist philosophies in accounting

Accounting is deeply at odds with itself over its preferred statistical logic. On one hand, accounting information theory (e.g., Feltham & Demski, 1970) is avowedly and overtly Bayesian. The users of financial statements are assumed to use Bayesian logic to interpret firms' financial disclosures, draw probabilistic inferences, and ultimately make decisions, all according to the inherently Bayesian ideal of economic rationality (subjective expected utility). In a philosophical paper on the role of accounting information, Chen and Schipper (2016) criticized the Ohlson accounting valuation framework and the empirical "value relevance" literature on grounds that neither embeds a normative "rational economics" description of accounting information users. According to that normative model, the users of financial reports are information assimilators who arrive at their subjective beliefs and consequent valuations of stocks by strictly Bayesian updating:

> This valuation approach does not model how investors use accounting information to update their beliefs about firms' future dividends. ... Absent a theory or at least an analytical structure explicitly considering investors' use of information (e.g., investors' prior, Bayes updating), the interpretations of these results must of necessity be ad hoc. ...We are not implying that the residual income frameworks revived by Ohlson (1995) and others have no value. In fact, we believe this research provides useful insights on the role of accounting measurement. Our point is that this research is not suitable to answer questions related to how investors use accounting data to update their assessments of estimates of future cash flows. (Chen & Schipper, 2016)

Modern analytical accounting theory, particularly all of the celebrated work on the "noisy rational expectations" theory of accounting disclosure**,** could not exist without the formal logic of Bayesian inference and choice (rational action).[3]

> In theory-based, economic analyses, reliance on Bayes rule is so routinized an assumption as rarely to warrant any justification. The compelling feature of Bayes rule is that it implies the most efficient use of information.

[3] See Verrecchia (2001) review paper and Johnstone (2018).

> Consequently, in market settings, investors who use information more efficiently (i.e., Bayesians) should be able to exploit and dominate their less efficient counterparts. (Verrecchia, 2001, p. 123)

Empirical research papers in financial accounting typically have no rigorous or "objective" theoretical derivation, and are often based on a wordy "story" propped up by references and testimony to previous empirical studies:[4]

> Accounting investigations often rest only on a story rather than on a theory. A major problem here is that a story, rather than a theory, can be changed or modified, which encourages data mining. (Dyckman & Zeff, 2019, p.2)

Anyone who sits in on accounting/finance seminars delivered by empirical researchers will commonly hear the paper's explanation called a "story" (e.g., "this paper is based on an information asymmetry story"). It is remarkable that this word, which is so close to "fairy story", is vital in an empiricist vocabulary that puts so much importance on rigour, precision and objectivity. Ultimately the paper's story ends up being translated into a rigorous looking regression equation which is tested in the same way as if had been derived from theory in the manner for example that Mendel derived his statistical hypotheses from his genetic theory.[5] Ohlson (2019, p.5) notes that "researchers are prone to get attached to their story as to how the world works, so they look for ways their emotional commitment can be validated".

When there is an attempt to justify the content of a regression equation or model with statistical theory, Bayesian financial economics and rationality is the fallback position prescribing how the users of financial statements infer and decide. The following quote shows a way of thinking, traced to Bayesian theorist DeGroot, which can be found in the narratives that lead into many empirical accounting research studies:

> Based on Bayesian decision theory research (e.g., DeGroot, 1970) that shows that loss-minimizing investors place less weight on noisier (i.e. more uncertain) information, we expect to observe more muted initial market reactions to unexpected earnings signals that have higher information uncertainty. (Francis, LaFond, Olsson, & Schipper, 2007, p.408)

It is by quotes like this that empirical accounting researchers require Bayesian rationality of others, yet their own statistical methods, virtually always *p*-levels, are ritualistically "frequentist" and rarely if ever mention any Bayesian or even quasi-Bayesian interpretation of the observed empirical/statistical results. Instead, their results papered all over the literature are all *p*-levels, stars or other frequentist reporting conventions for reporting frequentist (anti-Bayesian) statistical results; see Ohlson (2015) on "seeing stars".

The obvious question for insiders within the positivist statistical research paradigm is why empirical accounting research is so religiously frequentist, relying entirely on frequentist concepts like "error frequencies", "power" and "*p*-levels", when at the same time accounting information theory is always explicitly and committedly Bayesian. A field that upholds both Bayesian and frequentist statistical "logics" at once has methodological questions to resolve.

3. The replication crisis

The institutionalised use and misuse of *p*-levels across the social sciences has reached a point that the *American Statistical Association* (ASA) has published a statement effectively disowning common practices surrounding the reliance on and interpretation of significance tests (*p*-values) and calling for the application of alternative inference methods, particularly Bayesian research methods. See e.g., Greenland et al. (2016), Wasserstein and Lazar (2016) and Benjamin, Berger, [...], Johnson, and V.E. (2018), and the many other references below. That reckoning has been prompted in large part by the so-called replication or reproducibility crisis, where effects in psychology and other fields that were supposedly known and "statistically significant", and had been accepted as exemplars of what empirical research could achieve, have lately been retested and often failed to replicate.

> There is increasing discontent that many areas of psychological science, cognitive neuroscience, and biomedical research are in a crisis of producing too many false positive non-replicable results. This wastes research funding, erodes credibility and slows down scientific progress. Since more than half a century many methodologists have claimed repeatedly that this crisis may at least in part be related to problems with Null Hypothesis Significance Testing (NHST). However, most scientists (and in particular psychologists, biomedical scientists, social scientists, cognitive scientists, and neuroscientists) are still near exclusively educated in NHST, they tend to misunderstand and abuse NHST and the method is near fully dominant in scientific papers. (Szucs & Ioannidis, 2017)

[4] Schrand (2019) commented on the "bloat" in empirical research papers brought by authors multi-page homage to previous authors' results before beginning their own work.

[5] Interestingly, the production of "stories" in empirical research puts that research closer to critical accounting theory than empiricists might admit. The main difference is that in critical accounting theory, the story is itself the theory (often based on a more general theory from sociology or anthropology) and is not the mere stepping stone to what is ultimately presented as "objective" and anything but a story. Another obvious difference is that stories in empirical research are mainly from neo-classical rational financial economics rather than from, say, political economics or critical theory..

The replication[6] crisis gained its current momentum after Pashler and Wagenmakers (2012). Since then, multiple studies in psychological and medical sciences have found that published results could be replicated in only small proportions, typically only 20–50% (e.g., Camerer & Dreber, 2018). The Open Science Collaboration (2015) found that 36% of 100 psychology replications gave "significant at 5%" *p*-values, whereas the original studies gave 97% "significant" effects, and the average observed effect was near half the magnitude of the original.

It will be clear from how seriously the replication crisis and *p*-level controversy is taken in other, longer established, empirical research disciplines (e.g. medicine, psychology), that accounting as a discipline will ultimately have little choice but to join all social and behavioural sciences that have the same fundamentals – i.e., reliance on statistical models composed loosely from a montage of earlier models and intuitions rather than an underlying deductive theoretical derivation, innately noisy data affected by multiple factors and surrounding conditions, non-stationarity in underlying causal relationships, a physical inability to impose experimental controls, variables that can't be measured directly but only by often vaguely related or arbitrary proxies, ubiquitous reporting of *p*-levels rather than more revealing statistics or confidence intervals, historically little concern for replication, and typically few directly concerned readers outside an often internally motivated and jingoistic academic community producing the research.

In psychology, where the replication crisis arose after some famous and infamous empirical findings aroused suspicion and could not be replicated, critics suggest that the problem is not merely poor statistical methodology or discipline – instead, it is poor science based on poor theory. Because so much theory in psychology is vague, scant or ad hoc, as could be said also for many theories in empirical accounting, statistical hypotheses can arise on a publication whim or for any "unscientific" reason – e.g., they are politically acceptable within a journal's research community – rather than via an a priori derivation within an already established theoretical framework:

> Without an overarching theoretical framework that generates hypotheses across diverse domains, empirical programs spawn and grow from personal intuitions and culturally biased folk theories. (Muthukrishna & Henrich, 2019, p.221)

Without a theoretical framework, there is a largely unbounded set of possible "research hypotheses", even mutually inconsistent ones, some of which end up, if only by chance, validated by statistically significant results. Muthukrishna and Henrich (2019) hold that "having too many questions" leads to too little rigour spread too indiscriminately and ultimately an incoherent patchwork of results that amounts to no clearly identified disciplinary scientific advancement.

An econometric point underlying this concern is that a missing variable or other unintended bias in a model can "assist" the researcher to obtain a low *p*-level, such as by (for example) biasing a regression coefficient upwards, even when in reality the true regression coefficient is zero. Missing-variable bias can have the same effect as an increase in power, i.e., an increase in the probability of rejecting a false null hypothesis, but it can also have the effect of increasing the probability of rejecting a true null hypothesis. It is common for empiricists to attribute failure to reject the null to insufficient statistical "power", but it is less often emphasized that an observed rejection might in fact be rejection of a true null brought by bias in the model.

In other words, "wrong" or inherently biased models can be, and must surely often be, the source of "statistically significant" results, especially if models are mended or slanted so as to achieve a significant result. Since a *p*-level is a conditional probability, and is conditioned on the conjunction of the model and the null hypothesis *both* being true, it is as much evidence against the model as against the null hypothesis. That is usually forgotten when the aim of the research in the first place is to reject the null hypothesis, treating the associated model as given.

An issue in replicating an effect is whether it must be replicated using the same model. A bias in that model can make it appear as a replicable (i.e., repeatable) effect even when it does not really exist. Conversely, if it truly exists under one model or set of control variables but does not hold under another, the issue is whether it really exists in any practical application (e.g., will a drug work on patients or only work under narrow unrecognised conditions or interactions?).

4. Significance testing and research culture

There has long been a silent majority of accounting academics who feel that accounting as a discipline lost its relevance to the accounting profession when empirical research methods came to dominate *The Accounting Review* and other top ranked journals. That disquiet has only strengthened with time and publication of ever more of the same. The recent discussion paper by Schrand (2019) opens up to this complaint and lists some fundamentals that should change in the research and publication culture. The starting point is a statement that was made by AAA President David Burgstahler to provoke debate at the 2017 AAA meeting:

[6] R.A. Fisher, who invented many of the concepts in statistics, spoke of repetition and replication separately. Replication in the modern sense is what Fisher meant by repetition, and requires that the same experiment is run under different conditions, producing new data, but leading to the same conclusion. In Fisher's case, experiments and their statistical tests were in agriculture, and different experimental conditions meant different weather, soil, location and possibly different "treatments" such as fertilizers. Replication to Fisher meant a subset of the experiment such as a small garden plot with two plant types or treatments compared side by side, replicated by exact copies of each other in different parts of the garden, all within the same experiment. One experiment might include many replicas. In modern usage, the word "replication" is used in the sense of repeating the test under new conditions and possibly with better experimental control. In science, the assumption is that new data is drawn from a new experiment. See Hail et al. (2020) on different possible definitions of "replication" and "repetition".

> There is a broad perception that, despite the large volume of accounting research produced over the last half century, there has been relatively little impact on practice and public policy, and only limited lasting impact on the development of accounting thought. (AAA President David Burgstahler quoted by Schrand, 2019, p.11)

The Schrand (2019) diagnosis is that empirical "hypothesis testing" research based on easily available databases and pre-programmed software became fashionable and readily publishable. PhD students, often with little background in statistics, mimicked what they saw as the way to a career in academic accounting:

> . . . studies were viewed as more rigorous if they contained a hypothesis development section. Referees reject papers that observe practice, claiming the paper is "merely" descriptive. Doctoral students and junior faculty see the journal content and perceive "hypothesis testing" papers as the most publishable. Over time the concentration in hypothesis testing research intensifies. (Schrand, 2019, p.14)

Issues over the same ubiquitous hypothesis/significance testing have seethed in psychology since at least the 1990s and are now being raised across the social and medical sciences. There are now dozens of papers alleging that mechanical and opportunistic adherence to frequentist *p*-levels is a primary source of the "replication crisis" arising across the empirical sciences; e.g., Hunter (2001), John, Loewenstein, and Prelec (2012), Lambdin (2012), Amrhein, Greenland, and McShane (2019), Greenland et al. (2016), Boos and Stefanski (2011), Nuzzo (2014), Gelman and Loken (2014), McShane, Gal, Gelman, Robert, and Tackett (2019), Benjamin et al. (2018) and Begley (2013). The literature and extent of informed and expert cross-disciplinary discussion of this issue on scholarly blogs and in the popular media (Anvari & Lakens, 2018) is now too large and fast growing to survey.

Of all the statistical social sciences, the discipline of accounting should not be expected to lead the debate over significance testing, but should expect to confront the same issues of statistical logic and good practice that are now being debated openly and rigorously in psychology and statistics proper. The historical development of accounting as a research field is relatively recent and is not as heavily funded and advanced as a socially and economically vital field like medical statistics, but is just as much a user of statistical research methods.

Now that doubts about the reliability and integrity of empirical research programs are openly aired in many scientific disciplines, both in statistics proper and in applied statistics, empirical accounting research has been stirred into facing similar questions, fuelled it should be said by a genuine interest among some established empiricists (e.g., see Bloomfield et al., 2018; Hail et al., 2020). See also the earlier commentary on the significance testing cult by (Basu, 2012).

In accounting, along with other fields, there have been known episodes of outright faked empirical research, resulting in retractions (see *https://retractionwatch.com/*). The ability of researchers to use statistics to conjure and "document" results and narratives that are accepted into leading JAR/TAR/JAE type journals and later found to be fictions is apparently an undeniable sociological ill that has shaken some empiricists' long solid rejection of "the significance testing controversy".[7] In effect, a minority of fakers and opportunists pushed the empirical game in many social sciences too far.

The willingness of some researchers to game the publication and rewards mechanism so blatantly is their way of saying some of the same things that critics say: e.g., "it's not important, no external reader's actions or results depend on this being true, no one knows or cares", or "it's all just a publication and promotion game, referees will never know the difference, everyone does it one way of another", etc.

The essential concern is whether the overall body of accumulated empirical evidence, produced and summarized using orthodox "frequentist" hypothesis tests, is as substantive, well founded, "scientific" and reliable as has conventionally been assumed and "sold" to potential external users and other researchers by its proponents and their usually heavily indoctrinated PhD students. The dissenting view, now put officially and publicly by the *American Statistical Association* and groups of highly eminent statisticians, and first raised in multiple works in the 1960s and before (e.g. Rozeboom, 1960; Bakan, 1966; Meehl, 1967; Edwards, Lindman, & Savage, 1963; Lindley, 1957) – albeit unmentioned and unseen in accounting by generations of PhD students – is that significance tests are flawed in logic and effectively "designed to be abused", either by intention or by unknowing misuse and misunderstanding within communities of researchers whose overriding incentive is publication. See Hopwood (2008) "career instrumentalism".

Such "abuse" is not necessarily the intent of the researcher (Simmons et al., 2018a, p.518). If a technique is inherently weak or unreliable in its logic and conventions, at least in typical applications, then its widespread and virtually compulsory adoption will naturally lead to much wrong or weak evidence and publication, regardless of the researcher's intentions. That overall deficiency is the most general complaint that has been made by Bayesians (e.g., Lindley, 1987) regarding significance tests – they give out "significant results" too easily (see earlier example) and they don't impound the strength of evidence that is presumed of them and alluded by phrases like "the evidence is statistically significant".[8] Many users of significance tests just don't know of this criticism and were never told anything of the sort.

It is ironic that the evidential meaning or weight of evidence carried by sayings like "significant at 5%" or "statistically significant" is now openly questioned, because empiricists have always used those words as their egoistic logo, and often made the point when flaunting their own scientific rigour that evidence is credible if and only if it is statistically significant.

[7] This controversy dates to before Morrison and Henkel (1970) and therefore before accounting first used significance tests).

[8] This is relative, Bayesian methods are also not fully immune to "hacking". See Johnstone (2018) on how Bayesians can attempt to "sample to a foregone conclusion".

Although Bayesian statistical logic is implemented routinely in analytic "noisy rational expectations" accounting information theory, there has, at least until very recently,[9] been no similar application of Bayesian statistical mechanics in empirical accounting research or research training. Beginning with Jeffreys (1939), Lindley (1957), Savage (1954), Bayesian logicians rejected *p*-levels not only for their illogic but also because they are easily abused (a practice recently now known as "*p*-hacking" after Nuzzo, 2014). These charges come from the same statistical school of thought as the Bayesian information theory models that Feltham, Demski and others brought to accounting information theory. But that's where Bayesianism ended in accounting – the politically charged part of Bayesian theory that rejected *p*-levels, and decried their widespread abuse in social sciences, has for decades gone virtually unknown in accounting literature, despite being often debated in empirical econometrics (see e.g., Leamer, 1983; Lovell, 1983), empirical psychology, sociology, health and medical sciences, and other research disciplines.[10] See especially the highly authoritative and early critique by Edwards et al. (1963).

The most critical points of statistical logic and practice that Bayesians have aimed at *p*-levels are now beginning to be aired in accounting literature, but with deep-seated resistance. In recent times, James Ohlson, the highly eminent developer of analytical accounting valuation theory, whose considered thoughts on research cultures might be expected to be welcome at any accounting journal, has attempted to publish papers that deal directly with what's questionable about significance tests and their use. One paper has been published in *Abacus* in 2015, and (as of August 2020) is cited 30 times albeit largely by other critics (e.g., by Dyckman, 2016; Kim et al., 2018; Dumay & de Villiers, 2019), but not by any leading US accounting empiricist.

More disturbingly, in terms of how conservative sectors of the accounting research community evade the issue, Ohlson's (2019) critique was "desk rejected" at both the *Accounting Review* and *Accounting Horizons*. Papers can be rejected for many reasons, but desk rejection of salient methodological commentary by someone of Ohlson's stature and research experience would seem to imply that there are few readers or accounting empiricists open to this level of methodological cum sociological and philosophical critique, or alternatively there are at least some highly influential readers who would take umbrage and prefer to silence debate. Also, since both journals are *American Accounting Association* journals, such summary dismissal amounts to an "official" censoring by the accounting academic community.

The past unwillingness of empirical accounting researchers to engage with the critique of significance testing in other disciplines, including nearby econometrics, has long been a red flag indicating poor science, political censorship and scientific immaturity in accounting empiricism. Significance tests have been the accounting empiricists' tool of choice, and bread and butter, for 50 or more years, and must surely therefore warrant methodological appraisal, especially when they are criticized so deeply in so many other disciplines and by renowned theorist-statisticians, including not only Bayesians but also eminent frequentists.[11] In any field where results should actually matter, rigorous questioning of research foundations should presumably occur naturally. That is the most troublesome aspect of accounting's failure to front.

The empirical fact is that readers cannot go to JAR/TAR/JAE and find papers published in the last 50 years of significance testing that discuss what's wrong with significance tests or give any indication of the vast literature outside accounting, and in leading statistics journals, that shows in theory and by example that there is a lot wrong with significance testing. The obvious conclusion is that these journals do not entertain critique of their own foundations. The recent JAR re-positioning (Bloomfield et al., 2018; Hail et al., 2020) indicates that this long period of wilful neglect or censorship is no longer a united front, at least not under the lead of some influential insiders who are willing to at least modify the JAR/TAR/JAE culture and rules.

It will be clear from how seriously the replication crisis and *p*-level controversy is now being taken in more established empirical research disciplines (e.g. medicine, psychology) that accounting as a discipline will ultimately face the same criticism met by any social and behavioural science that has the same inherent barriers to reliable research, including the micro factors listed above and basically boiling down to causal complexity, non-stationarity, inability to control experiments and lastly a research motive that prioritizes or rewards publication per se, regardless of virtually anything else apart from perhaps the virtue of opening further avenues for further publication by others.

In a considered historical overview of research in accounting, Zeff (2019) arrived inexorably at the replication issue:

> In 1994, *CAR* Editor Michael Gibbins introduced a special section entitled "Improvements and Updates," which was intended to encourage replications and extensions. Yet the response by researchers was tepid. Gibbins' successor as editor, Steven E. Salterio, 2014, 1134, in a thoughtful article on the subject of replications in social science research, observed that *CAR*'s call for replication research "was met with a resounding 'thud' by the academic community." The section lasted but two years (Gordon & Boland, 2015: 474). Philip Brown (2013: 856) labels the actions of editors and reviewers who discourage replications as a signal of their "disciplinary immaturity." (Zeff, 2019)

That failed effort at CAR, likely prompted by Lindsay (1994), Lindsay (1995), was before its time in the sense that accounting was never positioned or motivated to front run the social sciences into recognising a "replication crisis".

There is no doubt that many inside the stronghold of US empirical accounting research note what is happening in other disciplines. Hints of this appeared in the inaugural issue and editorial statements of the "fresh approach", but apparently

[9] e.g. see Breuer and Schütt (2019).
[10] e.g. see the readings in Morrison and Henkel (1970)
[11] D.R. Cox (1982, p.327), who is as strong as any statistician of his pre-eminence in supporting classical frequentist testing as a legitimate technique, wrote "It is very bad practice to summarize an important investigation solely by a value of P (i.e., *p*-level)".

faltering in that ambition, *Journal of Financial Reporting*. One of the editors, Welker (2016) stated that the replicability of empirical research findings in accounting is a looming question, prompted by some low replication rates seen in other disciplines, and also that models with missing variables may often produce results (low *p*-levels) that don't stand up once controls are improved.

Methodological and sociological questions over the use and validity of *p*-levels have lately been raised on the boundaries of accounting. Issues to do with how the sample size greatly affects the evidential meaning attributed to an observed significance level have come to light, mainly via being aired in finance where very large sample sizes are now commonplace (e.g., Hoepner, McMillan, Vivian, & Simen, 2021; Harvey, 2017; Harvey & Zhou, 1990; Harvey, Liu, & Zhu, 2016; Chordia, Goyal, & Saretto, 2017; Kim & Ji, 2015), but also in some accounting literature; see Basu (2015), Basu and Park (2014), Ohlson (2015), Ohlson (2019), Dyckman (2016), Dyckman and Zeff (2014), Dyckman and Zeff (2015), Dyckman and Zeff (2019), Kim et al. (2018), Stone (2018) and Khan and Trønnes (2019). Earlier ad hoc critiques include Lindsay (1994), Lindsay (1995) and Johnstone (1995), Johnstone (1997). See also Bailey, Hasselback, and Karcher (2001). Lindley (1957) paradox is now being witnessed and discussed by finance researchers whose ultra-large samples expose what Lindley explained theoretically; i.e., that in a large sample you need to observe only a tiny "non-zero" and manifestly immaterial effect to be able to claim and report a "significant" *p*-value less than 0.05.

The wider problem is that "statistically significant" spurious correlations exist in infinite ways, either systematically or by a fortunate chance sample, yet are open for interpretation as "significant" evidence of a causal or replicable relationship:

> ... Bakan (1966) subdivided the data of 60,000 persons according to completely arbitrary criteria, like living east or west of the Mississippi river, living in the north or south of the USA, etc. and found all tests coming up statistically significant. Waller (2004) examined the personality questionnaire data of 81,000 individuals to see how many randomly chosen directional null hypotheses can be rejected. If sample size is large enough, 50% of directional hypothesis tests should be significant irrespective of the hypothesis. As expected, nearly half (46%) of Waller's (2004) results were significant. Simulations suggest that in the presence of even tiny residual confounding (e.g., some omitted variable bias) or other bias, large observational studies of null effects will generate results that may be mistaken as revealing thousands of true relationships (Bruns & Ioannidis, 2016). Experimental studies may also suffer the same problem, if they have even minimal biases. (Szucs & Ioannidis, 2017, p. 10)

Random yet formally "significant at 5%" associations would be less open to interpretation as causal if publication standards required: (i) more "deductive" and less rhetorical derivations of why such a causal relationship should exist, or (ii) replication or at least re-testability[12] of the observed effect with new data drawn under different conditions, (iii) reporting of confidence interval estimates of observed effects so as to prevent exploitation of the Lindley paradox by which tiny effects are "statistically significant" when samples are large (see the simple explanation in Johnstone, 2018, pp.42–44), (iv) testable predictions of future values of the variables in question.

Neither remedy (i) nor (ii) is likely. First, many hypotheses in accounting research are conjectures derived via an often post hoc adjusted story line rather than pre-study theoretical anticipation of likely results. Second, replication would likely occur if the effect observed had some overtly valuable application and potentially direct economic value, but results like that rarely exist in empirical accounting research, with even their own producers losing interest in any replication after publication. Remedy (iii) requires transparency by reporting of confidence intervals and is one of the oldest and most widely suggested antidotes to the misuse of significance tests, but has proven itself unpopular with researchers by simply not happening. Lastly, tests by prediction and ex post verification of predictions are not institutionalized "normal science" in accounting, despite the "explain and predict" ethos of physical science.

Being latecomers to statistical research methods relative to similar applied disciplines like psychology and medicine, accounting researchers in the 1970–80s naturally took on the existing automated statistical software, which at that time was entirely frequentist, and made a large personal investment in tooling up. Accounting empiricists from then on were committed to the orthodox frequentist methods and language, and each new crop of PhD students was instructed and inculcated in the frequentist empirical accounting research methods that remain stock standard in empirical accounting PhD programs today.

Not only does accounting as a research discipline have decades of sunk cost in its training of researchers in frequentist methods. It also has an empirical research culture that in many schools has been highly successful in material ways, including obviously in producing elite journal publications, placements of PhD graduates in high salary jobs, lucrative promotion and tenure decisions, the generation of research funding and all the related honours and consultancies that flow from these status symbols. It is natural that outcomes such as these appear to justify their means. Even the mere physical appearance of empirical research, with its technical language, symbols, models and aura of science, is seductive enough to draw respect and believers. When combined with the associated personal rewards, any reconsideration is unlikely to be a priority or seem beneficial.

For deeper consideration of how natural sociological attractions in accounting drive what is published and accepted, see Williams and Rogers (1995) and Chua (2019). One driver, more at play in accounting and finance than in other disciplines, is

[12] Empiricists cannot be required to replicate others' works unless they warrant that work, but should provide sufficient detail in their own papers to allow rigorous retesting and verification. Often that is not the case.

the profit generated for all the vested interest groups around and dependent on publication as a game, including publishers' subscription fees, authors' (re)submission fees and also the very large international and local student fees flowing to highly ranked research active finance and accounting schools (Moizer, 2009).

Moizer (2009) gives a detailed explanation of the publication process in the social sciences as a game between authors, referees and some elite journal editors, many of whom have a joint interest in achieving at least the appearance of scientific rigor and advancement. The frequentist hypothesis testing regime suits this objective, not merely because it is a mathematically sophisticated apparatus,[13] but especially because it produces so-called "significant results" very easily out of data that is often not so obliging on a more coherent statistical examination. Typically, for example, the 95% confidence interval associated with a 5% *p*-level in a large study shows virtually zero observed effect, but that result is reported instead as a 5% *p*-level rather than more revealingly as a confidence interval centred tightly next to zero (see example calculations in Johnstone, 2018, pp.42–44).

Much of what has been taken as rigorous or acceptable empirical research practice is now coming into question. The "replication crisis" has brought a sea change across multiple disciplines in the way that common practices are viewed (Simmons et al., 2018a; Staddon, 2017). In finance, which is before accounting in recognising issues surrounding significance tests, Harvey (2017), Harvey (2019) recognises "soft misconduct" where researchers seek out a "significant" effect by strategic sample selection (start and end dates), choice of when to stop sampling, winsorization or exclusion of "outliers", arbitrary variable transformations, model selection, choices of control variables, the choice of model fitting criteria and the selective (non)reporting of results. Researchers will typically go to great lengths when the referee's favorable report calls for "robustness checks" or extra analysis. The incentive, when publication is so near and sunk cost is so high, is to "find" what is expected (Ohlson, 2019).

In an unpublished note called the "garden of forking paths", Gelman and Loken (2013) maintain that even if we accept that a "fishing expedition" for low *p*-levels is not a fair representation of most empirical research, although how we could we be sure of that when significance fishing is not likely to be advertised, a replication crisis can still arise. They hypothesize that the replication crisis has occurred because each published study is based on many "random" or discretionary methodological choices along the way, often midstream as new results point in different directions and reveal potential experimental needs (e.g. drawing different data, better experimental control, sample stratification, new hypotheses etc.) and is therefore inherently "biased" or directed towards the desired realized "statistically significant" outcomes, outcomes that differ from those that would have arisen had the researchers taken other equally justifiable methods and choices.

Hence bias occurs inherently even when the researcher does not intend any ultimate publication bias or convergence to significance.[14]

> *P*-hacking is a pervasive problem precisely because researchers usually do not realize that they are doing it... (Simmons et al., 2018a, p.518)

In an extreme example, if statistical significance or "nice" results are not easily forthcoming, the whole project might be dropped or completely restructured, which of itself, possibly unintentionally and with no aim other than to move on and learn something new, misrepresents to the outside world what has not been found (the "file drawer problem"). Conversely, when results turn out significant, the researcher feels rewarded and correct (Gelman and Loken, 2014), suffering perhaps from the illusion of control effect known in psychology.

The absence in statistical enquiry of a program of replication is itself a source of bias. The usual argument against replication has been that we want "new" knowledge. But the underlying reason is likely that replication of findings that are of interest only to insiders, and not the outside world, has a loss function by which successful replication brings little of value, and failed replication brings many costs, political and real. By not replicating published results, researchers can unknowingly or opportunistically build an impressive looking literature of results that are, if tested more rigorously, often untrue.

Many accounting researchers have for decades attended weekly empirical research seminars that "sound and look the same", during which era there have been many over touted empirical discoveries (e.g., efficient markets can "see through" accounting policy choices), often based on a "maintained hypothesis", and also often an overt self-satisfaction inside the coterie of self-proclaimed "scientific" accounting researchers. That sameness and general lack of obvious scientific progress, combined with the wide indifference towards empirical accounting research from any group apart from the researchers themselves, has led to widespread cynicism:

> Some people may even argue that the current state of affairs is bizarre: the abundance of "validated" stories has been growing exponentially over the years. While the RQ [research questions] have become increasingly esoteric, as far as I can tell no evidence suggests that researchers have found it increasingly difficult to validate the RQs. Rejection of the null hypotheses, when desired, generally pose no recognized problems, and, as on cue, robustness tests always work

[13] Bayesians do not dispute frequentist mathematics, instead they dispute the inductive logic that is applied within that mathematics and the way that the calculated results (e.g., *p*-values) are interpreted.

[14] Simmons, Nelson, and Simonsohn (2011) note that these choices are within the researcher's "degrees of research freedom", which is a clever analogy because in frequentist statistics the sampling distribution of the observed result must be conditioned on the data's degrees of freedom.

> out per want. From what I can tell, robustness tests never overturn the paper's basic conclusion and researchers can so announce with forthright satisfaction. Even more amazing, I have never heard of any paper finding that some effect is not economically material. . . . Few members of the community get fooled: to a considerable extent papers are not taken all that seriously. Individuals who attend seminars on a regular basis end up being less and less impressed by the totality of the research effort. It would not be easy to argue convincingly that the great bulk of papers published generate much interest (setting aside captive audiences). (Ohlson, 2019)

The ingrained conservatism in accounting teaching and doctoral training is to be expected. Sunk costs and cosy equilibria take hold in all research communities (see the sociological explanations and references in Williams and Rogers (1995) and Chua (2019)). Accounting researchers have historically eschewed or simply not known about the Bayes versus "frequentist" logical and philosophical debate, choosing in effect to proceed on the basis that conventional frequentist methods have the one imprimatur that matters, namely publication in one of the high-end US accounting journals.

In fields like medicine, where the consumers of empirical evidence (e.g., drug companies) can be left with large legal liabilities after relying on wrong science (e.g., abuse or misinterpretation of *p*-levels), the motivation to take account of the weaknesses of conventional statistical methodology is obvious. It may well be an indicator of how much apart from journal publication rides on the truth of empirical research findings as to whether researchers question the veracity or reliability of their own statistical methods, and whether they consider possible evidential interpretations beyond the mechanistic "H_0 is rejected at 5%" and the like.[15] Economic rationality would demand that potentially valuable results are tested and replicated by practical measures before use.

Accounting researchers' imperviousness or opposition to the significance test controversy is a worthy subject of itself for sociological study and explanation. There is undoubtedly potential for a positivist explanation of the behaviour of accounting researchers. We can imagine a "market for statistical significance" (Johnstone, 1988) within a "positivist theory of accounting researcher*s*", that would be based on the same agency self-interest framework used by accounting researchers to explain the self-interested behaviour of managers and firms.

Perhaps the most remarkable reaction of all to the attack on *p*-levels and the empirical research factories that have been built all over the social sciences came from Professor Fiske (a Princeton University psychology researcher and past-president of the APA) who characterized the more effective commentators on the un-reproducibility in empirical psychology as akin to "methodological terrorists", and called for their criticism to remain private, or be made through the same journals' reviewing processes and standards as the very ones being criticized, rather than discrediting the field publicly and without any scientific "decorum". That reaction drew strong rebuke from the highly eminent statistician Andrew Gelman (2016), who called the psychology research culture based on poor statistics a "dead paradigm" and accused editors of not retracting studies that were known to be wrong.

In the rethink and retribution now underway across both social and hard sciences over *p*-hacking and irreproducible results, the question that arises is why so much unreliable and un-replicated empirical work happened and was published. There are suggestions from professional statisticians that one simple answer is that data bases and statistical software became available for downloading and statistical software (like SPSS) became too easy to use, requiring only the "click" of a menu. In earlier times, statistical software was run only by trained statisticians with computer programming skills, so there was a natural and economic rationing and quality assurance in what work was done. In accounting, beginning research students were quickly able to get empirical results up and running, and write up a paper or thesis by emulating the style and language of similar work in journals. At the same time "empiricism" became the vogue style of research and, in some schools, little else was on the agenda or tolerated.

Around the mid-1980s, at the University of Sydney where I had just finished a PhD thesis on the foundations of significance testing, some academic staff who did not do empirical/statistical research were encouraged by senior faculty to look elsewhere for their futures. The 1970–1980s empirical zeal was at a high point in Australia at that time, and virtually anything "empirical" (i.e., typically regression analysis) was valued more highly than anything "normative". Typically, for example, arguments about how to value assets or measure income, which meant something to the profession, were viewed by empiricists as old hat and "not scientific".

One conversation from that time sits in my memory. I had never believed in the reliability of the empirical work that I did in my last year as an undergrad. I had seen how I could manufacture different looking results by changing the variables in a regression equation or the sample size and period. We had run many regressions in my undergraduate econometrics major and I had noticed how my teachers would react very credulously and favourably when the results I showed them fitted their preconceptions (e.g., variables having the "right" sign and statistical significance). It was as if results matching expectations validate the process by which they were obtained. In my PhD work, out of curiosity and a wish to better understand statistical evidence and its weight, I had read nearly all the existing literature in statistics and philosophy of science on

[15] Interestingly, there was a short-lived attempt in the *Accounting Review* in 1987 towards explaining classical hypothesis tests in a Bayesian way. Burgstahler (1987) made no mention of what was already known in statistics about this very problem, and gave a conclusion that was wrong by its own Bayesian logic and the opposite of the accepted Bayesian analysis. The preceding paper in the *Accounting Review* by Kinney (1986) had the same limitation. See the re-analysis of Burgstahler (1987) in Johnstone (2018).

what was right and wrong with "classical" (frequentist) statistical hypothesis tests. At the end of that study, I was effectively in my own world[16] in the sense that when I told anyone who asked about the disdain many professional statisticians and logicians had for common significance tests there was generally incredulity and often annoyance. However, on this one occasion, a researcher, who was not an empiricist and whose career was in question, became frustrated enough with my liberation to blurt "... look, what are you saying, are significance tests bad in principle or just bad in the way that they are used?". My too immediate answer was "both", at which point he simply raised his eyebrows as if to say that surely that cannot be right. It has to be remembered that anyone in accounting at that time who had ever studied any level of statistics had been taught regression and frequentist hypothesis testing in a totally uncritical and usually naïve way, and usually absorbed what they learned as simply "right", much like they had learned to view arithmetic and mathematics in their childhood educations.

Those events, in hindsight, say much about the prevailing research culture in accounting at the time, and also that essentially all of the current uprising over past empirical research methods and *p*-hacking has been a very long time coming. Virtually every modern criticism of *p*-levels and *p*-hacking was already in the literature by the 1980s, for anyone to absorb. But the times did not welcome that level of methodological introspection. Papers like the Campbell Harvey (2017) Presidential Address to the *American Finance Association*, in which he brings the spectre of *p*-hacking and unreproducible empirics into the direct view of the "empirical finance" community, were 30 years over the horizon.

In an invited Editorial in the *Journal of Portfolio Management*, López de Prado (2017) put a list of what's wrong with publication-driven empirical finance research programs in Universities. The overall criticism is that empirical finance is a University tenure competition driven by the currently prevailing academic "conclaves" rather than a developing empirically testable science:

> Empirical finance is a unique discipline. First, academics cannot repeat experiments; hence, their discoveries cannot be independently validated. For all we know, the great majority of the discoveries published may be false because of selection bias. ... Second, unlike physical phenomena, financial markets are an adaptive system – the product of changing regulations, institutions and agents. There are no positive laws to be discovered. (López de Prado, 2017, p.5)

It has taken 50 years of a methodologically and culturally flawed empirical research apparatus to produce enough examples of the significance testing routine, rehearsed over and over, to prompt a change in how we feel about more of the same.

The one modern event that has precipitated the current change of base sentiment towards significance tests is "big data". It has been known since Berkson (1938), Nunnally (1960), Meehl (1967) that a very large sample size is virtually guaranteed to make the smallest physical effect "statistically significant" (Szucs & Ioannidis, 2017). Finance now has sample sizes routinely in the hundreds of thousands of observations, and finance researchers are often themselves frustrated with "everything coming out significant". That issue, which is essentially what underlies Lindley's paradox, was not apparent to earnest and often largely untrained empirical researchers when they were using much smaller samples, when getting a "significant" result was not a certainty. It was much easier when samples were smaller to retain faith in the significance testing method, because it was a little harder to get a significant result (harder might of course mean that it took more trial and error *p*-hacking).

There is a push now from journals with their lifeblood in empirical research to clean up, or shore up, the empirical research protocol and political dominance. One increasingly accepted enterprise is to promote contractual pre-registered research designs and hypotheses, so as in principle to prevent *p*-hacking and post hoc story telling around "significant results" (Nosek, Spies, & Motyl, 2012).

The JAR 2017 conference was based on this attempt at reform, and implicitly made admissions regarding what is seen to have often been a publication-seeking rather than truth-seeking research culture:[17]

> This process encourages researchers to engage in innovative research and to gather new data because the realization of the data and results of the analysis do not affect the evaluation of the Registered Report. Authors of accepted proposals can undertake the hard work of data gathering without concern that their Registered Report will be rejected simply because the data did not support their hypotheses. The process also helps readers evaluate the strength of the empirical evidence being reported, because the authors have no ability to modify their hypotheses or planned analyses after observing their data, and have no publication incentive to distort unplanned analyses in order to support their hypotheses. This process further encourages replications of well-known results with new data that have since become available. (Call for Papers, JAR Conference, 2017)

[16] Publications from my thesis were in statistics and the philosophy of science; e.g. Johnstone (1986, 1987a, 1987b). My PhD was examined by Bayesians Lindley and Pratt along with philosopher Henry Kyburg, and later I published several papers with Lindley. PhD research, like mine in Australia in the 1980s that is so far "off point" would nowadays be seen as "risky" or worse. Typically work like this, which questions the neo-classical economics rational/empirical movement in accounting (e.g., Williams, 1989, Ravenscroft & Williams, 2009, Williams & Ravenscroft, 2015) has been dismissed by mainstream insiders, typically without response, as "unhelpful" and disrespectful. That high-handed refusal to engage with critical philosophical commentary is now revealing itself as part of why there is a "replication crisis" in the empirical social sciences, amounting to a crisis of research method, ethics and malpractice.

[17] In medicine, Kaplan and Irvin (2015) found that by having authors pre-register the primary hypotheses in clinical studies, the proportion of positive findings fell from 57% (17 of 30 studies) to 8% (2 out of 25 studies).

The concession underlying preregistered studies does not say anything critical other than that empirical research should ideally have been built on better researcher behaviour, or on better enforced or motivated research protocols. Nonetheless, that is a major back-pedal in what has always seemed a highly self-satisfied research culture.

The JAR editors and team, to its credit, has rocked the boat by attempting to build a preregistration protocol that improves empirical research customs and integrity. However these are changes that preregistration can't necessarily or easily ensure. In regard to researchers' behaviours, a preregistered study might have been preregistered only after some initial data snooping and analysis, which would imply that preregistered studies are not an objective, random or unbiased subset of possible results on the topic in question. In fields where preregistered experiments were first used, much of the work proposed could not possibly have been "sampled" already, because the fixed cost, and the laboratory equipment and work required, do not allow any meaningful data collection without funding and approval. That is not so when researchers have the necessary archival data and software at any time before designing and proposing a study.

Beyond issues surrounding researchers' behaviour and research integrity, how can preregistered studies overcome the far greater issues of: (i) logically flawed statistical measures of evidence, (ii) limited or no theory underlying the researchers regressions or models, (iii) the innate limitations in much of the social sciences of highly noisy and nonstationary data, gathered with little possibility for experimental control and often only via proxy variables rather than direct measurement of the variables of interest? A very complicated, largely unmeasurable and noisy statistical world cannot be overcome without much more developed theory (often unlikely to eventuate) and greater experimental control (often physically impossible). The following comment by leading "critical theorist" and "critical statistician" Andrew Gelman (2018), relating to some earlier analysis, captures the inherent problem:

> Forking paths and *p*-hacking do play a role in this story: forking paths (multiple potential analyses on a given experiment) allow researchers to find apparent "statistical significance" in the presence of junk theory and junk data, and *p*-hacking (selection on multiple analyses on the same dataset) allows ambitious researchers to do this more effectively. . . . in the absence of forking paths and *p*-hacking, there'd be much more of an incentive to use stronger theories and better data. But I think the fundamental problem in work such as Wansink's is that his noise is so much larger than his signal that essentially nothing can be learned from his data. And that's a problem with lots and lots of studies in the human sciences. The forking paths and *p*-hacking are relevant only in the indirect way that it explained how the food behavior researchers (like the beauty-and sex-ratio researchers, the ovulation-and-clothing researchers, the embodied-cognition researchers, the fat-arms-and-voting researchers, etc etc etc) managed to get apparent statistical success out of hopelessly noisy data. The indirect role of preregistration etc. is potentially important, but I don't want people to think that if they just preregister (or, more generally, that if they just act virtuously) that research results will just stream in. (Gelman, 2018)

Gelman's description sums up what I have seen over the last 40 years in much empirical accounting research. Specifically, trained empirical researchers assume that merely by following their training, much like cooking from a recipe book, they will be protected from drawing unjustified inferences. In their PhDs and later work, researchers learn to run statistical software, enter data and get results, and then presume that this "accepted" approach – i.e., merely being "empirical", running statistical methods with large samples on fast computers and generally emulating other empirical researchers – can by its own virtues produce reliable conclusions in whatever context or research question. All that the researcher needs to do, albeit not without much work and often with disappointments, is to get the results that fit the loosely developed theory and then reproduce the computer output tables of results and *p*-levels in a paper that follows the well worn journal format and language. That research script, if acted out well, will ensure that "new scientific knowledge" has arrived, and of course publication.

Such automation and mechanisation of inference would not be assumed by the same researchers if they found themselves on jury duty. In that analogous task, i.e., one of inference under uncertainty, they would concern themselves with what the evidence means or says, how strong it is, and whether there is enough meaningful information to ever come to a clear conclusion in a complicated set of circumstances. Yet in accounting empirical research, the stated *p*-level (and the underlying model, whatever its lack of theoretical derivation) is taken to excuse the researcher from ever wondering out loud whether the evidence represented means much. Its strength is summarized and proven by the formality of it being "statistically significant".

An innate problem for accounting research is that accounting phenomena involve human behaviours and social and political interactions (in firms, stock markets, corporate regulation, etc.) which are much harder to model (explain and predict) than many of the physical phenomena that are studied in "hard" sciences. For example, the ways by which firms choose accounting policies would seem far more debateable and changeable (statistically "non-stationary") than whether a medicine works or whether some genetic characteristic or environmental exposure is unhealthy.

Some of the problems studied in accounting – e.g., whether a firm's better financial disclosure practices impress the stock market so much that investors offer the firm a lower cost of capital – are so complicated theoretically and practically, and so hard to model and test empirically, that the philosophical question has to be raised of whether any result in such a study can be given great evidential credence. I raise this one particular research question (information and the cost of capital) because it is often described as one of the most interesting and important questions in accounting, and also because in my own survey of published work on that topic I find that authors repeatedly describe the results of past empirical research as

being "mixed", that word being apparently the accepted euphemism for having reached different conclusions in different studies and hence no particular or reliable conclusion.

Reading the related empirical accounting research literature for the purposes of Johnstone (2016), Johnstone (2018) I was struck by the reappearing description of empirical results as "mixed". I started to keep a list of twenty or more related papers using this term after I realized that "mixed" was not a word that occurred so regularly by accident. Instead, it appears to be the word "on call" in the empiricists' vocabulary to say that there are indeed results, certainly not a dead-end, but not the "right" results or not consistent results. By describing results as "mixed" rather than as contradictory or unreliable, empiricists avoid saying that there is no result other than a lot of messy data and statistics, or that the different studies in the related literature might be producing noise rather than one replicable conclusion.

When embarking on an empirical research study, there is no guarantee, no matter what empirical research protocols are in place, and no matter what Bayesian or non-Bayesian statistical methods are applied, that there will ever be a rigorous and reliable research finding. That point is not usually emphasized in PhD methods training. Rather, one empirical research question is portrayed as the same as another, each existing to be answered to the same evidential standard, to a referee's satisfaction and to the same scientific and publication end by the same statistical methods and stylized write-up. The innate presumption that empirical progress is always possible is empowered by how easily conventional significance tests produce "statistically significant" results. A side benefit of tests that yield significant results very easily, identified by Chua (2019, p.9), is that empirical research careers are possible for nearly anyone with the will and a positivist training manual, and a PhD from a "top school".

After the preregistered "REP" papers in the JAR 2017 conference were published, Bloomfield et al. (2018) wrote a partly apologetic (to JAR empiricists) appraisal suggesting that papers published by this mechanism lacked the "polish" of those that went through the more usual "TEP" review and revision route before acceptance:

> We find that REP increases up-front investment in planning, data gathering, and analysis, but reduces follow-up investment after results are known. This shift in investment makes individual results more reproducible, but leaves articles less thorough and refined. REP could be improved by encouraging selected forms of follow-up investment that survey respondents believe are usually used under TEP to make papers more informative, focused, and accurate at little risk of overstatement. (Bloomfield et al., 2018)

Rather than being so obviously desirable, the follow-up polishing or "investment" described here can often be the source of unreproducible research. Using what Gelman and others call "researchers' degrees of freedom" (i.e., discretion to change models, data sets etc.), the natural incentive of researchers is to satisfy reviewers by presenting a clean, objective, and robust looking set of results. That can mean wallpapering over cracks and generally stretching both the bounds of scientific practice and ultimately the empirical truth.

In his critique of significance testing practices and "star gazing", Ohlson (2015), Ohlson (2019) asks how rarely do we see in print a statement that says that the author was asked to do robustness checks and found that they were unfortunately *not* "qualitatively similar" with the first results. Whenever a paper is near acceptance, after much investment by the authors, the human tendency is to push the paper over the finish line, whatever contortions that takes. In any noisy statistical environment, that will often be possible merely by selective reporting or by benign looking options within researchers' personal "degrees of freedom".

It is hard to overstate this problem. It is a problem for good research as well as bad, because the external appearance of ill-motivated *p*-hacking is often just as imposing and convincing as a highly ethical "no games" empirical study producing results that are as genuinely untainted as they look. The absence of deductive theoretical relationships in accounting, the discretionary choices of proxies for variables that can't be observed and the noisy nature of underlying accounting, economic and social phenomena leaves wide latitude for opportunism in empirical accounting research, and, especially given the rewards at stake, simply invite malpractices, ranging from faking data and results (as in the publicized retracted *Accounting Review* papers) to practices that might equally occur when the researchers genuinely want more evidence.

The problem for genuine research is that the process of learning while experimenting, gathering more data, improving proxies and experimental or statistical controls, and even things like cleaning data and removing outliers, can be explained equally by either "good" or "bad" science. To the outsider, they often look for all intents and purposes the same, which is a problem that will not go away in any research environment where work is published in elite journals even when it will not or cannot be replicated with new and independent data. The mechanics of information asymmetry, adverse selection and the market for lemons suggest that genuine research effort will go relatively unrewarded.

A problem for evaluating accounting research and its reliability is that few results are tested by being applied. In medicine, a drug that is tested and found to work, or found to have no apparent side effects, is tested more and more extensively when it is prescribed to many and varied patients. In accounting, many results are tested once, published, and not subjected to any independent re-testing. The truth or falsity of those published results remains unknown and often very uncertain.

That distrust begs a more general question, one that will make sense to empiricists because it is an empirical question – to wit, what proportion of rejected hypotheses in accounting are in fact false or false for all practical purposes? Or, in other words, what proportion of "found" and published statistically significant effects are in fact substantially true?

This question is unanswerable, for several reasons. Even if we know the error probabilities of tests, conditional that is on our models being correct, we do not know the proportions of true effects and true models that are picked out for testing by researchers. Researchers will be biased in what they choose to test, selecting "interesting" or "hot" research questions perhaps rather than ones that involve any particular frequency of true or false hypotheses. And then, when testing, they might be biased towards finding a significant effect, we won't know from what is published:

> ... usually it is impossible to decipher how much data dredging by the reporting authors or other research teams has preceded a reported research finding. (Ioannidis, 2005b, p.700)

Still deeper empirical unknowns are the extent to which true "accounting effects" actually exist and the extent that they can actually be proven or demonstrated using extant or new accounting theories, models, data, statistical methods and research skills we have available. Put together these unknowns raise a fundamental empirical question – to wit, is it possible in accounting related subject matter to correctly find true "law like" replicable effects? One answer to this question is "yes, look at Ball and Brown". But what of all the vast numbers of empirical findings that are not so manifest in simple plots and not so widely and independently replicated? How true, lasting or reliable are they all – that is simply unknown and in most cases unknowable.

In a very widely cited paper in medicine, Ioannidis (2005b) gave a coherent explanation of the hypothesis that most empirical research papers are actually wrong. In that paper he considers the extreme case where a research field contains no true effects but some effects are found to be statistically significant. The proportion of tested "statistically significant" effects in this field is thus a measure not of scientific progress but of researcher bias:

> Let us suppose that in a research field there are no true findings at all to be discovered. History of science teaches us that scientific endeavor has often in the past wasted effort in fields with absolutely no yield of true scientific information, at least based on our current understanding. In such a "null field," one would ideally expect all observed effect sizes to vary by chance around the null in the absence of bias. The extent that observed findings deviate from what is expected by chance alone would be simply a pure measure of the prevailing bias. (Ioannidis, 2005b, p.700)

Bias that leads to many statistically significant effects need not be intentional. Wrong models will often produce apparent but false effects, and will of themselves often produce statistical significance without any need for the researcher to intentionally "culture" statistical significance in a petri dish.

It is a remarkable technical point concerning *p*-levels in common contexts that even in the case of a true null hypothesis and an unbiased experiment, the probability of obtaining any given (possibly small) level of significance by sampling until that occurs is equal to one (Kadane, Schervish, & Seidenfeld, 1996). This is known as "sampling to a foregone conclusion". With a biased or wrong model, the probability of sampling until significant is also one, and "statistical significance" will often occur without much sampling, especially with a "lucky" bias inherent in the model:

> We do not know what percent of the published statistically significant findings are lucky false positives explained *post-hoc* (Szucs & Ioannidis, 2017)

In the "market for statistical significance" (Johnstone, 1988; Quiggin, 2019), which is lit up brightly by its tiny *p*-levels and rows of stars, the sellers of significance appear to have frequently duped the buyers. John et al. (2012) found in a survey that about 65% of psychology researchers admitted *p*-hacking by dropping dependent variables from their models. In their recent JAR based survey of accounting researchers, Hail et al. (2020) give results that overall say that researchers do not have ready confidence in the reproducibility of others' published empirical results. The reasons offered by insiders for why published work will not stand up are summarized as follows:

> Few respondents thought that outright fraud ("fabricated or falsified results") was common but few also thought that bad luck (random chance) was a major cause. Instead, the leading explanations were based on researchers' incentives with selective reporting of results" and "pressure to publish for career advancement" thought to always or very often contribute to the irreproducibility of extant findings. (Hail et al., 2020, p.7)

These explanations of how results that are either wrong or misleading, at least relative to what another researcher with the same data or new data would find, constitute the beginnings of an "agency theory of accounting researchers". The recurrent themes, as in agency theory, are self-interest and monitoring, not that individual researchers can be blamed in the sense that what is required within accounting research schools for appointments, tenure and respect is now hard wired into the culture and not necessarily relished by researchers who embark with good intentions yet find themselves subject to cultish behaviours and rules.

An overall way to understand the replication crisis in terms of its social and economic causes is one of "market failure" in the market for statistical evidence:

> In relation to the replication crisis, the core of the problem, as it is generally perceived, is that the price of including a variable of interest as statistically or economically significant in the reported model is too low. On the one hand, the bias against publication of negative results means that the benefit to researchers of reporting models with no significant variables of interest is limited and, in many fields, close to zero. On the other hand, the availability of the set of techniques pejoratively referred to as "p-hacking" means that the test statistics reported in published studies

> are more likely to arise through chance than would be suggested by the classical hypothesis-testing framework. Taken together, these observations suggest that the price to researchers of including a variable of interest in a published model is lower than would be socially optimal. As a result, too many positive results are reported. (Quiggin, 2019, p.6)

Clearly aware of concerns around a literature built on little theory and lots of *p*-levels, the Hail et al. (2020) survey must be praised for its openness. It exhibits not only pessimism about what has been learned reliably in empirical accounting research, but also about the inbuilt motivations and incentives that drive "unscientific" behavior by the researchers, or by those whose work gets into print.

The Hail et al. (2020) survey was based on a matching survey conducted by journal *Nature*, with mostly the same questions and response options. One potential driver of poor research reproducibility that was not considered in the *Nature* survey was along the lines of "no one and nothing depends on us getting this right". In fields like biology and biomedicine, that may well be a reasonable statement for less important research, but in accounting it must be typically true. In engineering, empirical tests of an empirical issue like how different concrete mixtures affect strength will have potentially dire consequences if wrong results become established. That practical relevance and economic significance cannot be attached to most research in accounting. For one thing there is no profession of accountants looking to the "elite accounting research journals" to base their day to day work upon. The general lack of impact on society of empirical accounting research is an invitation to researchers to please themselves and their referees. There is no comeback, and as pointed out by Ioannidis (2005b) it is rarely clear whether an empirical research paper is a window-dressed and sanitized version of what really happened or an earnest depiction, nor is it clear without reproduction of results (i.e. with new data and better experimental control) that the results shown are true in the physical or positivist sense that they won't fail when acted on.

The fact that readers cannot discern the difference between published Study A which was rigorous and honest, and happened to produce interesting results, and Study B that was fiddled from the start so as to find interesting and significant results while still looking like it has the same rigor as Study A, is a nearly insurmountable problem, especially in the absence of a culture of retesting and replication, which itself is unlikely to ensue in circumstances where the results are of no economic value or even much interest to external users.

Preregistration is designed to preclude Study B. The problem is that tight pre-registration prevents learning and re-modelling as data arrives and the study deepens (Gelman and Loken, 2014). An interesting aspect of the philosophical debate that occurred in the 1930s and later between the Neyman-Pearson and Fisherian branches of frequentist statistics was that Fisher accused Neyman's hypothesis tests as being "non-scientific" because it locked experimenters into predesignated statistical protocols before experiments started and any data was seen (see Johnstone, 1987b for details). That same "it's not science" criticism applies to key elements of preregistration.

Perhaps a good way to say this is that if we truly wanted to investigate some empirical hypothesis, for our own ends, we would not lock ourselves to any one way of doing that before we started, because that would limit remodelling and resampling when evidence mounts and knowledge increases. The fact that preregistration has gained general support, against traditional ideas of learning and adjusting along the way, is driven by researchers' beliefs that other researchers have conflicts of interest and cannot be trusted to be guided by evidence rather than personal rewards. This leads back to agency theory and the role of contracts and monitoring.[18] It seems that empirical research fields need to incur all the costs and lost opportunities of hamstringing research so as to avoid the costs of an inherently biased research methodology and culture. Preregistration is akin to using contracts to reduce monitoring costs.

A problem in the market for statistical significance is that the intrinsic qualities of the thing being sold are not observable to the buyer, much as in the Akerlof (1970) "market for lemons". Many of the Akerlof corollaries apply. Information asymmetry between authors (sellers) and readers (buyers) will allow and reward opportunistic behaviours by authors and leave an adverse selection problem for journal editors in the case of papers rejected at other journals or sent to journals in more need of copy. Genuine researchers may be driven out of production, if they have no way to effectively "signal" the true quality of their work. In a limiting case, all papers published will be assumed to be "lemons", as could hold true if enough genuine researchers find more rewarding applications for their skills and honesty, and the market will collapse.

Another finance analogy paints statistical researchers as akin to noise traders, where statistical noise masquerading as reliable evidence or a meaningful pattern can tempt belief and investment in a false lead. Even the investigator does not know when false assumptions (e.g. a false model) or pet hypotheses have been confirmed merely by luck or noise.

It is almost a paradox that positivist empirical research, sold from the outset as being superior because it is "scientific", is less transparent in what it found, and how it came to that result, than most of the schools of accounting research that it relegated. Old fashioned "normative" accounting research, like the works of May, Baxter, Canning and Chambers, showed in full view, in writing, its full makeup. The reader could see for herself what was questionable, weak or missing in all

[18] An expensive form of monitoring is an audit of results by repeating the analysis with the same data. It is often noted that published empirical papers provide too little detail to allow others to simply repeat the analysis. This form of auditing should not to called "replication", which according to R.A. Fisher requires the same experiment under different conditions (e.g. with new data, not the same data). In Fisher's statistical model, replication would mean something like comparing the yields of two seed types by growing them in matched pairs across the same plot ("repetition" would mean a different farm and year altogether).

the authors' arguments. By stark comparison, readers cannot observe or sometimes even guess what is missing or omitted in an empirical research study.

Good normative research cannot be faked, no different in that regard to a good novel, play, cross-word, computer program or mathematical proof. The same applies to an accounting analytical model like Fischer and Verrecchia (2000) because virtually everything about that model's analysis and conclusions is evident and can be checked by the reader. Interestingly, the sociological/critical research published in journals like *CPA*, which is vastly different in its underlying epistemological precommitments to game-theoretic rational economics, has the same transparency, because the reader sees the whole derivation. Its entire argument is there, to be viewed and questioned from any chosen angle. Transparency is not one of the selling points of most empirical research. Results that don't fit the story are easily left out.[19]

Whether out of dissatisfaction or monotony, empirical researchers have begun to reveal their own distaste with their own research methods. Simmons, Nelson, and Simonsohn (2018b) explained that they wrote their influential paper "False-Positive Psychology" (2011) after realizing their own research sins:

> We knew many researchers – including ourselves – who readily admitted to dropping dependent variables, conditions, or participants so as to achieve significance. Everyone knew it was wrong, but they thought it was wrong the way it's wrong to jaywalk. We decided to write "False-Positive Psychology" when simulations revealed it was wrong the way it's wrong to rob a bank. (Simmons et al., 2018b, p.255)

Opinion is however deeply split. In 2012, the Levelt Committee, Noort Committee, and Drenth Committee (2012) enquiring into fraudulent research by highly reputed social psychologist Diederik Stapel found numerous researchers defending unscientific biased research methods such as the usual ways of "*p*-hacking" and excluding unwanted data or sampling until getting the "right" conclusion.

> Another clear sign is that when interviewed, several co-authors who did perform the analyses themselves, and were not all from Stapel's 'school', defended the serious and less serious violations of proper scientific method with the words: that is what I have learned in practice; everyone in my research environment does the same, and so does everyone we talk to at international conferences. (Levelt Committee et al., 2012, p.48)

5. Conclusion

It is a curiosity in accounting research that analytical research papers that rest explicitly on Bayesian reasoning, and empirical research papers that are totally reliant on frequentist non/anti-Bayesian methods (*p*-levels), sit equally side by side in the same journal issues, presumably satisfying the same refereeing and scientific standards, if not the same referees.

In some empirical papers, hypotheses are hatched on a Bayesian depiction of investor behavior but are tested using frequentist logic. Such apparent cognitive dissonance within the accounting research community has been buried by different sleights of hand. Typically, for example, empirical researchers misrepresent significance tests by twisting their meaning in ways that sound sensible and pseudo-Bayesian or (more commonly) by saying nothing explicit as if to let low *p*-values speak for themselves.

However expressed, the rejection of the null hypothesis in a 5% significance test is held out as "evidence against the null hypothesis", which is a de facto Bayesian statement, and gets at what the reader is actually interested to know. The problem is that a correct Bayesian interpretation shows that rejection at 5% can in fact imply strong evidence in favour of the null hypothesis being true, either literally or in the practical sense that the observed effect is only a miniscule different from zero. A genuine Bayesian interpretation would lead to pointed questions such as whether the evidence is really strong and what alternative hypotheses (small or large effect sizes) are now most probable given what has been observed. The most probable hypothesis might now be so close to the hypothesis "rejected" that it makes no difference.

After decades of Bayesian criticism of significance tests, including repeated explanations of how conventional tests are inherently open to abuse, there has lately been a tide change in mainstream statistics and the empirical social sciences. The conservative *American Statistical Association* has issued formal statements advising statisticians in applied fields against relying on significance tests and against the *p*-level orthodoxy that has held sway in psychology and other social sciences for 50 years. Many professional statisticians and even some accomplished empirical researchers are now saying in effect that "the game is up" for empirical research in the social sciences. Consumers of research papers are increasingly aware that the simplistic mechanical significance testing methods that have been rehearsed for generations in PhD programs have created problems for all the social sciences, to the point of being repudiated by eminent theorists in published petitions, disowned by the *American Statistical Association* and blamed widely for the lately growing "replication crisis".

The next phase for empirical research in the social sciences might be a shift towards using Bayesian methods explicitly or at least interpreting frequentist results in ways that are more consistent with Bayesian inference. That shift of statistical logic is desirable, but will not come easily in the face of an entrenched commitment and virtual love affair with *p*-levels. Their

[19] In "real" science, transparent lab practice involves keeping lab notes that record by time and steps the researchers' progress, all for later viewing and to assist attempts to replicate results when replication turns out warranted, as it will when the result claimed is important enough.

elegant appearance and convenience, their routine production in statistical software, and the history of journals and respected papers adorned by them, not forgetting the careers built on them, has *p*-levels deeply ensconced in the empirical psyche. To denounce them and much of the work based on them would be akin to saying that 50 years of empirical papers need to be reassessed and possibly largely discounted, and that we as a research community are not as clever or "scientific" as our students have been led to believe. The alternative is to go on with the current undesirable "equilibrium" (Ohlson's word) with relatively minor patch ups (e.g., setting the required significance hurdle at two or three stars instead on one).

The easiest modification, one that has been proposed repeatedly (e.g., Gardner & Altman, 1986) but not taken up by researchers, is to report confidence intervals rather than *p*-levels. The obvious disincentive for researchers is that *p*-levels make results appear far more impressive than confidence intervals. A completely immaterial effect, i.e., an effect that is barely even measurable, can have a very small *p*-level, and will do so virtually always when the sample is large. That is the Lindley paradox – a large sample "should" mean that the evidence of a material effect is stronger, but for a given *p*-level (say $p = 0.05$) the evidence is actually weaker (the confidence interval is centred more tightly closer to zero). The obvious question is "how can that be?" and the answer, according to Bayesians, is that it occurs because the common logic of significance tests is illogical.

Going beyond methodological issues of Bayesian versus frequentist methods, the bigger and far more daunting issue for empirical accounting research is whether empirical researchers can expect to ever make significant progress in such an inherently complicated, changing and noisy statistical environment, wherein robust statistical relationships, especially ones that are surprising and practically important, do not come easily. A plausible hypothesis is that it has taken 50 years or more for the empirical social sciences to discover generally, in multiple different social and behavioral fields of enquiry, that the empirical research task is intrinsically harder to progress than imagined, and may be so inherently difficult that replicable and practically meaningful empirical truths or laws, apart from obvious ones that require little or no testing, are naturally few or non-existent.

That conjecture, which is itself an empirical hypothesis, says simply that empirical methods that often do well in say medicine or agriculture (e.g., in testing whether a drug or fertilizer works) do not work as easily in the inherently more complicated, noisy and changeable social, psychological and political sciences. If that is so, as a crude rule, then it has been obscured by the millions of published *p*-levels that all insinuate reliable evidence of something important.

As a way forward, it seems time for a survey and audit of empirical accounting research, with the objective of identifying accounting relationships and effects that are well supported by evidence, in the sense that they can be said to exist and be material in magnitude by either practical or scientific criteria. Fifty years after Ball and Brown (1968), it would be of interest to gauge what empirical "law like" generalizations have been established. For example, Chua (2019) critique of the positivist research culture in accounting suggests that the proposition that "accounting information has value relevance to stock markets" might be one such product. Note however that this proposition is weak until the practical size, and not mere existence, of that accounting effect is evident empirically, and is proved not to be a mere chance correlation driven by underlying confounding forces such as other overlapping information sources. Science is ultimately about the practical value or size of marginal effects (Cohen, 1988, 1994) if only because small effects are usually no help and economically unexploitable.

An audit and taxonomy of accounting empirical research findings would be of interest to all researchers, empiricists or not, and would have the implicit benefit of bringing into light the big unspoken statistical/methodological issues of what constitutes reliable statistical modelling and evidence and what makes successful and worthwhile empirical research. These questions and mindset occur naturally in Bayesian inference. The Bayesian approach allows for probability distributions on models as well as hypotheses, which is the holistic way to think about using data to make inferences about effects and their sizes. There is no Bayesian panacea, however, for the inherent difficulty of forming confident beliefs about inherently noisy and poorly understood economic and social phenomena. Gelman (2007) himself has commented that in his empiricist career he has inevitably been modelling only known unknowns or "white swans".[20]

The significance test crisis, or replication crisis, arose with venom in psychology, where some previously important and ostensibly evidence-based results were found not to be replicable, and where research results appeared that were ludicrous yet satisfied common *p*-level criteria (Carey, 2011; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). It is a mark of the practical importance of the results in question, and of the potentially self-correcting attribute of a genuine science, that such results were re-tested. The question in accounting, essentially the same question that Chua (2019) raised, is whether there are empirical results in "accounting science" that have this established acceptance and practical importance, and can they be replicated with new data?

The replication crisis in psychology is of immediate relevance to accounting. If parts of the accounting research literature open up to the replicability and related methodological critique, the chances are that the academy will fragment further. The difference however is that there is likely to be division within the empiricists themselves, and almost assuredly a backlash by researchers who resent their work, methods and ideology coming into question, especially when the positivist-empiricist creed has always been held out in accounting as the antidote to accounting's historically "normative", "unscientific" and "pre-science" roots:

[20] https://statmodeling.stat.columbia.edu/2007/04/09/nassim_talebs_t/

> This week, a global network of nearly 100 psychologists unveiled the results of their attempt to repeat 27 well-known psychology studies. In more than half of the cases, the replication was a partial or complete failure. Some are heralding the replications as a renaissance for the social sciences. But some on the receiving end of the replication are calling it an inquisition. (Bohannon, 2014, p.788)
>
> . . .those authors whose research is the subject of a replication effort may interpret the very act of replication as a challenge to their professional competence and integrity. . . (Dewald, Thursby, & Anderson, 1986, p.601)

Ultimately the test of a science is its reaction to criticism and openness to being questioned (Williams & Rogers, 1995, p.264). The signs at least from JAR editors (Hail et al., 2020; Leuz, 2018) are that in accounting the "rules of the game" are changing and that new research protocols emphasizing both better methodology and improved ethics, all designed to motivate better research habits and more reliable results, are being developed. The same optimistic sentiments are coming from the editors of *Science* (Berg, 2018, 2019). The claimed benefits for all statistical sciences are first a more common and intuitive understanding of the limitations of statistical techniques, and then a newfound interest in obtaining results that hold and can be replicated, rather than results designed foremost to gain the acceptance of referees and journal editors.

It is hard, however, to foresee how statistical-methodological reforms and in-house rule changes can bring such fundamental ideological and sociological change in empirical research communities. Romero (2019) and Nosek et al. (2012) argue that initiatives designed to improve researchers' technical understanding of statistical methods treat the symptoms rather than the disease. In their view the underlying problem is behavioral and social,[21] and not merely methodological:

> For the social reformist, it is too optimistic to expect scientists to follow good practices (in particular, to do replication work) if the right incentives are not in place. This is because science today is a professionalized activity. As such, scientists are constrained not only by the ethos of science but also by more mundane and arguably more forceful pressures, such as the requirement to produce many novel findings to have a career and continue playing the game. (Romero, 2019, p.9)
>
> Because we have directional goals for success, we are likely to bring to bear motivated reasoning to justify research decisions in the name of accuracy, when they are actually in service of career advancement. Motivated reasoning is particularly influential when the situation is complex, the available information is ambiguous, and legitimate reasons can be generated for multiple courses of action. . . . with flexible analysis options, we are more likely to find the one that produces a more publishable pattern of results to be more reasonable and defensible to others. (Nosek et al., 2013, p.617)

One main impediment to reform is that the significance testing pro forma has served a social purpose for many interested parties. Part of why it survives is that it covers for the lack of scientific progress in many fields by letting fields at least look and feel like science. Remarkably this political interpretation, which will clearly sit easily with many critical accounting theorists, has been directed at modern economic theory and empiricism by editors in a journal which is decidedly not on the usual reading list of critical theorists:

> The role of economic theory is not to explain the world; its role is to legitimize a political subject under the false cover of scientific knowledge. (López de Prado & Fabozzi, 2017, p.4)

Social reforms in research communities that de-emphasize researchers' need for publications, personal recognition, research grants, and individual financial rewards are in many ways counter to the law of the jungle. If we consider, for example, ways of encouraging greater funding and publication of replication studies, so as to ideally accelerate the "self-correction" mechanism in science, those reforms would run up against the deeply rooted natural respect that researchers have for only new and interesting discoveries:

> The primary objective of science as a discipline is to accumulate knowledge about nature. Learning something new advances that goal; reaffirming something known does not. (Nosek et al., 2013, p.617)
>
> Within the social sciences, only the discovery of a new fact is credited. (Schmidt, 2009 p.95)

In what is now a vast literature on what's wrong with significance tests, and how so many research programs have arrived at a replication crisis, most discussion on how results can be made more reliable is aimed at (i) the families of statistical methods used, and (ii) better protocols for designing and running statistical experiments. These suggestions are for the most part decades old and never previously taken up systematically. It may seem that researchers' resistance to better statistics, beginning with the Bayes-versus-frequentist debate, is mainly based on merely arguable "technical" matters. Indeed, many users of frequentist tests and *p*-levels have put counter-arguments for why these are not a problem, or are not their problem. But the deeper and more latent resistance, explaining why reforms of statistical methods and their uses and use have not been acted upon, is not that bothersome statistical retooling is necessary, but that any true reforms would rearrange the underlying social order in research fields and would potentially limit the rewards and social advancement available to the brand of researchers who live and work well under the existing culture of *p*-levels.

[21] See Williams and Rogers (1995) on the social production of accounting knowledge.

It seems axiomatic that if statistical methods and practices were more demanding and the evidential hurdle raised, entire fields would come into question, not merely in what has been achieved to date, but in ways that make it harder for researchers to justify new conclusions and continued funding and rewards.

Current research protocols have allowed research fields like accounting, which at its worst can be considered as a Feynman "cargo cult" science (Basu, 2012) or a "folk science" (Williams & Rogers, 1995), to build the imagery of a hard science but without any of the external testing that traditional science faces, both in other scientists' laboratories or in the more telling laboratory called the "real world", where for example a new drug or vaccine is tested in a massive "experiment" by widespread usage.

It is an enviable aspect of the life of an "accountics"[22] scientist that inferences like "better accounting reduces the firm's cost of capital" are never clearly tested or potentially refuted in a real world application, unlike for example when a drug company releases a new drug that medical scientists held to be effective and also to have no bad side effects. That threat of imminent or eventual real-world testing does not apply to empirical accounting research, which is part of why researchers are easily able to select a model or statistical-method option (within their wide "degrees of freedom") or compose a theoretical narrative, possibly ad hoc after results are known, which they foresee as acceptable to referees. Without any general opportunity for demonstrable real-world testing or benefit, the remaining claimable proof of the quality of the work is its having been published in a "top journal", which comes down to its political acceptance inside a community that explains itself politically to the outside world by its appearance and claims as a science, but operates internally like a "church" with ordained elders and high priests.

My paper will be discounted by many empiricists in accounting as "polemical" and unmannerly, which is itself a way to imply a lack of objectivity, scientific detachment and reason. If so, it should be said that as a polemic it does not rate alongside the critique of the significance testing culture that is now openly aired and published in medical sciences and psychology.

Similar polemics have in the past been easily dismissed or ignored. Like in other disruptions of long-standing industries and vested interests, progress against an established culture appears to require a "Me Too" style social movement that brings all of the issues into public view. The next steps in what is brewing as disruption of accepted research methods and academic power bases are evident in modern day eminent statistician-activist Gelman's blog. In one of many blog discussions under the heading "Retire statistical significance", Gelman (2019) notes that John Ioannidis, whose (2005b) paper has been cited 10,000 times, was uneasy about the group-signed petition in Greenland et al. (2016), and wrote of it as follows:

> I am afraid that what you are doing at this point is not science, but campaigning. Leaving the scientific merits and drawbacks of your Comment aside, I am afraid that a campaign to collect signatures for what is a scientific method and statistical inference question sets a bad precedent. It is one thing to ask for people to work on co-drafting a scientific article or comment. This takes effort, real debate, multiple painful iterations among co-authors, responsibility, undiluted attention to detailed arguments, and full commitment. Lists of signatories have a very different role. They do make sense for issues of politics, ethics, and injustice. However, I think that they have no place on choosing and endorsing scientific methods. Otherwise scientific methodology would be validated, endorsed and prioritized based on who has the most popular Tweeter, Facebook or Instagram account. I dread to imagine who will prevail.

Gelman (2019) noted that Sander Greenland replied:

> YES we are campaigning and it's long overdue ... because YES this is an issue of politics, ethics, and injustice!

Gelman (2019), who was one of the signatories to the petition, summed up as follows:

> My own view is that this significance issue has been a massive problem in the sociology of science, hidden and often hijacked by those pundits under the guise of methodology or "statistical science" (a nearly oxymoronic term). Our commentary is an early step toward revealing that sad reality. Not one point in our commentary is new, and our central complaints (like ending the nonsense we document) have been in the literature for generations, to little or no avail... Single commentaries even with 80 authors have had zero impact on curbing such harmful and destructive nonsense. This is why we have felt compelled to turn to a social movement: Soft-peddled academic debate has simply not worked. If we fail, we will have done no worse than our predecessors in cutting off the harmful practices that plague about half of scientific publications, and affect the health and safety of entire populations.

References

Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., et al (2015). Estimating the reproducibility of psychological science. *Science, 349*, 943.
Akerlof, G. A. (1970). The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics, 84*, 488–500.
Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, 305–307.
Anvari, F., & Lakens, D. (2018). The replicability crisis and public trust in psychological science. *Comprehensive Results in Social Psychology, 3*, 266–286.

[22] This term was coined by retired accounting professor Bob Jensen.

Bailey, C. N., Hasselback, J. R., & Karcher, J. N. (2001). Research misconduct in accounting literature: A survey of the most prolific researchers' actions and beliefs. *Abacus, 37*, 26–54.
Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*, 423–437.
Ball, R., & Brown, P. R. (1968). An empirical evaluation of accounting income numbers. *Journal of Accounting Research, 6*, 159–178.
Basu, S, & Markov, S (2004). Loss function assumptions in rational expectations tests on financial analysts' earnings forecasts. *Journal of Accounting and Economics, 38*, 171–203.
Basu, S (2012). How can accounting researchers become more innovative?. *Accounting Horizons, 4*, 851–870. https://doi.org/10.2308/acch-10311.
Basu, S., & Park, H. -U. (2014). Publication bias in recent empirical accounting research. 2014 Canadian Academic Accounting Association (CAAA) Annual Conference; Fox School of Business Research Paper No. 14-027. Available at SSRN: https://ssrn.com/abstract=2379889 or http://dx.doi.org/10.2139/ssrn.2379889
Basu, S. (2015). Is there a scientific basis for accounting? Implications for practice, research, and education. *Journal of International Accounting Research, 14*, 235–265.
Basu, S., Vitanza, J., & Wang, W. (2020) Asymmetric loan loss provisioning. Journal of Accounting and Economics. In press.
Begley, C. (2013). Reproducibility: Six red flags for suspect work. *Nature, 497*, 433–434.
Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature, 483*, 531–533. https://doi.org/10.1038/483531a.
Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods, 2*, 131–160.
Benjamin, D. J., Berger, J. O., [. . .], Johnson, & V.E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*, 6–10.
Berg, J. (2018). Progress on reproducibility. *Science, 359*, 6371–6379.
Berg, J. (2019). Replication challenges. *Science, 365*. 6457: 957.
Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York: Springer-Verlag.
Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the Chi-Square Test. *Journal of the American Statistical Association, 33*, 526–542.
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.
Bohannon, J. (2014). Replication effort provokes praise - and 'bullying' charges. *Science, 23*(344), 788–789.
Boos, D. D., & Stefanski, L. A. (2011). P-Value precision and reproducibility. *The American Statistician, 65*, 213–221.
Breuer, M., & Schütt, H. H. (2019) Accounting for uncertainty: An application of Bayesian methods to accruals models. https://ssrn.com/abstract=3417406 or http://dx.doi.org/10.2139/ssrn.3417406
Brown, P. (2013). How can we do better? *Accounting Horizons, 27*, 855–859.
Bruns, S. B., & Ioannidis, J. P. (2016). p-Curve and p-Hacking in observational research. *PloS One, 11*(2) e0149144.
Burgstahler, D. (1987). Inference from empirical research. *The Accounting Review, 62*, 203–214.
Camerer, C. F., Dreber, A., et al (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour, 2*, 637–644.
Carey, B. (2011). Journal's paper on ESP expected to prompt outrage. *The New York Times*. Jan. 5.
Chen, Q., & Schipper, K. (2016). Comments and observations regarding the relation be-tween theory and empirical research in contemporary accounting research. *Foundations and Trends in Accounting, 10*, 314–360.
Chordia, T., Goyal, A., & Saretto, A. (2017). *P-hacking: Evidence from two million trading strategies* Working Paper. Emory University, University of Lausanne, and University of Texas at Dallas.
Chua, W. F. (2019). Radical developments in accounting thought? Reflections on positivism, the impact of rankings and research diversity. *Behavioral Research in Accounting, 31*, 3–20.
Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Academic Press.
Cohen, J. (1994). The earth is round p < 0.05. *American Psychologist, 49*, 997–1003.
Cox, D. (1982). Statistical significance tests. *British Journal of Clinical Pharmacology, 14*, 325–331.
Dewald, W. G., Thursby, J. G., & Anderson, R. G. (1986). Replication in empirical economics: The journal of money, credit and banking project. *The American Economic Review, 76*, 587–603.
Dumay, J., & de Villiers, C. (2019). Qualitative accounting research: Special issue introduction. *Accounting and Finance, 59*, 1449–1458.
Dyckman, T. R. (2016). Significance testing: We can do better. *Abacus, 52*, 319–342.
Dyckman, T. R., & Zeff, S. A. (2014). Some methodological deficiencies in empirical research articles in accounting. *Accounting Horizons, 28*(3), 695–712.
Dyckman, T. R., & Zeff, S. A. (2015). Accounting research: Past, present, and future. *Abacus, 51*(4), 511–524.
Dyckman, T. R., & Zeff, S. A. (2019). Important issues in statistical testing and recommended improvements in accounting research. *Econometrics*. https://doi.org/10.3390/econometrics7020018.
Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*, 193–242.
Feltham, G. A., & Demski, J. S. (1970). The use of models in information evaluation. *The Accounting Review, 45*, 623–640.
Fischer, P. E., & Verrecchia, R. E. (2000). Reporting bias. *The Accounting Review, 75*, 229–245.
Francis, J. R., LaFond, R., Olsson, P., & Schipper, K. (2007). Information uncertainty and post-earnings-announcement-drift. *Journal of Business Finance and Accounting, 34*, 403–433.
Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than P values: Estimation rather than hypothesis testing. *British Medical Journal, 292*, 746–750.
Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian. Analysis, 3*, 445–450.
Gelman. (2016). https://statmodeling.stat.columbia.edu/2016/09/21/what-has-happened-down-here-is-the-winds-have-changed/
Gelman. (2018). https://statmodeling.stat.columbia.edu/2018/02/28/fear-many-people-drawing-wrong-lessons-wansink-saga-focusing-procedural-issues-p-hacking-rather-scientifically-important-concerns/#respond
Gelman. (2019). https://statmodeling.stat.columbia.edu/2019/03/20/retire-statistical-significance-the-discussion/
Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. https://www.semanticscholar.org/paper/The-garden-of-forking-paths-%3A-Why-multiple-can-be-a-Gelman-Loken/b63e25900013605c16f4ad74c636cfbd8e9a3e8eff.
Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*, 460–465.
Gordon, I. M., & Boland, L. A. (2015). Anatomy of a journal: A reflection on the evolution of *Contemporary Accounting Research*, 1984–2010. *Accounting History, 20*, 464–489.
Greenland, S. Senn, S. J., Rothman, K. J., Carlin, J. B. Poole, C., Goodman, S. N., & Altman, D. G. (2016) Statistical tests, P-values, confidence intervals, and power: A guide to misinterpretations. Online Supplement to the ASA Statement on Statistical Significance and P-values. Also published in European Journal of Epidemiology. 31: 337-350.
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. *Journal of Finance, 72*, 1399–1440.
Harvey. (2019). Replication in financial economics. https://ssrn.com/abstract=3409466 or http://dx.doi.org/10.2139/ssrn.3409466.
Harvey, C. R., Liu, Y., & Zhu, H. (2016). and the cross-section of expected returns. *The Review of Financial Studies, 29*, 5–68.
Harvey, C., & Zhou, G. (1990). Bayesian inference in asset pricing tests. *Journal of Financial Economics, 26*, 221–254.
Hoepner, A. G. F., McMillan, D., Vivian, A., & Simen, C. W. (2021). Significance, relevance and explainability in the machine learning age: An econometrics and financial data science perspective. *The European Journal of Finance, 27*, 1–7.

Hopwood, A. (2008). Changing pressures on the research process: On trying to research in an age when curiosity is not enough. *European Accounting Review, 17*, 87–96.
Hunter, J. E. (2001). The desperate need for replications. *Journal of Consumer Research, 28*, 149–158.
Ioannidis, J. P. A. (2005a). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association, 294*, 218–228.
Ioannidis, J. P. A. (2005b). Why most published research findings are false. *PLoS Medicine, 2*, 696–701.
Ioannidis, J. P. A. (2019). What have we (not) learnt from millions of scientific papers with *P* values? *The American Statistician, 73*(Suppl. 1), 20–25.
Jeffreys, H. (1939). *Theory of probability*. Oxford: The Clarendon Press.
John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524–532.
Johnstone, D. J. (1986). Tests of significance in theory and practice (with discussion by D.V Lindley and G. Barnard). *Journal of the Royal Statistical Society. Series D (The Statistician), 35*, 491–504.
Johnstone, D. J. (1987a). Tests of significance following R. A. Fisher. *The British Journal for the Philosophy of Science, 38*, 481–499.
Johnstone. (1987b) On the interpretation of hypothesis tests following Neyman and Pearson. In: Viertl, R. (Ed.) Probability and Bayesian Statistics. Boston, MA.: Springer. Pp. 267–277.
Johnstone. (1988). Comments on Oakes on the foundation of statistical inference in the social and behavioral sciences: the market for statistical significance. Psychological Reports. 63: 319-331.
Johnstone. (1995) Statistically incoherent hypothesis test in auditing. Auditing: A Journal of Theory and Practice. 14: 156–176.
Johnstone (1997) Comparative classical and Bayesian interpretations of statistical compliance tests in auditing. Accounting and Business Research. 28: 53-82.
Johnstone (2016) The effect of information on uncertainty and the cost of capital. Contemporary Accounting Research. 33: 752-774.
Johnstone (2018) Accounting theory as a Bayesian discipline. Foundations and Trends in Accounting. 13: 1-266.
Johnstone, D. J., & Lindley, D. V. (1995). Bayesian inference given data "significant at α": Tests of point hypotheses. *Theory and Decision, 38*, 51–60.
Kadane, J. B., Schervish, M. J., & Seidenfeld, T. (1996). Reasoning to a foregone conclusion. *Journal of the American Statistical Association, 91*, 1228–1235.
Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLoS ONE, 10*. https://doi.org/10.1371/journal.pone.0132382 e0132382.
Khan, M. J., & Trønnes, P. C. (2019). *p*-Hacking in experimental audit research. *Behavioral Research in Accounting, 31*, 119–131.
Kim, J. H., Ahmed, K., & Ji, P. I. (2018). Significance testing in accounting research: A critical evaluation based on evidence. *Abacus, 54*, 524–546.
Kim, J. H., & Ji, P. I. (2015). Significance testing in empirical finance: A critical review and assessment. *Journal of Empirical Finance, 34*, 1–14.
Kinney, W. R. (1986). Empirical accounting research design for Ph.D. students. *The Accounting Review*, 338–350.
Kupferschmidt, K. (2018). More and more scientists are preregistering their studies. Should you? https://www.sciencemag.org/news/2018/09/more-and-more-scientists-are-preregistering-their-studies-should-you
Lambdin, C. (2012). Significance tests as sorcery: Science is empirical - significance tests are not. *Theory and Psychology, 22*, 67–90.
Leamer, E. E. (1983). Let's take the con out of econometrics. *American Economic Review, 73*, 31–43.
Leuz, C. (2018). Evidence-based policymaking: Promise, challenges and opportunities for accounting and financial markets research. *Accounting and Business Research, 48*, 582–608.
Levelt Committee Noort Committee & Drenth Committee (2012). *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel* Stapel Investigation: Joint Tilburg/Groningen/Amsterdam investigation of the publications by Mr. Stapel. Tilburg University.
Lindley, D. V. (1957). A statistical paradox. *Biometrika, 44*, 187–192.
Lindley, D. V. (1987). *Bayesian Statistics, a Review. (CBMS-NSF Regional Conference Series in Applied Mathematics)*. Society for Industrial Mathematics.
Lindsay, R. M. (1994). Publication system biases associated with the statistical testing paradigm. *Contemporary Accounting Research, 11*, 33–57.
Lindsay, R. M. (1995). Reconsidering the status of tests of significance: An alternate criterion of adequacy. *Accounting, Organizations and Society, 20*, 35–53.
López de Prado, M. (2017). Invited editorial comment: Finance as an industrial science. *Journal of Portfolio Management, 43*(4), 5–9.
López de Prado, M., & Fabozzi, F. (2017). Who needs a Newtonian finance? *Journal of Portfolio Management, 44*(1), 1–4.
Lovell, M. (1983). Data mining. *Review of Economics and Statistics, 45*, 1–12.
McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician, 73*, 235–245.
Meehl, P. E. (1967). Theory testing in psychology and physics: a methodological paradox.
Moizer, P. (2009). Publishing in accounting journals: A fair game?. *Accounting, Organizations and Society, 34*, 285–304.
Morrison, D. E., & Henkel, R. E. (Eds.). (1970). *The significance test controversy - A reader*. London: Butterworths.
Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour, 3*, 221–229.
Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia II: Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*, 615–631.
Nunnally, J. (1960). The place of statistics in psychology. *Education and Psychological Measurement, 20*, 641–650.
Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature, 506*, 150–152.
Ohlson, J. A. (2015). Accounting research and common sense. *Abacus, 51*, 525–535.
Ohlson. (2019). Researchers' data analysis choices: An excess of false positives? (June 2, 2019). Available at SSRN: https://ssrn.com/abstract=3221894 or https://doi.org/10.2139/ssrn.3221894
Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science. 349: aac4716.
Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence?. *Perspectives on Psychological Science, 7*, 528–530.
Quiggin, J. (2019). The replication crisis as market failure. *Econometrics, 7*, 44. 10.3390.
Ravenscroft, S., & Williams, P. (2009). Making imaginary worlds real: The case of expensing employee stock options. *Accounting, Organizations and Society, 34*, 770–786.
Romero, F. (2019). Philosophy of science and the replicability crisis. *Philosophy Compass, 14*. https://doi.org/10.1111/phc3.12633 e12633.
Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57*, 416–428.
Salterio, S. E. (2014). We don't replicate accounting research – or do we?. *Contemporary Accounting Research, 31*, 1134–1142.
Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.
Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*, 90–100.
Schrand, C. M. (2019). Impediments to relevant research. *Accounting Horizons, 33*, 11–16.
Seidenfeld, T. (1979). *Philosophical problems of statistical inference: Learning from R.A. Fisher*. Dordrecht: Reidel.
Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significance. *Psychological Science, 22*, 1359–1366.
Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2018a). Psychology's renaissance. *Annual Review of Psychology, 69*, 511–534.
Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2018b). False-positive citations. *Perspectives on Psychological Science, 13*, 255–259.
Szucs, D., & Ioannidis, J. P. A. (2017). When null hypothesis significance testing is unsuitable for research: A reassessment. *Frontiers Human in Neuroscience*. https://doi.org/10.3389/fnhum.2017.00390.
Staddon, J. (2017). *Scientific method: How science works, fails to work or pretends to work*. London: Taylor and Francis.

Stone, D. N. (2018). The "new statistics" and nullifying the null: Twelve actions for improving quantitative accounting research quality and integrity. *Accounting Horizons, 32*, 105–120.
Tukey, J. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
Verrecchia, R. E. (2001). Essays on disclosure. *Journal of Accounting and Economics, 32*, 97–180.
Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100*, 426–432.
Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-Values: Context, process, and purpose Available from. *The American Statistician, 70*, 129–133 http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XlOaE2MN.
Welker, M. (2016). Commentary on: A re-examination of the cost of capital benefits from higher-quality disclosures. *Journal of Financial Reporting, 1*, 97–99.
Williams (1989). The logic of positive accounting research. *Accounting, Organizations and Society, 14*, 455–468.
Williams, P. F., & Ravenscroft, S. P. (2015). Rethinking decision usefulness. *Contemporary Accounting Research, 32*, 763–788.
Williams, P. F., & Rogers, J. L. (1995). The accounting review and the production of accounting knowledge. *Critical Perspectives on Accounting, 6*, 263–287.
Zeff, S. A. (2019). A personal view of the evolution of the accounting professoriate. *Accounting Perspectives, 18*, 159–185.