JOURNAL OF PRAGMATIC CONSTRUCTIVISM

Is the validity of positivistic management accounting research exposed to questionable research practices?

Kristian Mohr Røge
Middelfart Kommune; Kristian.MohrRoge@middelfart.dk

Abstract

A recent paper in Management Accounting Research (MAR) claimed that the validity of positivistic management accounting research (PMAR) has increased significantly during the last four decades.
We argue that this is a misrepresentation of reality as the current crisis of irreproducible statistical findings is not addressed. The reliability and validity of statistical findings are under an increasing pressure due to the phenomenon of Questionable Research Practices (QRPs). It is a phenomenon argued to increase the ratio of false-positives through a distortion of the hypothetico-deductive method in favour of a researcher's own hypothesis. This phenomenon is known to be widespread in the social sciences. We therefore conduct a meta-analysis on susceptibility of QRPs on the publication practices of PMAR, and our findings give rise to reasons for concern as there are indications of a publication practice that (unintentionally) incentivises the use of QRPs. It is therefore rational to assume that the ratio of false-positives is well-above the conventional five-per cent ratio. To break the bad equilibrium of QRPs, we suggest three different solutions and discuss their practical viability.

Keywords: Philosophy of science; Questionable Research Practices (QRPs); hypothetico-deductive method; Publication practices

1 Introduction

Statistical inferences independent of scientific field are proving to be fragile, unreliable, lacking replicability and facing emerging problems with generalising 'lab' findings to a real-world setting[1] (Gelman & Loken, 2014b; Ioannidis, 2005; Simmons, Nelson, & Simonsohn, 2011). The irreproducibility of statistical findings has been found to be widespread (e.g., strategic management (Bergh, Sharp, Aguinis, & Li, 2017), economics (Camerer et al., 2016), general management (Banks, O'Boyle, et al., 2016; Banks, Rogelberg, Woznyj, Landis, & Rupp, 2016), medicine (Begley & Ellis, 2012; Ioannidis, 2005; Prinz, Schlange, & Asadullah, 2011), neuroscience (Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014)). The irreproducibility of statistical research is argued to be caused by Questionable Research Practices (QRPs), which is a phenomenon claimed to distort the hypothetico-deductive method in favour of a researcher's own hypothesis with the side effect of increasing the probability of a false-positive (Chambers et al., 2014; Ioannidis, 2005; Simmons et al., 2011).
Studies have confirmed the QRPs activities among business school researchers (Butler, Delaney, & Spoelstra, 2017; O'Boyle, Banks, & Gonzalez-Mulé, 2017) arguing that playing with numbers, playing with models and playing with hypotheses are not uncommon. Scholars have explained the existence of QRPs in three ways: the inadequate training of

[1] The October 19, 2013 issue of The Economist 'Unreliable research: Trouble at the lab' provides a lengthy critique of the inability to replicate scientific research. While ScienceNews wrote: It's Science's dirtiest secret: the "scientific method" of testing hypotheses by statistical analysis stands on a flimsy foundation" (Siegfried, 2010). Furthermore, a study by Camerer et al. (2016) published in Science tried to replicate 18 studies that were published in the American Economic Review and the Quarterly Journal of Economics. They "found a significant effect in the same direction as in the original study for 11 replications (61%); on average, the replicated effect size is 66% of the original…" (Camerer et al., 2016, p. 1433).

researchers, the pressure and incentives to publish in certain outlets, and the demand and expectations of journal editors and reviewers.

As the sine qua non method of positivistic management accounting research (PMAR) is null-hypothesis testing (NHST) (Chua, 1986; Lachmann, Trapp, & Trapp, 2017; Lindsay, 1994; Merchant, 2010; Shields, 1997; Van der Stede, Young, & Chen, 2005), it implies that QRPs is a potential threat to the validity of causal claims that PMAR intends to make (Ittner, 2014; Lachmann et al., 2017; Luft & Shields, 2014). It would be naïve to assume that researchers within PMAR are somehow resilient to QRPs if institutional actors, such as academic journals and publication counting deans, (unintentionally) incentivise QRP activities.

Unfortunately, the very attempt to uncover a 'true' rate of QRPs in published research is obscured by the very practices that make them questionable, as they tend not to be conducted transparently due to either misreporting or lack of reporting (Banks, O'Boyle, et al., 2016). But, if the publication practice of academic journals allows QRPs, then we must assume that QRPs have already invaded that particular scientific field.

In this light, we find it reasonable and necessary to investigate if the publication practices of high-ranking journals in the field of management accounting allow QRPs or whether they somehow have managed to be resilient to QRPs. As these high-ranking journals represent the ideal of scientific integrity and as the editors and reviewers are gatekeepers to scientific integrity, it implies that the solution to QRPs also lies here. QRPs are a paradox at work, because to live up to the positivistic image of 'pure science', academic journals and researchers may find themselves transgressing this very ideal (Butler et al., 2017).

The purpose of this paper is therefore to shed light on the likelihood of QRPs within PMAR and, in doing so, we try to answer the following research question: How susceptible are the publication practices of PMAR to the phenomenon of QRPs? To develop an analytical framework that can provide an answer to this question, we draw on the accumulated knowledge on QRPs from a wide range of scientific fields (e.g. Banks, O'Boyle, et al., 2016; Garud, 2015; Gelman & Loken, 2014a; Gigerenzer & Marewski, 2015; Ioannidis, 2005; Kerr, 1998; Nuzzo, 2014; Simmons et al., 2011).

The next section provides an extensive clarification of the phenomenon of QRPs and their consequences for the hypothetico-deductive method. The third section outlines the analytical framework developed oi section two and clarifies the data selection. Section four presents and discusses the findings, while section five concludes on the analysis and suggests how to mitigate a further development of QRPs in PMAR.

2 Questionable Research Practices – what is that?

The scientific system, or the 'publish or perish' culture, is claimed to have fostered a range of QRPs within hypothetico-deductive method (Chambers et al., 2014; Leung, 2011; Lindsay, 1994) and, as a result, the research community has expressed two main concerns. The first concern is a research bias: "*the combination of various design, data, analysis, and presentation factors that tend to produce research findings when they should not be produced*" (Ioannidis, 2005, p. 41). The second concern addresses the 'publish or perish' thought in research, where "*QRPs are the steroids of scientific competition, artificially enhancing performance and producing a kind of arms race in which researchers who strictly play by the rules are at a competitive disadvantage*" (John, Loewenstein, & Prelec, 2012, p. 524). Unfortunately, it appears that QRPs are not limited to a small subsection of the scientific community but are, in fact, widespread and by some considered 'defensible' (Butler et al., 2017; John et al., 2012; Martinson, Anderson, & de Vries, 2005; Starbuck, 2016).

By convention, researchers are ultimately perceived to be 'pure', that is, individuals motivated solely by the acquisition of knowledge. However, this utopian idea of a researcher is in reality flawed as researchers have human needs, desires and motives, just like non-researchers (Mahoney, 1977). Furthermore, the scientific ecosystem, in which researchers work, is built and maintained by a number of stakeholders (i.e. universities, funding bodies, industry stakeholders and publishers) whose interest may diverge from pure knowledge accumulation (Hardwicke, Jameel, Jones, Walczak, & Weinberg, 2014). Unfortunately, it appears that the incentives embedded in the scientific system do not adequately account for these human factors and, thus, reward individuals that are lucky or willing to 'play the game' of QRPs (Chambers et al., 2014; Hardwicke et al., 2014; Nosek, Spies, & Motyl, 2012; Starbuck, 2016).

QRPs are defined as a range of activities that, intentionally or unintentionally, distorts the hypothetico-deductive method in favour of a researcher's own hypothesis (Chambers et al., 2014; Hardwicke et al., 2014; John et al., 2012). These practices are typically *P-hacking, low statistical power, Hypothesising After the Results are Known (HARKing), publication bias, a lack of data sharing and lack of replications*, and they are claimed to increase the probability of making a false-positive finding (Ioannidis, 2005; Maniadis, Tufano, & List, 2014; Simmons et al., 2011). If QRPs are employed on a large scale, it would have a devastating impact on the validity of an entire field of scientific inquiry.

In the next section, we explore each QRP in a greater detail.

2.1 P-hacking and low statistical power: A search for significance

> "*If you torture the data long enough, it will confess."*
> Ronald Coase (Tullock, 2001, p. 205)

In research, NHST has become equated with scientific rigour and perceived as the touchstone for establishing knowledge. It is considered the *sine qua non* of the 'scientific method' for making scientific inferences (Gigerenzer & Marewski, 2015; Hubbard & Lindsay, 2013; Lindsay, 1994).

To clarify, NHST is a method where a researcher seeks to reject a straw-man null hypothesis as evidence in favour of some favoured alternative hypothesis based on the obtained P value. The *American Statistical Association* informally describes P value as the probability of a statistical summary of the data being equal to or more extreme than its observed value under a specified statistical model (Wasserstein & Lazar, 2016). A P value therefore measures the incompatibility between a set of data and a proposed model for the data. That is, the smaller the P value, the greater the statistical incompatibility of the data with the null hypothesis – if the underlying assumptions used to calculate the P value hold. However, P values do not measure the probability for the studied hypotheses to be true, nor the probability for the data to be produced by random chance alone (Wasserstein & Lazar, 2016). Instead this depends on a range of parameters, for instance, the prior probability of it being true (before doing the study), the statistical power of the study and the level of significance (Ioannidis, 2005; Nuzzo, 2014).

The luring danger to NHST is false-positives (Simmons et al., 2011; Simonsohn, Nelson, & Simmons, 2014). Replication of studies would ideally protect science against false-positives as intuitively a result should only be trusted if it is corroborated by many different studies, such as "*when a pattern is seen repeatedly in a field, the association is probably real, even if its exact extent can be debated*" (Ioannidis, 2008, p. 640). The intuition of this argument is the premise that false-positive findings tend to require many failed attempts at the prevailing rule of a significance level of 0.05. From this perspective, a researcher studying a non-existent effect would on average observe a false-positive finding only once in 20 studies (Simonsohn et al., 2014), and corroborated studies would therefore in fact appear to be true.

However, the seminal work by Ioannidis (2005) in *PLoS Medicine* questioned this premise, an argument which since has been reclaimed in *American Economic Review* (Maniadis et al., 2014). Ioannidis presented a Bayesian argument for why the irreproducibility crisis in science should not come as a surprise, as the false discovery rate (FDR) is likely to be much higher than the assumed ratio of five percent. This claim has received enormous attention from the broad scientific community manifested by over 4,900 Google Scholar citations[2]. The theorem is constructed around the two terms: positive predictive value (PPV) and FDR:

$$PPV = 1 - FDR = \frac{(1-\beta)R + \mu\beta R}{(1-\beta)R + \mu\beta R + \alpha + \mu(1-\alpha)}$$

The PPV is the probability that a finding is true, while the FDR is the probability that a finding is false. The theorem informs us that as either type I error rate (α) increases, the statistical power (1 – β) decreases or the ratio of false to true hypotheses (R) decreases, the FDR will increase. In addition, Ioannidis modelled all sources of bias into a single factor μ, which is the proportion of null hypotheses that would not have been claimed as discoveries in the absence of bias, but which ended up as such, because of it. Thus, if μ increases, the PPV would go down and the FDR would go up. Therefore, with an increasing bias, the chance of a research finding being true diminishes. Table 1 illustrates different calculations of PPV as a function of R and for various settings of β, α and μ to evidence the significance of the ratio of false to true hypotheses (R) and the level of bias (μ), as even with a very low alpha and high statistical power (1 – β), it proves difficult to attain a satisfactory PPV:

[2] Google Scholar – 31/08-2017.

Table 1. The PPV estimate as a function of prior R and for various settings of β, α and μ

	Power 80%, Alpha = 0.01				Power 60%, Alpha = 0.01			
	u = 0.05	*u = 0.1*	*u = 0.3*	*u = 0.5*	*u = 0.05*	*u = 0.1*	*u = 0.3*	*u = 0.5*
R	PPV				PPV			
0.1	0.58	0.43	0.22	0.15	0.51	0.37	0.19	0.14
0.2	0.73	0.60	0.36	0.26	0.68	0.54	0.32	0.24
0.3	0.80	0.69	0.46	0.35	0.76	0.64	0.41	0.32
0.4	0.84	0.75	0.53	0.42	0.81	0.70	0.48	0.39
0.5	0.87	0.79	0.58	0.47	0.84	0.75	0.54	0.44
0.6	0.89	0.82	0.63	0.52	0.86	0.78	0.58	0.49
0.7	0.91	0.84	0.66	0.56	0.88	0.80	0.62	0.53
0.8	0.92	0.86	0.69	0.59	0.89	0.82	0.65	0.56
0.9	0.92	0.87	0.72	0.62	0.90	0.84	0.68	0.59
	Power 80%, Alpha = 0.05				**Power 60%, Alpha = 0,05**			
	u = 0.05	*u = 0.1*	*u = 0.3*	*u = 0.5*	*u = 0.05*	*u = 0.1*	*u = 0.3*	*u = 0.5*
R	PPV				PPV			
0.1	0.45	0.36	0.20	0.15	0.39	0.31	0.18	0.13
0.2	0.62	0.53	0.34	0.26	0.56	0.47	0.30	0.23
0.3	0.71	0.63	0.44	0.34	0.66	0.57	0.39	0.31
0.4	0.77	0.69	0.51	0.41	0.72	0.64	0.46	0.38
0.5	0.81	0.74	0.56	0.46	0.76	0.69	0.52	0.43
0.6	0.83	0.77	0.61	0.51	0.79	0.73	0.56	0.48
0.7	0.85	0.80	0.64	0.55	0.82	0.76	0.60	0.52
0.8	0.87	0.82	0.67	0.58	0.84	0.78	0.63	0.55
0.9	0.88	0.84	0.70	0.61	0.85	0.80	0.66	0.58

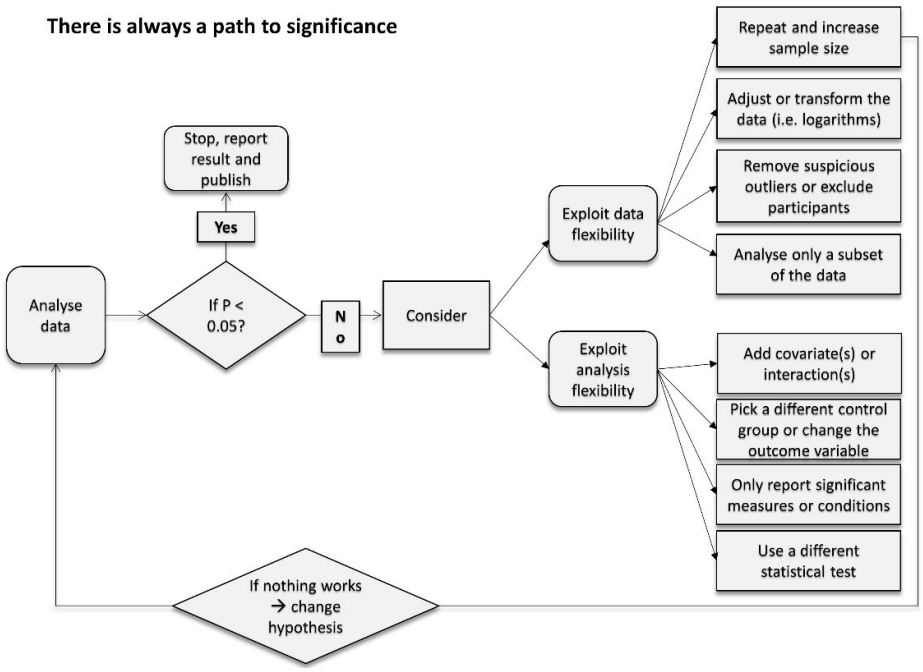
The argument by Ioannidis (2005) has been further substantiated by evidence produced by simulations demonstrating that, in a singly study, only a few changes in data analysis decisions could increase the FDR to approximately 60 percent (Simmons et al., 2011). The study shows that P-hacking activities upends the assumption about the number of failed studies that is required to produce a false-positive finding. Furthermore, Simmons et al. (2011) argued that P-hacking, in the form of undisclosed data flexibility in collection and analysis, allowed them to present almost any hypothesis as significant.

P-hacking should be understood as a term that is confined by a researcher's collective actions taken to exploit ambiguity in a pursuit for statistical significance (Halsey, Curran-Everett, Vowler, & Drummond, 2015; Motulsky, 2015; Nuzzo, 2014; Simmons et al., 2011); in other words, P-hacking is any and all post-hoc changes to the data analysis. Figure 1 demonstrates the process of P-hacking.

P-hacking seems to have emerged as a long-term consequence in a world of publication-counting universities along with a positive publication bias in academic outlets. These two factors provide near perfect incentives for researchers to maximise publications by chasing their data for significance. To protect science against P-hacking, it has been suggested that transparency be increased in terms of providing access to raw data and data analysis (Nuzzo, 2014); the American Statistical Association further argues that proper inferences require full reporting and transparency:

> *"P-values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain p-values… renders the reported p-values essentially uninterpretable…. data dredging, significance chasing, significance questing, selective inference and "p-hacking," leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided… Whenever a researcher chooses what to present based on statistical results, valid interpretation of those results is severely compromised if the reader is not informed of the choice and its basis. Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted and all p-values computed. Valid scientific conclusions based on p-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including p-values) were selected for reporting."*
>
> (Wasserstein & Lazar, 2016, pp. 9-10)

Figure 1. What counts as P-hacking? Any and all post-hoc changes, as with enough flexibility, there will always be a path leading to significant effects



(Motulsky, 2015; Simmons et al., 2011)

2.2 HARKing: 'Hypothesising After the Results are Known'

"A reader quick, keen and leery
Did wonder, ponder and query
When results clean and tight
Fit predictions just right
If the data preceded the theory"
Anonymous (Kerr, 1998, p. 196)

In psychology research, HARKing has been a topic of debate for quite some time (Kerr, 1998). HARKing involves generating a hypothesis from the dataset, by uncovering an intriguing significant relationship, typically through innovative statistical analyses (Motulsky, 2015), and then presenting it as an *a priori* hypothesis (Anonymous, 2015; Chambers et al., 2014; Kerr, 1998). The HARKing debate has since reached general management research, where a provocation and provocateur's piece in *Journal of Management Inquiry* discusses the process and ethics of HARKing. The piece is an anonymous author's reflection on the ethical aftermath of engaging in HARKing[3]. The author recounts how he and his co-authors were able to find intriguing significant results through an innovative analytical approach but were unable to find support for their original *a priori* hypotheses (Anonymous, 2015). After finding these intriguing results, they rewrote the article as if they had formulated these new hypotheses in advance, justifying this by acknowledging that everybody does it (Butler et al., 2017). However, in the end, the anonymous author felt uncomfortable with this, but one of his co-authors was not tenured, and for his sake, the anonymous author felt highly motivated to see the paper published. The paper ended up being published in an A-journal in their field.

When conducting HARKing, the researcher uses the same dataset for generating the hypothesis and testing it (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009). But, a hypothesis should not be tested with the same data from

[3] The identity is known to the editors.

which it was derived; this is essential scientific misconduct, as such hypotheses would be no more than empirical findings disguised as hypotheses (Leung, 2011). In blunter terms, Garud (2015, p. 452) frames it as: "*the practice of presenting post hoc hypothesis as a priori can end up compromising what we know. That is, epistemology compromises ontology. Specifically, this practice can increase the possibility of Type I and Type II errors, misrepresent the truth-value of hypotheses that never were*". HARKing allows for fictitious results becoming an immutable truth, which in the long run could contaminate the entire knowledge pool of a scientific field (Garud, 2015). HARKing also counteracts the communication of valuable information about what did not work, which can challenge the development of valid scientific theories (Kerr, 1998).

HARKing appears to be researchers' long-term response to publication pressure along with a positive publication bias where significant outcomes are more likely to be published. (Francis, 2014; Garud, 2015; Ioannidis, 2005; Simmons et al., 2011). Nevertheless, HARKing is still a practise that violates the percepts of basic scientific method but appears to be accepted as normal, and which is practically impossible to verify empirically (Garud, 2015).

2.3 Publication bias: possible misrepresentation of reality

Significance testing of null hypotheses has long been considered a touchstone for establishing knowledge, and a publication bias appears to have emanated from holding this episteme (Hubbard & Lindsay, 2013; Lindsay, 1994; Sterling, 1959). The bias originates from journals that reject manuscripts because they fail in attaining statistical significant results or in removing insignificant findings during the review process (Chambers et al., 2014). Allegedly this situation has prompted a '*file drawer problem*' where researchers' filing cabinets are believed to be chuck-full of insignificant findings (Simonsohn et al., 2014); consequently, failure to report non-supported hypotheses may lead others to continue their efforts to test the very same hypotheses in subsequent research (Bedeian, Taylor, & Miller, 2010; Garud, 2015; Greenwald, 1975). In the end, false-positives may potentially emerge from the continuous efforts in trying to prove the same hypotheses thereby causing misrepresentation of reality, as reality would predominantly consist of significant findings (Ioannidis, 2008)

Statistical significance is not always a prerequisite, nor is it ever sufficient for establishing research findings as scientifically valuable or meaningful (Gigerenzer & Marewski, 2015; Lindsay, 1994; Ziliak, 2016). This is due to statistical significance not being equivalent to scientific, human or economic significance (Wasserstein & Lazar, 2016); the basis for publication should therefore not be statistical significance but rather the scientific significance of a finding. A positive publication bias is therefore, in truth, an undesirable situation for science. Lykken (1968, pp. 158-159) shares this perception: "*The moral of this story is that the finding of statistical significance is perhaps the least important attribute of a good experiment; it is never a sufficient condition for concluding that a theory has been corroborated, that a useful empirical fact has been established with reasonable confidence – or that an experimental report ought to be published*".

The episteme of NHST has created frustration and tension among journals and editors and an outgoing editor of Journal of Applied Psychology expressed his dissatisfaction as: "*Perhaps P values are like mosquitos. They have an evolutionary niche somewhere and no amount of scratching, swatting, or spraying will dislodge them…Investigators must learn to argue for the significance of their results without reference to inferential statistics*" (Campbell, 1982, p. 698). Others saw NHST as a decline in statistical thinking and blamed it on Fisher: "*Sir Ronald has befuddled us, mesmerized us, and led us down the primrose path. I believe the almost universal reliance on merely refuting the null hypothesis… is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology*" (Meehl, 1978, p. 817). In an effort to oppose the dominance of NHST, the editors of *Basic and Applied Social Psychology* decided to ban P values (Wasserstein & Lazar, 2016) and it is not the first time in history that a scientific journal has tried an approach of banning P value. *The New England Journal of Medicine* (in the 1970s), *Epidemiology and the American Journal of Public Health* (in the 1990s) and the *Publication Manual of the American Psychological Association* have all experimented with bans (Ziliak, 2016).

The common thread of these claims and bans is that we, as researchers, must be able to argue for the scientific significance of our findings (Ziliak & McCloskey, 2004a, 2004b), instead of indulging in what is described by Gigerenzer and Marewski (2015) as *mindless statistical inference*. In the end, a systematic omission of insignificant results would impede the advancement of a scientific field as these findings can contribute to science.

2.4 Lack of data sharing: A neglected possibility to increase integrity and credibility

The natural sciences have always acknowledged reproducibility as a cornerstone of scientific practice and a fundamental form of validation. It is therefore alarming for the validity of science when natural scientists recognise that there are issues as regards the reproducibility of published results, an awareness epitomised by a series in *Nature*

entitled "Challenges in irreproducible research"[4] and by the "Reproducibility Initiative"[5], a project intended to identify and reward replications (Halsey et al., 2015). In self-correcting science, valuation of replications is essential, and data sharing might therefore be an all-important aspect of research conduct, as it would allow researchers to verify original analyses, conduct novel analyses or carry out meta-analyses that can corroborate the reliability and magnitude of reported effects (Hardwicke et al., 2014; Tenopir et al., 2011).

In an effort to illustrate how data sharing could corroborate research findings, Silberzahn and Uhlmann (2015) recruited 29 research teams and asked them to answer an identical research question with an identical dataset. What was soon to be evidenced was that research teams approach a research question and a dataset in different ways, and interestingly, produce highly varying results. The research teams were asked to investigate whether dark-skinned players were more likely to receive a red card in football than white-skinned ones. Of the 29 research teams, 20 teams found a statistically significant correlation between skin colour and red cards. However, the findings varied enormously in effect sizes, from a slight (non-significant) tendency to a strong (significant) tendency for referees to give more red cards to dark-skinned than white-skinned players. The variation in results indicates that any single team's results are highly influenced by their subjective choices taken in the analysis phase, a result also theoretically evidenced through a simulation study by Simmons et al. (2011).

The Silberzahn and Uhlmann (2015) study, published in Nature, illustrates that data sharing would in fact not only reinforce our research findings through corroboration but also allow self-correcting irreproducible findings. On the other hand, a lack of data sharing would impede the detection of QRPs and prevent detailed meta-analyses with the overall result of impeding the replication of previous findings. If journals begin promoting a culture for data sharing or retention, we could increase transparency and reproducibility of the scientific process (Munafò et al., 2014).

2.5 Lack of replication: The hallmark of science is self-correction

The importance of replication can be learned from an age-old prayer "*Lord, protect us from what we only think we know*" and from an equal ancient hypothesis that God only helps those who help themselves (Kane, 1984). Replication is what ensures credibility and integrity in research, it is about creating rigorous theories or hypotheses by opposing the accumulation and dissemination of false knowledge (Merton, 1942, 1973).

By convention, 'true' or genuine replication requires a process where the exact same finding is re-examined in the exact same way (Moonesinghe, Khoury, & Janssens, 2007). Often, therefore, genuine replication is impossible, and instead corroboration or indirect supporting evidence are more realistic (Goodman, Fanelli, & Ioannidis, 2016). In financial accounting, a replication is defined as: "*redoing the identical study in the same way but for another sample period, or periods*" (Dyckman & Zeff, 2014, p. 698). A corroboration approach might be the only sustainable way of doing replications in the social sciences; however, according to Moonesinghe et al. (2007), a lurking danger is the tiny distance from this type of replications to QRPs, which in the end might contribute to 'pseudo' replications, which are false-findings corroborated by other false-findings.

In the social sciences it is rather rare to see a replication being published, which is typically accredited to two main factors (Dyckman & Zeff, 2014; Goodman et al., 2016). First, the high cost of searching for errors in empirical research; and second, it is notably less rewarding in terms of reputation and ability to publish. According to Kane (1984, p. 3), choosing a task of replication is "*widely regarded as prima facie evidence of intellectual mediocrity, revealing a lack of creativity and perhaps even a bullying spirit*". Status and promotion are only granted to the publication which is published first (Dyckman & Zeff, 2014; Gigerenzer & Marewski, 2015). Nevertheless, reproducibility remains the cornerstone of the hypothetico-deductive method, because if an empirical finding is to be considered a 'fact', other researchers must be able to observe it, thus strengthening the credibility of the 'fact' (Kane, 1984).

Maniadis et al. (2014) provided theoretical evidence for the importance of replications by demonstrating through Bayesian statistics that a few independent replications dramatically increase the chances of an original finding being true, and they therefore claim that replications are the best solution to the current inference problem. On the other hand, a limited number of replications would force researchers to generalise from a small number of published studies consequently limiting the ability of research to accumulate 'true' knowledge, in particular if the studies did not 'tell the whole story', or proved unreliable (Bamber, Christensen, & Gaver, 2000). If we consider the importance of replications in creating rigorous theories or hypotheses, it should be defined as *the* 'scientific gold standard' (Jasny, Chin, Chong, & Vignieri, 2011).

[4] For more information on the series see http://www.nature.com/content/nature/24974-01.html
[5] For more information on this project see http://validation.scienceexchange.com/#/

3 Research Method

3.1 Previous attempts at identifying the prevalence of QRPs in published research

Various attempts have been made at uncovering an approximation of the 'true' ratio of QRPs in published research.

For instance, O'Boyle et al. (2017) found that in management research the ratio of supported to unsupported hypotheses more than doubled from defended dissertation to journal publication. They evidenced that the increase in predictive accuracy was a result of QRPs through the practices of dropping nonsignificant hypotheses, the addition of significant hypotheses, the reversing of predicted direction of hypotheses and alterations to the data. Another example of QRPs is from a survey of over 2,000 academic psychologists at major U.S. universities, which found that almost half admitted to have selectively decided to report studies that 'worked' or decided to collect more data after having examined whether the results were significant (John et al., 2012). A study by Jager and Leek (2013) produced a conservative estimate of the ratio of false-positive findings in published studies in five major medical journals. Their meta-study found the false-positive ratio to be 14 percent (Jager & Leek, 2013), however, in a commentary Ioannidis stated that Jager and Leek (2013) fall into the same "false result" category because "their approach is flawed in sampling, calculations and conclusions" (Ioannidis, 2014). Another commentary by two statisticians substantiated Ioannidis's claims, arguing that the false discovery rate is probably closer to 30 or even 50 percent when adjusting for the sample (Benjamini & Hechtlinger, 2014). As a final example of QRPs, a recent survey of 1,500 researchers, initiated by Nature, investigated the current reproducibility crisis in the natural sciences (Baker, 2016). They found that more than 70 percent of researchers have failed in reproducing another researcher's experiment, and more than half had failed in reproducing their own results. It was therefore not surprising, when 52 percent argued for a significant replication crisis, while 38 percent saw it as only a slight crisis, and 7 percent argued for the non-existence of a crisis (3 percent did not know). When asked which factors contributed to irreproducible research, the main contributing factors argued were selective reporting, pressure to publish, low statistical power and/or poor analysis, etc.; and when asked which factors could boost reproducibility, a better understanding of statistics was emphasised by more than 80 percent of the respondents.

These studies illustrate that in many scientific fields QRPs are a reality, but the exact extent of this QRPs is unknown.

3.2 What are the characteristics of a publication practice that allows the existence of QRPs?

Based on section two, we can identify a set of characteristics that are necessary conditions in a publication system for QRPs to take root, and if the publication system of PMAR exhibits such characteristics, we can expect that it is already being distorted by QRPs.

We use descriptive statistics to visualise whether these features are present in PMAR, or at least in the journals analysed. The method is fairly elementary but under the right circumstances rather effective (Gigerenzer, 2004; Gigerenzer & Marewski, 2015) and has been used successfully (Dyckman & Zeff, 2014; Matthes et al., 2015; McCloskey & Ziliak, 1996; Ziliak & McCloskey, 2004b).

The framework looks at the frequency in experimental design, sample type, type of participants, sample selection, sample size, response rate, publication bias, and P values. A set of survey questions has also been developed from the theory presented on QRPs in the second section of the paper; for each article, the questions must be answered either with a yes or a no. By combining the survey questions and ratios, the aim is to shed light on whether the publication system of PMAR is susceptible and provides life-support for QRPs. In the following we will explain the relevance of each survey question.

1) *Does the paper replicate a former study to corroborate its findings?*
   Maniadis et al. (2014) have evidenced that a few independent replications will dramatically increase the chances that original statistical findings are in fact true, as replications decrease the FDR. In this analysis, we follow the broader definition of Dyckman and Zeff (2014) stating that a replication should redo an identical study in the same way but for another but similar sample.
2) *Does the paper disclose information on data availability?*
   Promoting a data sharing or retention culture would bring transparency and reproducibility to the scientific process, as data sharing or retention would allow other researchers to corroborate or self-correct irreproducible findings (Munafò et al., 2014). It is therefore of interest to investigate whether journals are dedicated to this practice.
3) *Does the paper state the statistical power of the test?*

Statistical power is the likelihood that a study will detect an effect when there is an effect to be detected. High statistical power therefore reduces the probability of making a type II error and is directly related to the FDR (Halsey et al., 2015; Ioannidis, 2005; Maniadis et al., 2014).

4) *If the paper mentions power, does it then examine the statistical power of the test?*
If the paper does indeed mention statistical power, how many papers do actually examine it? It would be sound scientific conduct if statistical power was investigated *a priori* to determine the required sample size for detecting the expected effect size (Matthes et al., 2015; Simmons et al., 2011).
5) *Does the paper engage in "sign econometrics"?*
"Sign econometrics" is about stating the direction of the coefficient but not its size (McCloskey & Ziliak, 1996; Ziliak & McCloskey, 2004b). However, 'sign' is not economically significant unless the magnitude is large enough to matter, and statistical significance does not indicate whether it is large or small (Carver, 1993; Sullivan & Feinn, 2012; Wasserstein & Lazar, 2016; Ziliak, 2016). Low P values do not necessarily imply large or more important effects (Wasserstein & Lazar, 2016). Coefficients and effect sizes should therefore be carefully interpreted in relation to the hypothesis under investigation to determine whether the significant statistical findings have scientific or economic relevance. Specifically, an effect size provides quantitative information about the magnitude of the relationship studied; it is therefore much more precise than making a qualitative statement saying that "X increases positively with Y" (Halsey et al., 2015; Hubbard & Lindsay, 2013).
6) *If the paper discusses coefficients, does it then provide confidence intervals for the coefficients?*
Confidence intervals also play an important part in interpreting the relevance of coefficients as they convey what a P value does not, namely the magnitude and the relative importance of an effect (Nuzzo, 2014). Specifically, by presenting the range within which the true effect size is likely to lie, a confidence interval indicates the uncertainty of a measure. Ziliak and McCloskey (2004a, p. 673) refer to Jeffrey Wooldridge on the importance of these two measures. In his textbook *Introductory Econometrics*, Wooldridge suggests that "*Sign without size, and sign without size without confidence intervals, is mainly beside the point*". By reporting effect sizes and confidence intervals, the statistical interpretation of data emphasises both the importance and precision of the estimated effect size, which, in the end, allows for the importance and relevance of the effect to be judged (Halsey et al., 2015).
7) *Does the paper carefully compare and discuss the findings with previous similar studies?*
This question concerns the comparison of sign and effects with previous similar studies, because when findings are confirmed by an independent study, the credibility of the inference made is strengthened. This is also what Maniadis et al. (2014) describe as a solution to the current inferential problem of science; however, it would require that replications are common practice in the scientific field.

3.3 The dataset: published Positivistic Management Accounting Research papers

Our journal selection is focused on journals in which PMAR has been prominently published and it reflects two leading journals according to journal rankings and accounting faculty surveys (Ballas & Theoharakis, 2003; Bonner, Hesford, Van der Stede, & Young, 2006; Lachmann et al., 2017). The period is selected based on the argument that statistical validity has increased over time (Lachmann et al., 2017) and we wish to investigate the prevalence of QRPs in present-day research. We therefore analyse the publication practices of the journals of *Accounting, Organization and Society* (AOS) and *Management Accounting Research* (MAR) in the period 2010-2015 in order to determine if their practices allow QRPs. The approach of this paper follows that of Dyckman and Zeff (2014) in terms of data sampling and selection of articles. We chose to restrict the selection of articles to include only survey and experimental research papers. Consequently, the following types of papers are excluded: (i) theoretical papers, (ii) qualitative papers, (iii) archival papers, (iv) Bayesian estimation methods, and (v) purely explorative papers identifying constructs. This screening leaves us with 38 survey papers and 36 experimental papers, with experimental research being predominantly within AOS (33 out of 36) and survey papers predominantly within MAR (24 out of 38). The papers analysed are presented in appendix A. The sample of articles from two of the leading accounting research journals is argued to be representative of recently published research – an argument also claimed in a similar study by Dyckman and Zeff (2014). The data produced from the articles is a result of direct examination and not from using an electronic search engine. An extract of the coding of the papers is presented in appendix B; by providing the coding of each paper we try to provide transparency on the meta-analysis.

It should be noted that management accounting research is a subset of accounting research, just like financial accounting and auditing research. These accounting research subfields overlap and are not very distinguished, which is why it may be claimed that some of the articles analysed are part of another subset. However, we do not necessarily see

this as problematic for the analysis, because the research methods applied will be more or less equivalent due to being published in the same outlets, and the fields draw on the same theoretical base.

4 Analysis

The structure of the analysis follows the structure of the framework presented and discussed in the previous section. It is important to stress that it is neither the intention nor the purpose of this article to claim researcher malfeasance, but to analyse and discuss our current publication system and the related potential danger of QRPs. An extract of the data material for the figures and tables is provided in appendix B. The results for survey studies are presented in Figure 2 and Table 2 while the results for experimental research are presented in Figure 3 and Table 3.

4.1 Results from survey studies

The results indicated a presence of a positive publication bias in survey research, as 24 out of 38 articles confirmed more than 70 percent of their hypotheses and 34 percent confirmed all of their hypotheses (13 out of 38 articles). Furthermore, none of the published studies confirmed less than 40 percent of their hypotheses. In total, 71 percent (164 of 232) of all stated hypotheses were found significant and thus claimed to be true.

Concerning data sharing, not a single survey study provided access to the raw data, which unfortunately is not a surprising finding. Sharing raw data in management accounting research is uncommon as raw data tends to be protected due to confidentiality reasons. However, anonymising raw data might solve the confidentiality issues, but raw data also represents a publication value, illustrated by Bedford (2015) and Bedford and Malmi (2015) using the same survey raw data; both studies are published in MAR. Considering the common saying 'publish or perish', it would, in this light, be rather unwise to share raw data as it provides an opportunity for additional publications and hence career advancement.

Another finding is that replications are a rather rare exemption, as only one study claimed to reproduce or corroborate a previous finding of a similar study. Thus, of the 164 significant hypotheses, it would appear from this meta-analysis that not one single hypothesis has been genuinely replicated or replicates a previous empirical finding. The lack of replications is also illustrated by the fact that not a single study carefully compares coefficients and effect sizes with previous empirical findings. The scarcity of replications within AOS and MAR could be attributed to their high-ranking status, by only publishing what they perceive as novel findings. Similarly, by analysing *The Accounting Review* and *Journal of Accounting Research*, a study by Dyckman and Zeff (2014) found a clear lack of replications in financial accounting, which is a situation that appears to be symptomatic for social science as a whole (Gigerenzer & Marewski, 2015; Goodman et al., 2016; Yong, 2012).

So far, it appears that the publication environment for survey studies in PMAR unintentionally incentivise researchers' engagement in QRPs as it provides the space for P-hacking and HARKing to occur. This is due to a positive publication bias, a lack of replications and no raw data transparency. The likelihood of a published false-positive to be falsified or self-corrected is therefore, unfortunately, close to zero.

We now look at the flexibility in sampling and the process of analysing and reporting the data. We found that the typical study, 30 percent of survey studies, has a sample size in the interval of 51-100 respondents, while 66 percent are below 200 respondents. Furthermore, the typical sample type is a non-random sample by approximately 74 percent. Van der Stede et al. (2005) claim, as a rule of thumb, that a survey study should at least attain a sample size between 200-300 respondents to ensure an acceptable level of validity, which is higher than approximately 61 percent of the studies in this analysis. Considering the relatively small sample sizes, it would have been appropriate that the survey studies had estimated their statistical power. However, only 21 percent mention statistical power, and only 11 percent examine the power of the test. This situation is worsened by the effect sizes within management accounting typically being quite small, and the smaller the effect size, the larger the statistical power required (Borkowski, Welsh, & Zhang, 2001; Sullivan & Feinn, 2012).

In theory, a statistical study should a priori calculate the sample size needed by reflecting on the expected effect sizes to ensure that the statistical power is high enough for detecting the expected effect sizes. Henri and Journeault (2010, p. 69) follow this best practice: "*In the current study, the sample size is adequate to test the proposed model (n = 303) as well as the ratio of observations per parameter. Furthermore, based on the guidelines of MacCallum, Browne, and Sugawara (1996), this study has adequate statistical power (i.e. 0.93)*". Power analysis can also be used to *post hoc* analyse if a study correctly rejects a hypothesis. Artz, Homburg, and Rajab (2012, p. 454) illustrate this: "*As a valid theoretical rationale exists for expecting a significant positive interaction effect, an important question is whether our sample size is powerful enough to find an existing effect. Therefore, we analysed whether our sample has adequate statistical power to reject the null hypothesis of no effect… We found that we can detect a true effect size of .050 with about 90% power and a true effect size of .037 with about 80% power… The significant interactions (table 4) have effect sizes greater than .050. Therefore, we conclude that if an effect exists, our setting is powerful enough to find it. We still have to reject H2a*". Unfortunately, these two studies appear to represent a rarity in PMAR, and considering the

influence of statistical power on the FDR, it is worrying that statistical power receives so little attention from authors and reviewers.

Concerning bias, the response rate for the analysed survey studies typically ranges from 20 to 40 percent, representing 24 of 38 studies, and it is a well-known fact that a low response rate could induce a *non-sampling error* issue and a *non-response bias*. So, when a sample size is secured to be large enough for the statistical power to be satisfactory, any efforts left should be moved to increase the response rates (Van der Stede et al., 2005). A low response rate represents a potential for increased bias in the PPV equation resulting in an increased ratio of false-positives.

Figure 2. Dashboard of results from analysing 38 survey studies in AOS and MAR from 2010 to 2015

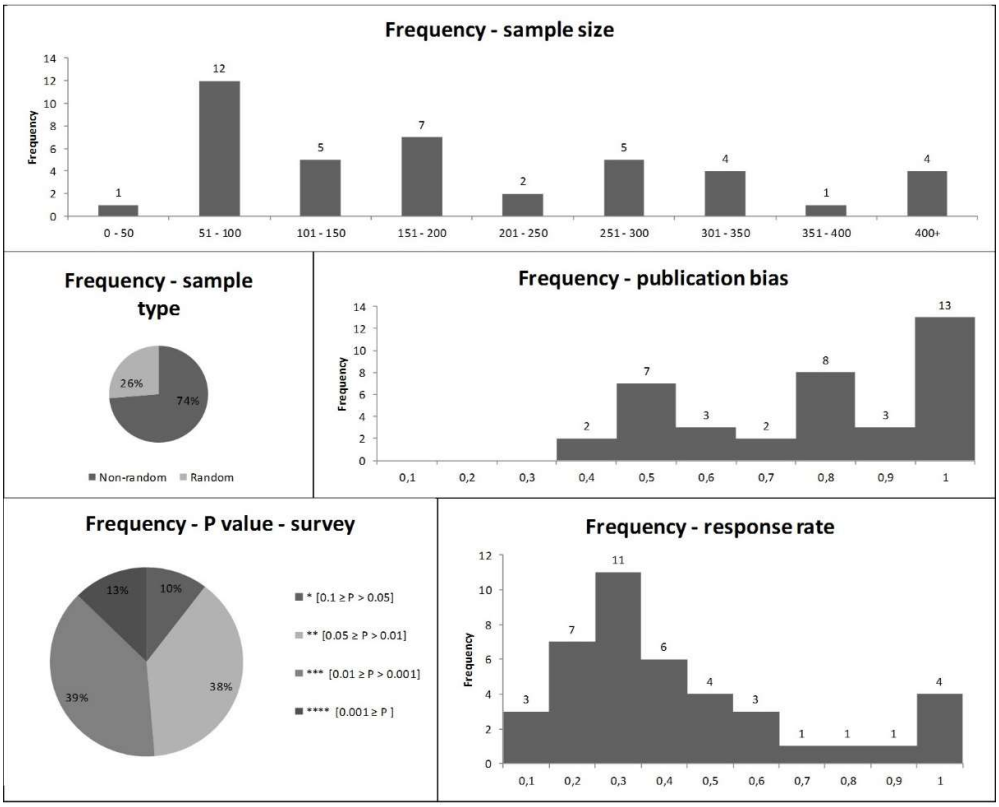


Table 2. Findings for survey studies in AOS and MAR

Survey Questions	Yes	No
1. Does the paper replicate a former study to corroborate its findings?	3%	97%
2. Does the paper disclose any information on data availability?	0%	100%
3. Does the paper state the statistical power of the test?	21%	79%
4. If the paper mentions power, does it then examine the statistical power of the test?	11%	89%
5. Does the paper engage in "sign econometrics"?	84%	16%
6. If the paper discusses coefficients, does it then provide confidence intervals for the coeff	5%	95%
7. Does the paper carefully compare and discuss the findings with previous similar studies?	0%	100%

Source: All the full-length papers using tests of statistical significance and published in *AOS or MAR* between 2010 and 2015
N (AOS) = 14, N (MAR) = 24, Total = 38 articles

We now analyse the practice of statistical reporting. We found that 84 percent of the papers engage in "sign econometrics", that is remarking on the direction of the effect but not its size. This is understood as papers basing the relevance of their findings on whether they were significant in the predicted direction instead of discussing the magnitude of the relationship studied. This practice allows P-hacking to occur, as P-hacking is about searching for significance with no regard to effect sizes. To illustrate the typical ways of reporting statistical evidence in the articles analysed, we present a set of quotes that are representative for the sample.

Bisbe and Malagueño (2012, p. 305) report their results as: "*Panel A in Table 3 displays the results of the causal steps procedure. It shows that SPMS have a positive effect on the strategic decision array variety ($p < 0.01$), which in turn, has a positive effect on organisational performance measured through ROS ($p < 0.05$). Analogously, results show that SPMS have a positive effect on the strategic decision array size ($p < 0.01$) as well as on ROA ($p < 0.05$). Overall, these results suggest that, as predicted by H1a, SPMS are positively associated with the comprehensiveness of the strategic decision arrays*". This quote illustrates a typical practice of reporting coefficients and effect sizes in tables but refrains from commenting or reflecting on the effect sizes. Another example of "sign econometrics" is from Ho, Wu, and Wu (2014, p. 48): "*Specifically, employees' tenure (EMP_TENURE) is positively and significantly associated with customer satisfaction (0.02, $t = 2.97$, $p < 0.01$ in Model 1 and 0.02, $t = 3.19$, $p < 0.01$ in Model 2), which suggest that senior salespeople provide more satisfying service and have earned greater trust and loyalty from their customers*". The researchers refrain from discussing the practical or scientific relevance of their findings despite their coefficients appearing to be very small and perhaps rendering the results irrelevant.

However, not all papers omit discussions of effect sizes. An example is Guerreiro, Rodrigues, and Craig (2012, pp. 493-494) who apply an *odds ratio* approach to differentiate between significant predictors by claiming that the most important predictors are those that change the odds of the outcome the most: "*As column Exp(B) of Table 4 reveals, firms with listed parent companies are 12.93 times more likely to adopt IFRS voluntarily than are firms with unlisted parent companies*". This study found the calculated odds ratios to range from 0.173 to 12.925 with all predictors being statistically significant thereby allowing researchers to judge the relative relevance between predictors.

The survey studies analysed argued their relevance on significance and there was an almost complete lack of discussing coefficients and effect sizes. A practice of "sign econometrics" is unfortunate, not only by allowing P-hacking, but also because small P values do not imply importance or relevance. Instead careful consideration of coefficients and effect sizes would be not only a prerequisite for claiming scientific or economic significance but also discourage P-hacking (Wasserstein & Lazar, 2016; Ziliak, 2016; Ziliak & McCloskey, 2004b). The focus on significance is further evidenced by a complete disregard for confidence intervals despite confidence intervals being superior to P values in every way.

Concluding on the findings, the publication practice related to survey studies give rise to concern. The findings indicate that the PMAR publication system on survey studies is susceptible to QRPs and unintentionally incentivise researchers to engage in QRPs. No evidence is found that the publication practice of PMAR is discouraging the activities of *P-hacking* or HARKing. In terms of publication practices and statistical reporting, PMAR therefore appears to be following in the footsteps of social science by allowing the hypothetico-deductive method to be distorted by QRPs. However, methodological critique is not unknown to survey studies (Van der Stede et al., 2005), and many see experiments as a response to this critique; experiments are therefore becoming increasingly popular.

The next subsection investigates if the publication practice of experimental research in PMAR is more resilient to QRPs, which unfortunately has not been the case for other scientific fields (Camerer et al., 2016; Matthes et al., 2015; Simmons et al., 2011).

4.2 Results from experimental research

Our results indicate an even more consistent positive publication bias than observed within survey studies, as 24 out of 36 articles confirm 100 percent of their hypotheses. In total, 86 percent (99 of 115) of all stated hypotheses were found to be statistically significant and claimed to be true. Concerning data availability, it is the same case as for survey studies; not a single study provided access to raw data, although it is quite common for studies to provide a detailed experimental design guide either as an appendix or as online supplementary material. A guide gives researchers the opportunity to replicate experimental findings, however only 6 percent (2 out of 38) claimed to replicate former empirical findings in a new setting by a new experimental design, thereby not representing a 'genuine' replication. This would require the precise same experimental design to be used (Dyckman & Zeff, 2014; Moonesinghe et al., 2007). As noted before, the reason for the lack of replications could, again, be that the prestigious journals prefer novel findings and therefore are reluctant to publish replications. However, another factor is possibly due to experiments within management accounting being in the making thus limiting the number of studies available for replication. Nonetheless, a lack of replications is unfortunate for the accumulation of knowledge, as evidenced by Maniadis et al. (2014) who demonstrate that the FDR diminishes drastically even after a few replications. The lack of replications therefore represents a concern for PMAR. For example, economic research has found itself limited in its ability to reproduce

previous experimental findings, as Camerer et al. (2016) tried to replicate 18 experimental studies published in the *American Economic Review* and in the *Quarterly Journal of Economics*; they only found a significant effect in the same direction as in the original study for 11 replications, and on average the replicated effect size was only 66 percent of the original.

Based on these findings, we draw the same conclusion for the publication environment for experimental studies as we did for survey studies, namely, that it provides the potential to engage in P-hacking and HARKing without researchers having to fear being 'caught'. This is because of a high publication bias, a lack of replications and no transparency on raw data resulting in the likelihood of false-positive to be falsified or self-corrected being, once again, close to zero.

Figure 3. Dashboard of results from analysing 36 experimental studies in AOS and MAR from 2010 to 2015

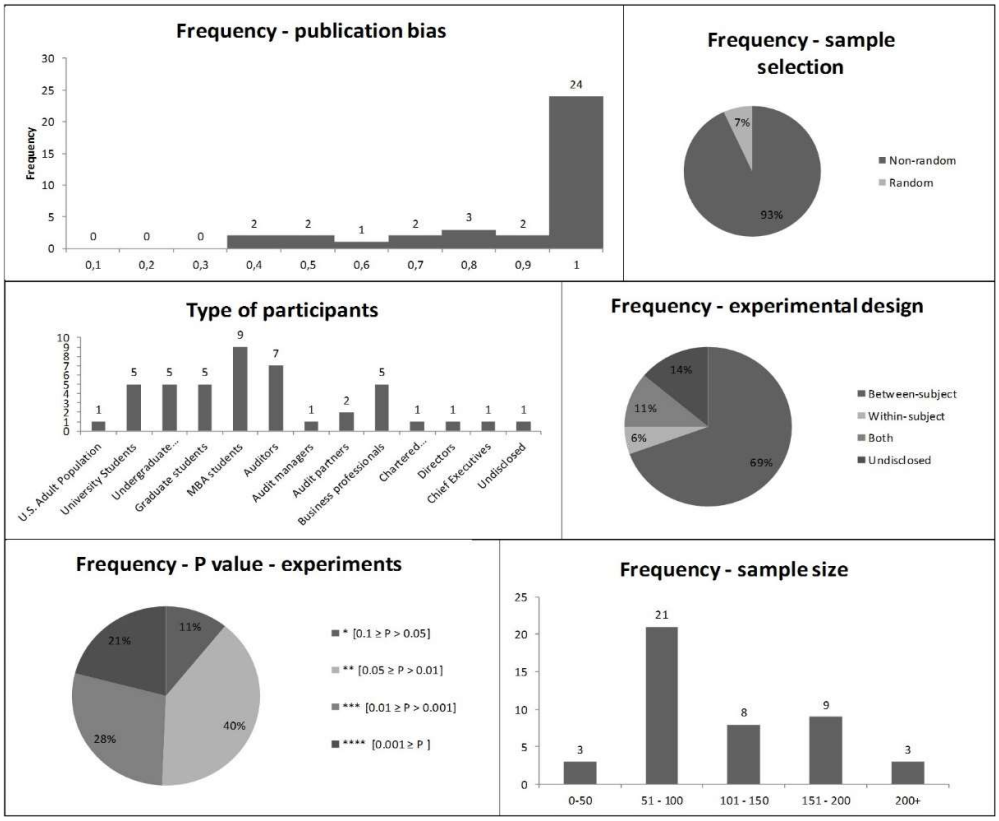


Table 3. Findings for experimental studies in AOS and MAR

Survey Questions	Yes	No
1. Does the paper replicate a former study to corroborate its findings?	6%	94%
2. Does the paper disclose any information on data availability?	0%	100%
3. Does the paper state the statistical power of the test?	8%	92%
4. If the paper mentions power, does it then examine the statistical power of the test?	0%	100%
5. Does the paper engage in "sign econometrics"?	94%	6%
6. If the paper discusses coefficients, does it then provide confidence intervals for the coeffi	0%	100%
7. Does the paper carefully compare and discuss the findings with previous similar studies?	0%	100%

Source: All the full-length papers using tests of statistical significance and published in AOS or MAR between 2010 and 2015
N (AOS) = 33, N (MAR) = 3, Total = 36 articles

We now look at the flexibility in sampling and the process of analysing and reporting the data. We found that 93 percent of the studies use a non-random sample. The typical participant is a student, as university students take part in more than 40 percent of the experiments (15 of 36), while MBA students account for 25 percent (9 of 36). It might be called into question whether these types of participants represent the population of interest. On the positive side, there is also a noteworthy high frequency of auditors (9 of 36).

Concerning the sample sizes, it is surprising that 66 percent of the studies analysed have a sample size below 100 participants; an *a priori* statistical power test would have been appropriate to confer that the study was large enough to detect the expected effect size. However, only eight percent of the studies mention statistical power and not a single study conducts an *a priori* analysis of the sample size needed, and also, no study conducts a *post hoc* statistical power test to uncover if the study failed to reject a H_0 hypothesis. Perhaps this is unsurprising, considering that 86 percent of all hypotheses were found to be significant. The ratio of 86 percent could be claimed to represent an impressive foresight in theoretical forecast, but it is more likely to represent some degree of *P*-hacking or HARKing if the publication environment permits it (Leung, 2011). It is, however, unfortunate that the authors and reviewers put so little emphasis on the calculation of statistical power tests.

We now analyse the practice of statistical reporting: 94 percent of the experimental studies were found to engage predominantly in "sign econometrics", demonstrated by a low emphasis on coefficients and effect sizes despite a wide range of effect size measures for experimental research being available, e.g. Cohen's d, Odds Ratio, Relative Risk or Risk Ratio, Pearson's r correlation, r2 coefficient of determination and Glass' Δ (Sullivan & Feinn, 2012).

To illustrate the typical way of reporting statistical evidence, we, again, draw on a selection of quotes. However, the reporting of statistical evidence is very similar to that in survey research and we therefore only two representative quotes. For example, Perreault and Kida (2011, p. 542) report the statistical findings as: "*Compared to an auditor using a cooperative communication style, those that negotiated with the contentious auditor liked the auditor less (3.13 vs. 4.72; t = 6.58; p < .001), were less happy with the auditor (3.36 vs. 4.84; t = 5.68; p < .001), were more frustrated with the auditor (4.44 vs. 3.59; t = 3.21, p < .001) and were more angry with the auditor (3.46 vs 2.67; t = 3.20; p < .001). As a result hypothesis four is supported*". While Chen and Tan (2013, p. 221) only focused on the P value: "*We find a significant exposure effect on the change in participants' third quarter earnings estimates in the absence of the performance cue (p = 0.04), and an insignificant exposure effect in its presence (p = 0.29)… In Sum Hypotheses 1 and 2 are supported for both credibility and earnings estimates measures*".

Once again, we observe that experimental studies tend to focus on the P values and related significance while refraining from discussing coefficients or effect sizes. In addition, not a single paper calculated and presented confidence intervals as a part of the discussion. Reviewers' reluctance to require a discussion of coefficients is not only unfortunate but also 'poor' statistical reporting, as the *American Statistical Association*, on the matter of QRPs, argues that scientific importance is not solely a matter of significance, but just as much of confidence intervals, coefficients and effect sizes (Wasserstein & Lazar, 2016).

As for survey studies, we also found reason for concern as regards experimental research. On various parameters, experimental research is even more indicative of a publication practice that favours QRPs, as we found a stronger publication bias, smaller sample sizes, bias in use of participants, and an even stronger tendency to engage in "sign econometrics". Based on these observations, we argue that the publication practices of experimental research in PMAR unintentionally encourage the phenomenon of QRPs which is why we are concerned about the size of the ratio of false-positives. As such, the theoretical advantage of experimental research in making causal claims appears to be offset by the likelihood of QRPs taking place. On the bright side, the method allows easier replication and hence the detection of false-positives as it would only require our outlets to begin to publish replications.

4.3 Discussion

The meta-analysis evidences that it is reasonable to argue that the publication practices of PMAR is susceptible to QRPs and, as such, it is likely that our field is also being distorted by this phenomenon. This is based on the identification of a set of current characteristics in the PMAR publication system where we found a positive publication bias, a non-random sampling, a lack of replications, no data sharing, no statistical power tests, and a common practice of "sign econometrics". From the meta-analysis, we highlight the following key ratios:

- High positive publication bias for survey studies [164 of 232 hypotheses confirmed, 71%] and experimental studies [99 of 115 hypotheses confirmed, 86%].
- Non-random sampling for surveys [74%] for experimental [93%].
- Sample size for survey studies [66% below 200 respondents and response rate of 20-40%] for experimental studies [typically 51-100 participants, and the typical participant is a student].
- A complete lack of replications, no data sharing and a negligence of statistical power tests [either a priori or post hoc].

- Common engagement in "sign econometrics" for survey studies [84%] and experimental studies [94%], while disregarding presentation and discussion of confidence intervals for survey studies [5%] and experimental studies [0%].
- A complete lack of comparing findings [coefficients and/or effect sizes] with previous, similar studies.

Research has evidenced the existence of QRPs in the social sciences and the natural sciences, and we found no evidence suggesting that QRPs have not taken foothold in PMAR. It is therefore rational to expect that the false-positive ratio is significantly higher than the conventional ratio of five percent; however, the precise extent of false-positives is impossible to verify without replications. We therefore expect PMAR to be producing research findings when they should not be produced, supporting a major concern, raised by Ioannidis (2005), for statistical research as a whole.

Our findings outline a different reality for the validity of PMAR than the one painted by Lachmann et al. (2017). However, we very much agree that in terms of those characteristics investigated by Lachmann et al. (2017) the validity of PMAR has been increasing over the last four decades. On the other hand, we argue that the validity of PMAR cannot be discussed without taking QRPs into account. It is therefore not enough to discuss only the criteria of internal validity, external validity, construct validity and statistical validity[6]. This will not provide a holistic picture of the validity of PMAR, i.e. the ability to replicate original findings. If we are right about the PMAR publication practices of, then it amounts to a serious problem for the reliability of PMAR in terms of whether research findings can be considered as 'true' (Ioannidis, 2005; Ioannidis & Doucouliagos, 2013; Young, Ioannidis, & Al-Ubaydli, 2008); also it indicates a paradox: To live up to the positivistic image of 'pure science' published in academic journals, researchers find themselves – ironically – transgressing this very ideal (Butler et al., 2017).

Genuine replications would have provided hard evidence of the existence of QRPs; unfortunately, this is a type of study that is underappreciated in most social sciences, including PMAR. Self-correction is a scientific feature that seems long forgotten in social sciences as a whole.

To put it bluntly, the publication practice of PMAR can be understood from a column written by Andrew Gelman and Erik Loken where they compare the practice of publishing in science with the subprime mortgage crisis in the United States. Their column is framed 'The AAA Tranche of Subprime Science' and was published in Change:

> *"The first step is statistical significance. Out of the primordial soup of all possible data analyses, the statistically significant comparisons float to the top. They represent the high-certainty statements selected out of the many less-reliable claims. The second step is publication in a scientific journal, ideally a high prestige outlet… -But, if not a top journal any outlet will do. The convention is to treat published claims as true unless demonstrated otherwise. The two-step process – first the achievement of statistical significance, then publication – corresponds with the movement of scientific hypothesis from the hazy zone of uncertain speculation to presumed certainty".*
> (Gelman & Loken, 2014a, p. 51).

They describe a system where the role of statistical significance and the peer review process represent a seal of approval for scientific claim (Bedeian, 2004; Dyckman & Zeff, 2014; Ohlson, 2015). However, this system is allegedly challenged as neither statistical significance nor the peer review process, as currently practiced, work quite as intended (Bamber et al., 2000; Banks, Rogelberg, et al., 2016; Garud, 2015; Maniadis et al., 2014; Starbuck, 2016).

The meta-analysis indicates that the publication practice of PMAR contains similarities with scientific fields that we know are being distorted by QRPs, as data analysis or scientific inference seem to have been reduced to a mechanical 'bright-line' rule, i.e. $P < 0.05$, when justifying for scientific claim. For example, only a very few studies argue for the relevance of their empirical findings on other information than the decisive factor of the P value being below 0.05, while not a single study counter-argues the relevance of a significant hypothesis due to a too small effect size rendering the finding irrelevant. Such a practice has the potential of leading to erroneous beliefs and poor decision-making (Wasserstein & Lazar, 2016). A use of 'significance', as the license for claiming a scientific finding has the potential of leading to considerable distortion in the scientific process of the hypothetico-deductive method. An empirical finding does not immediately become 'true' on one side of the divide and 'false' on the other (Wasserstein & Lazar, 2016) and to quote Sir R. A. Fisher:

---

[6] Internal validity is coded in terms of time frame. External validity is coded in terms of type of sample and primary occupation of participants. Construct validity is coded in terms of number of measures for construct validation, number of reliability measures, type of dependent variables, and number of data sources. Statistical validity is coded in terms of particular tests of multicollinearity, omitted variable bias, simultaneity bias, self-selection bias, heteroscedasticity, outliers and so on (Lachmann et al., 2017).

> *"No scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas."*
> (Fisher, 1956, p. 42).

Relying on significance provides the perfect incentives for P-hacking and HARKing, as invariably there is a data analysis path which leads to significance even in the absence of an underlying effect, and – if not – a researcher can always change hypothesis (Loken & Gelman, 2017). The publication practices of PMAR therefore are in danger of upending the assumption about the number of studies required to produce a false-positive finding, thereby questioning the validity of our knowledge base.

The American Statistical Association elaborates on the meaning of P values arguing that a P value does not measure the probability that a studied hypothesis is true or even relevant (Wasserstein & Lazar, 2016). Mainly reporting and concluding on significance and directional effects instead of arguing for scientific significance or economic significance therefore in reality and in essence constitute statistical misbehaviour (Dyckman & Zeff, 2014; Evans, Feng, Hoffman, Moser, & Van der Stede, 2015; Gigerenzer, 2004; Gigerenzer & Marewski, 2015; Lindsay, 1994; Ziliak & McCloskey, 2004b).

The purpose and ambition of PMAR are to develop reliable and valid causal explanations of management accounting phenomena and thus to draw inferences from a sample of specific observations to the general (Ittner, 2014; Lachmann et al., 2017; Luft & Shields, 2014). The prestigious academic journals of Accounting, Organizations and Society and Management Accounting Research should be beacons of scientific integrity and reliability. It is therefore problematic for PMAR as a whole when we find the publication practices of these two journals to be susceptible and perhaps even (unintentionally) incentivising QRPs.

In the next section, we will put forward suggestions for the adaptation of the publication practices of PMAR towards becoming more resilient towards QRPs hence strengthening the credibility of our scientific field.

5 Conclusion and suggestions

Our findings indicated that the current research traditions of PMAR rendered it possible for QRPs to take root. Based on our findings, we question the reproducibility of PMAR and therefore expect the false-positive ratio to be well above the conventional five-percent ratio. As a result we would expect PMAR to produce research findings when they should not, which was the main concern of QRPs and its distortion on the hypothetico-deductive method first raised by Ioannidis (2005).

5.1 The establishment of a bad equilibrium in PMAR?

Any scientific field should strive towards becoming an accumulative, iterative, self-correcting endeavour, where mistakes, such as false-positives, are a normal short-term effect of a long-term process of accumulating knowledge. However, for this to be the case, a publication system that does not incentivise researchers to stray from this path and indulge in activities of QRPs is required. A scientific field should avoid "sign econometrics", positive publication bias, lack of data sharing and a lack of replications. However, it appears that the 'pressure to publish' combined with the competitive advantage of QRPs have created a situation where:

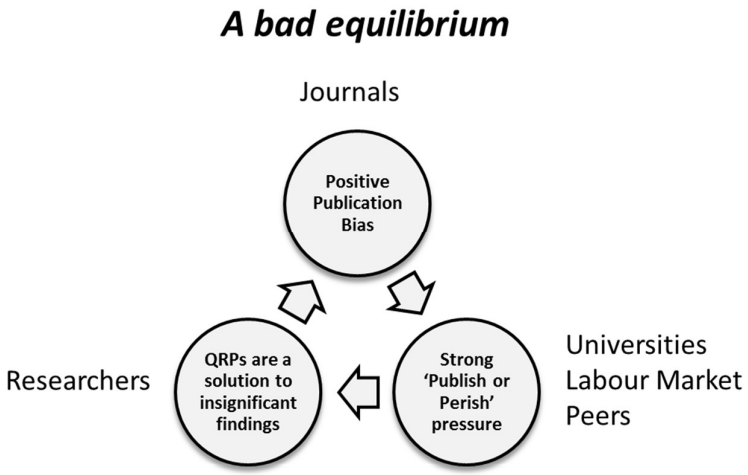
> *"… it is no longer worth questioning questionable research practices (QRPs). If these practices are widespread, then we have all become prisoners of a system that we have created but are unable to free ourselves from. Macro expectations and incentives have transformed into micro motives… Problematic practices that violate the percepts of basic methods appear now to be accepted as normal"*
> (Garud, 2015, p. 452).

We believe that the dysfunctionality of the current publication system in relation to QRPs has permitted a bad equilibrium to unfold (Butler et al., 2017).

The equilibrium is sustained due to the 'publish or perish' pressure and the fact that QRPs are a viable solution to insignificant findings. In other words, in a scientific field where P values prevail, researchers will tend to chase the asterisk that signal 'statistical significance' as it is the catalyst for justifying theoretical assertions or scientific relevance (Gelman & Loken, 2014a). However, a publication system where "statistical significance" + "publication" = "truth", provides the potential for P-hacking or HARKing to be the solution to insignificant findings. The current publication system of PMAR appears to be a system that unintentionally 'promotes' instead of 'intentionally' hinders false-positives (Ohlson, 2015). If this situation does not change, in the long run it will unquestionably lead to devastating

consequences for society, as it may lead researchers, policymakers and funding agencies down false paths, stifling and potentially eroding scientific progress while wasting society's resources (Simonsohn et al., 2014).

Figure 4. A self-sustaining equilibrium



The equilibrium is sustained due to the 'publish or perish' pressure and the fact that QRPs are a viable solution to insignificant findings. In other words, in a scientific field where P values prevail, researchers will tend to chase the asterisk that signal 'statistical significance' as it is the catalyst for justifying theoretical assertions or scientific relevance (Gelman & Loken, 2014a). However, a publication system where "statistical significance" + "publication" = "truth", provides the potential for P-hacking or HARKing to be the solution to insignificant findings. The current publication system of PMAR appears to be a system that unintentionally 'promotes' instead of 'intentionally' hinders false-positives (Ohlson, 2015). If this situation does not change, in the long run it will unquestionably lead to devastating consequences for society, as it may lead researchers, policymakers and funding agencies down false paths, stifling and potentially eroding scientific progress while wasting society's resources (Simonsohn et al., 2014).

Nature and *Science* have united in an effort to break this bad equilibrium by bolstering statistical research. Science is adding an extra round of statistical checks to its peer-review process, which is conducted in collaboration with the *American Statistical Association* and a new *Statistics Board of Reviewing Editors* (SBoRE) consisting of seven expert statisticians (McNutt, 2014). *Nature*, on the other hand, has developed a *statistical checklist*[7] to improve the statistical robustness of results and, furthermore, employs statistical consultants to the review process of certain papers; this is done at the discretion of the editors or, if suggested, by the referees (Van Noorden, 2014).

We hope that in the future PMAR will strive for the same robustness against QRPs. If incorrect causal claims are generalised to practice, it will not only erode scientific progress but be devastating to the society that we as researchers claim to serve.

In the next section, we will present three suggestions for breaking the bad equilibrium.

5.2 How to break the cycle?

The trigger of the bad equilibrium is arguably the "journals" (Butler et al., 2017), so for any solution to be viable, this is where the problem must be solved. At first glance, increasing transparency in mitigating for QRPs would be the natural thing to do as the issue is either misreporting or lack of reporting. In addition, a system built to appreciate reports of magnitudes instead of "sign econometrics" would include resistance towards P-hacking and HARKing; however, this disincentive might not be strong enough. We therefore propose three solutions for implementation by journals; and we will evaluate how realistic each of these solutions is and if their impact is forceful enough to counter QRPs.

Solution 1:

Researchers must submit research projects instead of final papers. Research projects contain research question(s), theory and hypotheses development along with a detailed research design. Data collection and results

[7] http://www.nature.com/authors/policies/checklist.pdf

should be produced after the project has been accepted for publication ensuring that also non-significant results will be published. The approach of this solution would completely discourage any engagement in QRPs; this approach resembles the 2017 *Journal of Accounting Research* (JAR) experiment on "Registered Reports". Whether the approach is realistic is questionable, it would change the current scientific system drastically, but it is bound to break the cycle of the bad equilibrium. It will be interesting to follow the JAR experiment.

Solution 2:

The second proposal concerns the publishing of replications. The knowledge that researchers are likely to reproduce from previous findings would, in the long run, trigger concerns for reputation and hence discourage QRPs. In addition, a few independent replications would also dramatically increase the probability that the original finding is true (Maniadis et al., 2014). On the other hand, we would risk overemphasizing replications thereby shifting attention away from exploiting innovations and novel findings and hence slowing down the progress of research. Furthermore, who would undertake the responsibility of doing replications if it indicates lack of creativity and originality? It might be unrealistic to expect major journals to change their strategy and actively publish replications. However, it might be an opportunity for "second-tier" journals to "move up" if the broader academic public would appreciate replications.

Solution 3:

The third proposal is a system of "variance investigation", which is defined as a system where a paper might be lifted from the archives for replication purposes. If researchers are aware that this might happen and if it is not too unlikely, it would trigger reputation concerns for the researcher if the replication fails to corroborate the original findings. This would introduce a level of inertia for a researcher to engage in QRPs. Also, this system is realistic as journals might use it as a seal of 'quality' thus enabling them to benefit from it.

None of these proposals should stand alone, and we therefore would like to point out the following for consideration independently of the above proposals. First, journals must make sure that their reviewers require authors to use more facets of the statistical toolbox in the argumentation of their results than solely judging on statistical significance. In this way, they would also encourage a discussion of magnitudes when claiming for scientific relevance (Dyckman & Zeff, 2014). Second, disclosing the number of hypotheses explored, all data collection decisions, all statistical analyses conducted and all P values computed is a precondition (Wasserstein & Lazar, 2016). These papers should be branded as '*P certified, not P hacked*' including the following wording: "*We report how we determined our sample size, all data exclusions (if any), all manipulations and all measures in the study*" (Nuzzo, 2014). Third, in preventing publication bias compromising our knowledge base, it is pivotal that research outlets move away from the seemingly positive publication bias: by appreciating the inherent knowledge in insignificant findings and by not encouraging their exclusion in the review process (Banks, O'Boyle, et al., 2016). Fourth, journals should develop a data sharing culture as it would give researchers the opportunity to verify original analyses. Commitment to complete transparency as regards original data is an important property of 'good' science, and journals should therefore develop policies on how to handle data sharing (Munafò et al., 2014; Nosek et al., 2015; Tenopir et al., 2011). Mitigating confidentiality issues trough anonymising raw data ought to be possible; this is already done in medical research on confidentiality concerns of private data. Fifth, a theme also present in previously proposed solutions is replications, being a fundamental requirement for the creation of rigorous theories and opposing the accumulation and dissemination of false knowledge (Merton, 1942, 1973). By pushing for replications, we allow science to be self-correcting and we would conform to a core requirement for claiming causality argued by David Hume in '*Enquiries of Human Understanding*' from 1748:

> *"Even after one instance or experiment, where we have observed a particular event to follow upon another, we are not entitled to form a general rule, or foretell what will happen in like cases; it being justly esteemed an unpardonable temerity to judge of the whole course of nature from one single experiment, however accurate or certain."*
> (Hume, 1975, p. 74)

By following these guidelines and considering the three solutions proposed, it should help ensure that a 'bright-line' rule does not steer the publication process and, in the end, ensure that future policy or business decisions in society are not indirectly based on whether a P value passed a certain threshold. Just because we do not necessarily enjoy the methodological precision as natural sciences does (Aguinis & Edwards, 2014), it does not mean that we cannot strive for the same rigour in statistical method, which is a basic prerequisite for theoretical progress and the accumulation of knowledge.

Acknowledgements
Hanne Nørreklit was my supervisor during both my master's thesis and PhD dissertation and it was her who encouraged me to apply for a PhD after finishing my master's degree at Aarhus University. If it was not for her, I would never have thought about applying for a PhD, as it was never in consideration for my future. However, the journey I was on during my time as a PhD student was a profound experience not only from a professional perspective but also on a personal level.
Hanne Nørreklit is truly one of the most visionary and innovative researchers in the field of management accounting and my collaboration with her during my PhD has been inspiring not only in relation to my PhD but also in my further professional career as an economic consultant at Middelfart Municipality.
My contribution to the 'festskrift' will be my third chapter of my PhD dissertation which I have presented at the 10th conference for New Directions in Management Accounting in Brussel. The paper has also been in a review process in Accounting, Organization and Society as well as Management Accounting Research, but it has unfortunately been outside my reach to refine it in a way that allowed for it to get through the review process.
The paper as it stands now is therefore more or less the same paper as when it was published in my PhD dissertation, and I know that this paper meant a great deal to my supervisor Hanne Nørreklit as it touches and discusses the integrity of science within research in Management Accounting. The paper does not include any research conducted since 2017 as I left academia when I finished my PhD. But it is my humble opinion that the topic the paper touches is still relevant and remains an underdiscussed topic, as it is quite sensitive topic to many researchers.
I therefore dedicate this paper to Hanne Nørreklit, and I am honoured to get the opportunity to contribute to her 'festskrift'.
In addition I would like to thank the participants at the 10th Conference on New Directions in Management Accounting, Brussel 2016, and in particular Frank Moers, Teemu Laine, Tuomas Korhonen, Rafael Heinzelmann, Morten Jakobsen and Hanne Nørreklit for their comments and suggestions.

Appendix A List of the seventy-four articles in the sample
Management Accounting Research (2010 – 2015) – Survey research
- Abernethy, Bouwens, and Lent – Leadership and control system design (2010)
- King, Clarkson, and Wallace – Budgeting practices and performance in small healthcare businesses (2010)
- Hall – Do comprehensive performance measurement systems help or hinder managers' mental model development? (2011)
- Lee and Yang – Organization structure, competition and performance measurement systems and their joint effects on performance (2011)
- Weißenberger and Angelkort – Integration of financial and management accounting systems: The meditating influence of a consistent financial language on controllership effectiveness (2011)
- Burkert, Fischer, and Schäffer – Application of the controllability principle and managerial performance: The role of perceptions (2011)
- Windolph and Moeller – Open-book accounting: Reasons for failure of inter-firm cooperation (2012)
- Hartmann and Slapničar – The perceived fairness of performance evaluation: The role of uncertainty (2012)
- Speckbacher and Wentges – The impact of family control on the use of performance measures in strategic target setting and incentive compensation: A research note (2012)
- Caglio and Ditillo – Opening the black box of management accounting information exchanges in buyer-supplier relationships (2012)
- Bisbe and Malagueño – Using strategic performance measurement systems for strategy formulation: Does it work in dynamic environments?
- Burkert and Lueg – Differences in the sophistication of Value-based Management – The role of top executives (2013)
- Dekker, Sakaguchi, and Kawai – Beyond the contract: Managing risk in supply chain relations (2013)
- Ding, Dekker and Groot – Risk, partner selection and contractual control in interfirm relationships (2013)
- Pondeville, Swaen, and De Rongé – Environmental management control systems: The role of contextual and strategic factors (2013)
- Marginson, McAulay, Roush, and Zijl – Examining a positive psychological role for performance measures (2014)
- Ylinen and Gullkvist – The effects of organic and mechanistic control in exploratory and exploitative innovations (2014)

- Speklé and Verbeeten – The use of performance measurement systems in the public sector: Effects on performance (2014)
- Ming Chong and Mahama – The impact of interactive and diagnostic uses of budgets on team effectiveness (2014)
- Janke, Machlendorf and Weber – An exploratory study of the reciprocal relationship between interactive use of management control systems and perception of negative external crisis effects (2014)
- Su, Baird and Schoch – The moderating effect of organisational life cycle stages on the association between the interactive and diagnostic approaches to using controls with organisational performance (2015)
- Bedford – Management control systems across different modes of innovation: Implications for firm performance (2015)
- De Baerdemaeker and Bruggeman – The impact of participation in strategic planning on managers' creation of budgetary slack: The mediating role of autonomous motivation and affective organisational commitment (2015)
- Lisi – Translating environmental motivations into performance: The role of environmental performance measurement systems (2015)

Accounting, Organization and Society (2010 – 2015) – Survey research

- Henri and Journeault – Eco-control: The influence of management control systems on environmental and economic performance (2010)
- Veen-Dirks – Different uses of performance measures: The evaluation versus reward of production managers (2010)
- Grafton, Lillis, and Widener – The role of performance measurement and evaluation in building organizational capabilities and performance (2010)
- Bol and Moers – The dynamics of incentive contracting: The role of learning in the diffusion process (2010)
- Herda and Lavelle – The effects of organizational fairness and commitment on the extent of benefits big four alumni provide their former firm (2011)
- O'Connor, Vera-Muñoz, and Chan – Competitive forces and the importance of management control systems in emerging-economy firms: The moderating effect of international market orientation (2011)
- Fayard, Lee, Leitch, and Kettinger – Effect of internal cost management, information systems integration, and absorptive capacity on inter-organizational cost management in supply chains (2012)
- Artz, Homburg, Rajab – Performance-measurement system design and functional strategic decision influence: The role of performance-measures properties (2012)
- Guerreiro, Rodrigues, and Craig – Voluntary adoption of International Financial Reporting Standards by large unlisted companies in Portugal – Institutional logics and strategic responses (2012)
- Fullerton, Kennedy, Widener – Management accounting and control practices in a lean manufacturing environment (2013)
- Ho, Wu, and Wu – Performance measures, consensus on strategy implementation, and performance: Evidence from the operational-level of organizations (2014)
- Mahlendorf, Kleinschmit, and Perego – Relational effects of relative performance information: The role of professional identity (2014)
- Arnold and Artz – Target difficulty, target flexibility, and firm performance: Evidence from business units' targets (2015)
- King and Clarkson – Management control system design, ownership, and performance in professional service organisations (2015)
- Management Accounting Research (2010 – 2015) – Experimental research
- Kelvin Liu and Leitch – Performance effects of setting targets and pay-performance relations before or after operations (2013)
- Denker, Schwartz, Ward, and Young – Voluntary disclosure in a bargaining setting: A research note (2014)
- Cheng and Coyte – The effects of incentive subjectivity and strategy communication on knowledge-sharing and extra-role behaviours (2014)
- Accounting, Organization and Society (2010 – 2015) – Experimental research
- Koch and Schmidt – Disclosing conflicts of interest – Do experience and reputation matter? (2010)
- Schultz Jr., Bierstaker, and O'Donnell – Integrating business risk into auditor judgment about the risk of material misstatement: The influence of a strategic-system-audit approach (2010)
- Knechel, Salterio, Kochetova-Kozloski – The effect of benchmarked performance measures and strategic analysis on auditors' risk assessments and mental models (2010)

- Cianci and Kaplan – The effect of CEO reputation and explanations for poor performance on investors' judgement about the company's future performance and management (2010)
- O'Donnel and Prather-Kinsey – Nationality and differences in auditor risk assessment: A research note with experimental evidence (2010)
- Norman, Rose, and Rose – Internal audit reporting lines, fraud risk decomposition, and assessments of fraud risk (2010)
- Gibbins, McCracken, and Salterio – The auditor's strategy selection for negotiation with management: Flexibility of initial accounting position and nature of the relationship (2010)
- Cardinaels and Veen-Dirks – Financial versus non-financial information: The impact of information organization and presentation in a Balanced Scorecard (2010)
- Seifert, Sweeney, Joireman, and Thornton – The influence of organizational justice on accountant whistleblowing (2010)
- Jackson, Rodgers, Tuttle – The effect of depreciation method choice on asset selling prices (2010)
- Ranking and Sayre – Responses to risk in tournaments (2011)
- Norman, Rose, and Suh – The effects of disclosure type and audit committee expertise on Chief Audit Executives' tolerance for financial misstatements (2011)
- Gaynor, McDaniel, and Yohn – Fair value accounting for liabilities: The role of disclosures in unravelling the counterintuitive income statement effect from credit risk changes (2011)
- Tan and Koonce – Investors' reactions to retractions and corrections of management earnings forecasts (2011)
- Brüggen and Luft – Capital rationing, competition, and misrepresentation in budget forecasts (2011)
- Perreault and Kida – The relative effectiveness of persuasion tactics in auditor-client negotiations (2011)
- Chen, Kelly, Salterio – Do changes in audit actions and attitudes consistent with increased auditor scepticism deter aggressive earnings management? An experimental investigation (2012)
- Church, Hannan, and Kuang – Shared interest and honesty in budget reporting (2012)
- DeZoort, Holt, and Taylor – A test of the audit reliability framework using lenders' judgements (2012)
- Chang, Chen, and Trotman – The effect of outcome and process accountability on customer-supplier negotiations (2013)
- Chen and Tan – Judgement effects of familiarity with an analyst's name (2013)
- Rose, Mazza, Norman, and Rose – The influence of director stock ownership and board discussion transparency on financial reporting quality (2013)
- Messier Jr., Quick, and Vandervelde – The influence of process accountability and accounting standard type on auditor usage of a status quo heuristic (2014)
- Newman – An investigation of how the informal communication of firm preferences influences managerial honesty (2014)
- Brown, Fisher, Sooy, and Sprinkle – The effect of rankings on honesty in budget reporting (2014)
- Newman and Tafkov – Relative performance information in tournaments with different prize structures (2014)
- Fanning and Piercey – Internal auditors' use of interpersonal likability, arguments, and accounting information in a corporate governance setting (2014)
- Managing audits to manage earnings: The impact of diversions on an auditor's detection of earnings management (2015)
- Lachmann, Stefani, and Wöhrmann – Fair value accounting for liabilities: Presentation format of credit risk changes and individual information processing (2015)
- Gopalakrishnan, Libby, Samuels, and Swenson – The effect of cost goal specificity and new product development process on cost reduction performance (2015)
- Arnold and Gillenkirch – Using negotiated budgets for planning and performance evaluation: An experimental study (2015)
- Agoglia, Hatfield, and Lambert – Audit team reporting: An agency theory perspective (2015)
- Church, Peytcheva, Yu, and Singtokul – Perspective talking in auditor-manager interactions: An experimental investigation of auditor behaviour (2015)

Appendix B Extract of data for the empirical analysis

#	Journal	Authors	Volume (Year)	Study type	Replication	Information disclosed for replication?	Sample type	N	Sample size	Response rate	# hypothesis	Hypothesis	Result (0/1)	PValue	Disclose data availability	Mentioning statistical power	Examining the statistical power of the test	Publication bias	Engage in "sign econometrics"	Provide Confidence Intervals	Carefully compare and discuss findings with previous similar studies?
1	MAR	De Baerdema	29 (2015)	Cross-sectional	No	No	Random	2045	249	12%	5	1	0		No	No	No	0.8	Yes	No	No
												2	1	***							
												3	1	***							
												4	1	***							
												5	1	**							
2	MAR	Eleonora Lisi	29 (2015)	Cross-sectional	No	No	Non-random	443	91	21%	5	1a	1	***	No	Yes	No	1	Yes	No	No
												1b	1	**							
												1c	1	***							
												2a	1	***							
												2b	1	*							
3	MAR	Bedford	28 (2015)	Cross-sectional	No	Yes	Random	911	421	46%	10	1	1	**	No	No	No	0.5	Yes	No	No
												2	1	**							
												3	1	**							
												4	0								
												5	0								
												6	0								
												7	1	**							
												8	0								
												9	1	**							
												10	0								
4	MAR	Su, Bairdand	26 (2015)	Cross-sectional	No	Yes	Random	1000	343	34%	4	1	1	**	No	No	No	1	Yes	No	No
												2	1	**							
												3	1	**							
												4	1	*							
														**							
5	MAR	Janke, Mahler	25 (2014)	o-point Longitudi	No	(Yes)	Random	722	361	50%	2	1	1	***	No	No	No	1	Yes	No	No
								850	277	33%		2	1	**							
6	MAR	Ming Chong a	25 (2014)	Cross-sectional	Yes	(Yes)	Non-random	2000	186	9%	5	1a	1	***	No	No	No	0.6	Yes	No	No
												1b	0								
												2a	1	**							
												2b	0								
												3	1	****							
7	MAR	Spekté and Vé	25 (2014)	Cross-sectional	No	No	Non-random	97	97	100%	3	1	1	**	No	No	No	0.67	Yes	No	No
												2	1	*							
												3	0								
8	MAR	Margirson, M	25 (2014)	Cross-sectional	No	No	Non-random	284	98	35%	5	1	1	***	No	No	No	1	Yes	No	No
												2	1	**							
												3	1	***							
														***							
														***							
												4a	1	***							
												4b	1	**							
														***							
														***							

References

Aguinis, H., & Edwards, J. R. 2014. Methodological wishes for the next decade and how to make wishes come true. ***Journal of Management Studies***, 51(1), 143-174.

Anonymous. 2015. The Case of the Hypothesis That Never Was; Uncovering the Deceptive Use of Post Hoc Hypotheses. ***Journal of Management Inquiry***, 24(2), 214-216.

Artz, M., Homburg, C., & Rajab, T. 2012. Performance-measurement system design and functional strategic decision influence: The role of performance-measure properties. ***Accounting, organizations and society***, 37(7), 445-460.

Baker, M. 2016. 1,500 scientists lift the lid on reproducibility: Survey sheds light on the 'crisis' rocking research. ***Nature***, 533(7604), 452–454.

Ballas, A., & Theoharakis, V. 2003. Exploring diversity in accounting through faculty journal perceptions. ***Contemporary Accounting Research***, 20(4), 619-644.

Bamber, L. S., Christensen, T. E., & Gaver, K. M. 2000. Do we really 'know'what we think we know? A case study of seminal research and its subsequent overgeneralization. ***Accounting, organizations and society***, 25(2), 103-129.

Banks, G. C., O'Boyle, E. H., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., . . . Adkins, C. L. 2016. Questions about questionable research practices in the field of management a guest commentary. ***Journal of Management***, 42(1), 5-20.

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. 2016. Editorial: Evidence on questionable research practices: The good, the bad, and the ugly. ***Journal of Business and Psychology***, 31(3), 323-338.

Bedeian, A. G. 2004. Peer review and the social construction of knowledge in the management discipline. ***Academy of Management Learning & Education,*** 3(2), 198-216.

Bedeian, A. G., Taylor, S. G., & Miller, A. N. 2010. Management science on the credibility bubble: Cardinal sins and various misdemeanors. ***Academy of Management Learning & Education***, 9(4), 715-725.

Bedford, D. S. 2015. Management control systems across different modes of innovation: Implications for firm performance. ***Management Accounting Research***, 28, 12-30.

Bedford, D. S., & Malmi, T. 2015. Configurations of control: An exploratory analysis. ***Management Accounting Research***, 27, 2-26.

Begley, C. G., & Ellis, L. M. 2012. Drug development: Raise standards for preclinical cancer research. ***Nature***, 483(7391), 531-533.

Benjamini, Y., & Hechtlinger, Y. 2014) Discussion: An estimate of the science-wise false discovery rate and applications to top medical journals by Jager and Leek. ***Biostatistics***, 15(1), 13-16.

Bergh, D. D., Sharp, B. M., Aguinis, H., & Li, M. 2017. Is there a credibility crisis in strategic management research? Evidence on the reproducibility of study findings. ***Strategic Organization***, 15(3), 423-436.

Bisbe, J., & Malagueño, R. 2012. Using strategic performance measurement systems for strategy formulation: Does it work in dynamic environments? ***Management Accounting Research***, 23(4), 296-311.

Bonner, S. E., Hesford, J. W., Van der Stede, W. A., & Young, S. M. 2006. The most influential journals in academic accounting. ***Accounting, organizations and society***, 31(7), 663-685.

Borkowski, S. C., Welsh, M. J., & Zhang, Q. M. 2001. An analysis of statistical power in behavioral accounting research. ***Behavioral Research in Accounting***, 13(1), 63-84.

Butler, N., Delaney, H., & Spoelstra, S. 2017. The gray zone: Questionable research practices in the business school. ***Academy of Management Learning & Education***, 16(1), 94-109.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., . . . Chan, T. 2016. Evaluating replicability of laboratory experiments in economics. ***Science***, 351(6280), 1433-1436.

Campbell, J. P. 1982. Editorial: Some remarks from the outgoing editor. ***Journal of Applied Psychology***, 67(6), 691-700.

Carver, R. P. 1993. The case against statistical significance testing, revisited. ***The Journal of Experimental Education***, 61(4), 287-292.

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

Chen, W., & Tan, H.-T. 2013. Judgment effects of familiarity with an analyst's name. ***Accounting, organizations and society***, 38(3), 214-227.

Chua, W. F. 1986. Radical developments in accounting thought. ***Accounting review***, 61(4), 601-632.

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

Evans, J. H., Feng, M., Hoffman, V. B., Moser, D. V., & Van der Stede, W. A. 2015. Points to Consider When Self-Assessing Your Empirical Accounting Research. ***Contemporary Accounting Research***, 32(3), 1162-1192.

Fisher, R. A. 1956. ***Statistical methods and scientific inference.*** Edinburgh: Oliver and Boyd.

Francis, G. 2014. The frequency of excess success for articles in Psychological Science. ***Psychonomic Bulletin & Review***, 21(5), 1180-1187.

Garud, R. 2015. Eyes wide shut? A commentary on the hypothesis that never was. ***Journal of Management Inquiry***, 24(4), 450-454.
Gelman, A., & Loken, E. 2014a. Ethics and Statistics: The AAA Tranche of Subprime Science. ***CHANCE***, 27(1), 51-56.
Gelman, A., & Loken, E. 2014b. The statistical crisis in science. ***American Scientist***, 102(6), 460.
Gigerenzer, G. 2004. Mindless statistics. ***The Journal of Socio-Economics***, 33(5), 587-606.
Gigerenzer, G., & Marewski, J. N. 2015. Surrogate science the idol of a universal method for scientific inference. ***Journal of Management***, 41(2), 421-440.
Goodman, S. N., Fanelli, D., & Ioannidis, J. P. 2016. What does research reproducibility mean? ***Science translational medicine***, 8(341), 341ps312.
Greenwald, A. G. 1975. Consequences of prejudice against the null hypothesis. ***Psychological bulletin***, 82(1), 1.
Guerreiro, M. S., Rodrigues, L. L., & Craig, R. 2012. Voluntary adoption of International Financial Reporting Standards by large unlisted companies in Portugal–Institutional logics and strategic responses. ***Accounting, organizations and society***, 37(7), 482-499.
Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. 2015. The fickle P value generates irreproducible results. ***Nature methods***, 12(3), 179-185.
Hardwicke, T. E., Jameel, L., Jones, M., Walczak, E. J., & Weinberg, L. M. 2014. Only Human: Scientists, Systems, and Suspect Statistics. ***Opticon***1826, 25(16), 1-12.
Henri, J.-F., & Journeault, M. 2010. Eco-control: The influence of management control systems on environmental and economic performance. ***Accounting, organizations and society***, 35(1), 63-80.
Ho, J. L., Wu, A., & Wu, S. Y. 2014. Performance measures, consensus on strategy implementation, and performance: Evidence from the operational-level of organizations. ***Accounting, organizations and society***, 39(1), 38-58.
Hubbard, R., & Lindsay, R. M. 2013. The significant difference paradigm promotes bad science. ***Journal of Business Research***, 66(9), 1393-1397.
Hume, D. 1975. ***Enquiries concerning human understanding and concerning the principles of morals*** (3. ed., repr. / with text rev. and notes by P.H. Nidditch ed.). Oxford: Clarendon.
Ioannidis, J. P. 2005. Why most published research findings are false. ***PLoS Med***, 2(8), e124.
Ioannidis, J. P. 2008. Why most discovered true associations are inflated. ***Epidemiology***, 19(5), 640-648.
Ioannidis, J. P. 2014. Discussion: Why "An estimate of the science-wise false discovery rate and application to the top medical literature" is false. ***Biostatistics***, 15(1), 28-36.
Ioannidis, J. P., & Doucouliagos, C. 2013. What's to know about the credibility of empirical economics? ***Journal of Economic Surveys,*** 27(5), 997-1004.
Ittner, C. D. 2014. Strengthening causal inferences in positivist field studies. ***Accounting, Organizations and Society***, 39(7), 545-549.
Jager, L. R., & Leek, J. T. 2013. An estimate of the science-wise false discovery rate and application to the top medical literature. ***Biostatistics***, 15(1), 1-12.
Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. 2011. Again, and again, and again…. ***Science***, 334(6060), 1225.
John, L. K., Loewenstein, G., & Prelec, D. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. ***Psychological science***, 23(5), 524-532.
Kane, E. J. 1984. Why journal editors should encourage the replication of applied econometric research. ***Quarterly Journal of Business and Economics***, 23(1), 3-8.
Kerr, N. L. 1998. HARKing: Hypothesizing after the results are known. ***Personality and Social Psychology Review***, 2(3), 196-217.
Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. 2009. Circular analysis in systems neuroscience: the dangers of double dipping. ***Nature neuroscience***, 12(5), 535-540.
Lachmann, M., Trapp, I., & Trapp, R. 2017. Diversity and validity in positivist management accounting research—A longitudinal perspective over four decades. ***Management Accounting Research***, 34, 42-58.
Leung, K. 2011. Presenting post hoc hypotheses as a priori: Ethical and theoretical issues. ***Management and Organization Review***, 7(3), 471-479.
Lindsay, R. M. 1994. Publication system biases associated with the statistical testing paradigm. ***Contemporary Accounting Research***, 11(1), 33.
Loken, E., & Gelman, A. 2017. Measurement error and the replication crisis. ***Science***, 355(6325), 584-585.
Luft, J., & Shields, M. D. 2014. Subjectivity in developing and validating causal explanations in positivist accounting research. ***Accounting, Organizations and Society***, 39(7), 550-558.
Lykken, D. T. 1968. Statistical significance in psychological research. ***Psychological bulletin***, 70(3p1), 151-159.
MacCallum, R. C., Browne, M. W., & Sugawara, H. M. 1996. Power analysis and determination of sample size for covariance structure modeling. ***Psychological methods***, 1(2), 130.

Mahoney, M. J. 1977. Publication prejudices: An experimental study of confirmatory bias in the peer review system. ***Cognitive therapy and research***, 1(2), 161-175.
Maniadis, Z., Tufano, F., & List, J. A. 2014. One swallow doesn't make a summer: New evidence on anchoring effects. ***The American Economic Review***, 104(1), 277-290.
Martinson, B. C., Anderson, M. S., & de Vries, R. 2005. Scientists behaving badly. ***Nature***, 435(7043), 737-738.
Matthes, J., Marquart, F., Naderer, B., Arendt, F., Schmuck, D., & Adam, K. 2015. Questionable research practices in experimental communication research: A systematic analysis from 1980 to 2013. ***Communication Methods and Measures,*** 9(4), 193-207.
McCloskey, D. N., & Ziliak, S. T. 1996. The standard error of regressions. ***Journal of Economic Literature***, 34(1), 97-114.
McNutt, M. 2014. Raising the bar. ***Science***, 345(6192), 9-9.
Meehl, P. E. 1978. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. ***Journal of consulting and clinical Psychology***, 46(4), 806.
Merchant, K. A. 2010. Paradigms in accounting research: A view from North America. ***Management Accounting Research***, 21(2), 116-120.
Merton, R. K. 1942. Note on science and democracy. ***Journal of Legal & Political Sociology***, 1, 115.
Merton, R. K. 1973. ***The sociology of science: Theoretical and empirical investigations***. Chicago and London: University of Chicago press.
Moonesinghe, R., Khoury, M. J., & Janssens, C. J. 2007. Most published research findings are false--but a little replication goes a long way. ***PLoS medicine***, 4(2).
Motulsky, H. J. 2015. Common misconceptions about data analysis and statistics. ***British journal of pharmacology***, 172(8), 2126-2132.
Munafò, M., Noble, S., Browne, W. J., Brunner, D., Button, K., Ferreira, J., . . . Lindquist, M. 2014. Scientific rigor and the art of motorcycle maintenance. ***Nature biotechnology***, 32(9), 871-873.
Nosek, B., Alter, G., Banks, G., Borsboom, D., Bowman, S., Breckler, S., . . . Christensen, G. 2015. Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility. ***Science***, 348(6242), 1422-1425.
Nosek, B., Spies, J. R., & Motyl, M. 2012. Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. Perspectives on Psychological Science, 7(6), 615-631.
Nuzzo, R. (2014). Statistical errors. ***Nature,*** 506(7487), 150-152.
O'Boyle, E. H., Banks, G. C., & Gonzalez-Mulé, E. 2017. The Chrysalis Effect: How Ugly Initial Results Metamorphosize Into Beautiful Articles. ***Journal of Management***, 43(2), 376-399.
Ohlson, J. A. 2015. Accounting research and common sense. ***Abacus,*** 51(4), 525-535.
Perreault, S., & Kida, T. 2011. The relative effectiveness of persuasion tactics in auditor–client negotiations. ***Accounting, Organizations and Society***, 36(8), 534-547.
Prinz, F., Schlange, T., & Asadullah, K. 2011. Believe it or not: how much can we rely on published data on potential drug targets? ***Nat Rev Drug Discov***, 10(9), 712-712.
Shields, M. D. 1997. Research in management accounting by North Americans in the 1990s. ***Journal of Management Accounting Research***, 9, 3.
Siegfried, T. 2010. Odds Are, It's Wrong: Science fails to face the shortcomings of statistics. ***ScienceNews***, 177(7). Retrieved from https://www.sciencenews.org/article/odds-are-its-wrong
Silberzahn, R., & Uhlmann, E. L. 2015. Crowdsourced research: Many hands make tight work. ***Nature***, 526(7572), 189-191.
Simmons, J. P., Nelson, L. D., & Simonsohn, U. 2011. False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. ***Psychological science,*** 22(11), 1359-1366.
Simonsohn, U., Nelson, L. D., & Simmons, J. P. 2014. P-curve: A key to the file-drawer. ***Journal of Experimental Psychology: General***, 143(2), 534.
Starbuck, W. H. 2016. 60th Anniversary Essay How Journals Could Improve Research Practices in Social Science. ***Administrative Science Quarterly***, 1-19.
Sterling, T. D. 1959. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. ***Journal of the American statistical association***, 54(285), 30-34.
Sullivan, G. M., & Feinn, R. 2012. Using effect size-or why the P value is not enough. ***Journal of graduate medical education***, 4(3), 279-282.
Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., . . . Frame, M. 2011. Data sharing by scientists: practices and perceptions. ***PloS one***, 6(6), e21101.
Tullock, G. 2001. A comment on Daniel Klein's" a plea to economists who favor liberty". ***Eastern Economic Journal***, 27(2), 203-207.

Van der Stede, W. A., Young, S. M., & Chen, C. X. 2005. Assessing the quality of evidence in empirical management accounting research: The case of survey studies. ***Accounting, Organizations and Society***, 30(7), 655-684.

Van Noorden, R. 2014. ***Science joins push to screen statistics in papers: New policy follows efforts by other journals to bolster standards of data analysis.*** Retrieved from http://www.nature.com/news/science-joins-push-to-screen-statistics-in-papers-1.15509

Wasserstein, R. L., & Lazar, N. A. 2016. The ASA's Statement on p-values: context process, and purpose. ***The American Statistician***, 70(2), 129-133.

Yong, E. 2012. Replication studies: Bad copy. ***Nature***, 485(7398), 298-300.

Young, N. S., Ioannidis, J. P., & Al-Ubaydli, O. 2008. Why current publication practices may distort science. ***PLoS Med***, 5(10), e201.

Ziliak, S. T. 2016. Statistical significance and scientific misconduct: Improving the style of published research papers. ***Review of Social Economy***, 74(1), 83-97.

Ziliak, S. T., & McCloskey, D. N. 2004a. Significance redux. ***The Journal of Socio-Economics***, 33(5), 665-675.

Ziliak, S. T., & McCloskey, D. N. 2004b. Size matters: the standard error of regressions in the American Economic Review. ***The Journal of Socio-Economics***, 33(5), 527-546.