NEW GUIDELINES FOR NULL HYPOTHESIS SIGNIFICANCE TESTING IN HYPOTHETICO-DEDUCTIVE IS RESEARCH

Willem Mertens, Woolworths Group Limited, Brisbane, QLD, Australia

Jan Recker (corresponding author), University of Cologne

Abstract

The objective of this Research Perspectives article is to promote policy change amongst journals, scholars and students with a vested interest in hypothetico-deductive information systems (IS) research. We are concerned about the design, analysis, reporting and reviewing of quantitative IS studies that draw on null hypothesis significance testing (NHST). We observe that debates about misinterpretations, abuse, and issues with NHST, while having persisted for about half a century, remain largely absent in IS. We find this an untenable position for a discipline with a proud quantitative tradition. We discuss traditional and emergent threats associated with the application of NHST and examine how they manifest in recent IS scholarship. To encourage the development of new standards for NHST in hypothetico-deductive IS research, we develop a balanced account of possible actions that are implementable short-term or long-term and that incentivize or penalize specific practices. To promote an immediate push for change, we also develop two sets of guidelines that IS scholars can adopt right away.

Keywords

Research methods, quantitative, statistics, null hypothesis significance testing, p-value, hypothetico-deductive research, open science.

*Paper accepted at the **Journal of the Association for Information Systems**, 23 September 2019.*

NEW GUIDELINES FOR NULL HYPOTHESIS SIGNIFICANCE TESTING IN HYPOTHETICO-DEDUCTIVE IS RESEARCH

"Statistical techniques for testing hypotheses – have more flaws than Facebook's privacy policies."
Siegfried (2014)

Introduction

Our paper extends a conversation across several of our top journals (e.g., Burton-Jones & Lee, 2017; Gregor & Klein, 2014; Grover & Lyytinen, 2015) that focuses on pushing a prominent information systems (IS) research tradition toward "a new state of play" (Grover & Lyytinen, 2015)–namely positivist, quantitative research on basis of the hypothetico-deductive model to science (Godfrey-Smith, 2003, p. 236). This conversation is bound to theory-based, quantitative empirical studies that seek to explain and predict IS phenomena (Gregor, 2006), which includes a large majority of IS research (Gregor, 2006; Grover & Lyytinen, 2015), such as survey and experimental research traditions. At the same time, it excludes several important traditions such as both interpretive and qualitative research, design science research, as well as some quantitative traditions such as purely data-driven predictive methods and analytical modeling.

Like our colleagues before us, we see the need to constantly assess and revisit all aspects of our scholarship to ensure that we as a community constantly perform and improve on our fundamental mission, understanding how information systems can be effectively developed and deployed in the human enterprise.

Moreover, like the contributions of our colleagues in this conversation, we have a specific focus: ours is the way the IS community[1] apply *null hypothesis significance testing* (NHST) within the

[1] I.e., the entire IS scholarly ecosystem of authors, reviewers, editors/publishers and educators/supervisors.

hypothetico-deductive tradition. NHST is a method of statistical inference by which an hypothesized factor is tested against a hypothesis of no effect or relationship on basis of empirical observations (Pernet, 2016). NHST is the dominant statistical approach in use in science today (Gigerenzer, 2004) and broadly permeates through society. For example, the concept p-value – a key component of the NHST lexicon – features in statistics or algebra courses in schools in many countries since the 1930s, and is part of SAT testing in the United States at least since the 1990s.

The proposal we make in this paper details changes to the way we apply NHST in hypothetico-deductive research in IS. We argue making this proposal is *important* because it affects research practices employed by large parts of the IS community. The issue, we argue, is not necessarily vested in NHST, but in ourselves.[2] We argue that *the way NHST is used* in the research practices employed in our ecosystem of authors, reviewers, editors/publishers and educators has become so deeply rooted and ritualized that it formed normed habits that are difficult to break. This is a potential threat to IS research on two counts: first, some practices in applying NHST (such as the use and interpretation of the p-value) have always been susceptible to misunderstanding and misuse (e.g., Cohen, 1994; Dixon, 2003; Fisher, 1955; Lang, Rothman, & Cann, 1998; Neyman & Pearson, 1928). Second, changes to the phenomena and research settings in which IS scholarship is situated (such as the advent of digital population data or the emergence of computational advances to data analysis, e.g., Berente, Seidel, & Safadi, 2019; Freelon, 2014; Lazer et al., 2009) have begun to challenge incumbent practices; some have led to the emergence of questionable research practices that skirt the line between ethical and unethical rather than being blatant misconduct (O'Boyle Jr., Banks, & Gonzalez-Mulé, 2017).

We also argue making our proposal is *timely*. Conversations around the correct application of NHST in the sciences date back to its origin in the proposals for significance testing by R.A.

[2] We will also discuss some of the problems inherent to NHST but our clear focus is on our own fallibilities and how they could be mitigated.

Fisher (1935b) on the one hand and for acceptance based on critical rejection regions by J. Neyman and E.S. Pearson (1928, 1933) on the other hand. Still, several recent developments have reinvigorated this debate, which has paradoxically remained both rampant and dormant for decades. For example,

1. the movement to quantify academic productivity and outcomes through journal rankings and citation analysis since the early 2000s as part of the now well established "*publish or perish*" mantra has demonstratedly led to the emergence of several questionable research practices such as HARKing or *p*-Hacking (Kerr, 1998; O'Boyle Jr. et al., 2017; Simonsohn, Nelson, & Simmons, 2014; Starbuck, 2016);
2. the *open science movement*, i.e., the idea that all scientific knowledge elements (including publications, data, physical samples, and software) should be openly shared as early as is practical in the discovery process (Nielsen, 2011), while dating back hundreds of years (David, 2004), has over the past ten years advanced rapidly on the basis of internet technologies providing a range of novel services including data sharing platforms, computationally-intensive data analytics, crowdsourcing for project funding, open access publishing, data and publication archiving, and others;
3. the unfolding and increasing availability of large-scale volumes of *digital trace data* (Freelon, 2014; Howison, Wiggins, & Crowston, 2011) through the increasingly ubiquitous digitalization of everyday life (Vodanovich, Sundaram, & Myers, 2010; Yoo, 2010) has led to a vast increase in opportunities to conduct studies with extremely large organic sample sizes, which draws into doubt statistical practices historically used to draw inferences from small-sample populations (Lin, Lucas Jr., & Shmueli, 2013; Starbuck, 2016; Xu, Zhang, & Zhou, 2019);
4. advances in *computational approaches to data analytics* and statistical software packages with respect to interfaces, computational power and usability have led to a vast increase in popularity and application (e.g., Hair, Sarstedt, Ringle, & Mena, 2012; Ringle, Sarstedt, &

Straub, 2012) and allow researchers to easily sift repeatedly through data in search of patterns (Bettis, 2012). This has led some to argue that the increase in application of such methods has not been matched with a similar attention to methodological details (e.g., Rönkkö & Evermann, 2013; Rönkkö, McIntosh, Antonakis, & Edwards, 2016); and

5. the "*replication crisis"* (Open Science Collaboration, 2015; The Economist, 2013; Yong, 2012) has led to renewed and heightened skepticism about commonly used statistical procedures, as well as confirmation, positivity, and publication bias, which traversed from psychology to virtually all disciplines in the social sciences. In the IS field, it has led to the establishment of a dedicated journal, the *AIS Transactions on Replication Research* (Dennis & Valacich, 2015; Saunders et al., 2017).[3]

Finally, we argue that making our proposal is *relevant* to the IS field. While some of the above developments (e.g., the publish or perish movement, the replication crisis) are not restricted to the IS field alone, several others, in particular the advent of digital trace data, the rise of computational approaches to data analytics, and the continued emergence of technologically-enabled open science initiatives, speak fundamentally to the core phenomena in our field.[4]

We develop our proposal as follows. We first review NHST and its role in the hypothetico-deductive model to science. We review historic and emergent threats that relate to how NHST is applied in this scientific model. We then analyze the 100 most impactful recent papers in our top journals to identify whether NHST is commonly applied in leading IS scholarship and whether indicators exist that the discussed threats also occur in our field. We then make suggestions for the IS field for moving forward with the application of NHST, with the view to stimulate reflection and change. We detail proposals for how we theorize for statistical testing, how we use

---

[3] Remarkably, contrary to several fields, the experiences at the *AIS Transactions on Replication Research* after three years of publishing replication research has been that a meaningful proportion of research replications have produced results that are essentially the same as the original study (Dennis, Brown, Wells, & Rai, 2018).

[4] This trend is evidenced, for example, in the emergent number of IS research articles on these topics in our own journals (e.g., Berente et al., 2019; Howison et al., 2011; Levy & Germonprez, 2017; Lukyanenko, Parsons, Wiersma, & Maddah, 2019).

statistics for analysis, how we report results, and how we publish. We also detail two concrete sets of guidelines that our field can adopt right away.

NHST and its role in the traditional hypothetico-deductive research cycle

The point of this paper is neither to describe the origin and development of the hypothetico-deductive research cycle and its use of NHST in detail, nor to focus on the perceived or actual weaknesses of NHST as a technique in isolation. There are several accounts of the origin and evolution of NHST as a heuristic method of inference (e.g., Pernet, 2016; Szucs & Ioannidis, 2017) as well as a multitude of reviews and analyses of various properties of the technique itself (e.g., Amrhein, Greenland, & McShane, 2019; Branch, 2014; Wasserstein & Lazar, 2016). For the point of the paper, we provide an idealized account of a typical research process so that we can illustrate where potentially problematic practices involving NHST have always existed or recently emerged. We do so because such practices can threaten the efficiency, validity and robustness of the hypothetico-deductive research cycle. Figure 1 shows a stylized version of the hypothetico-deductive research cycle.

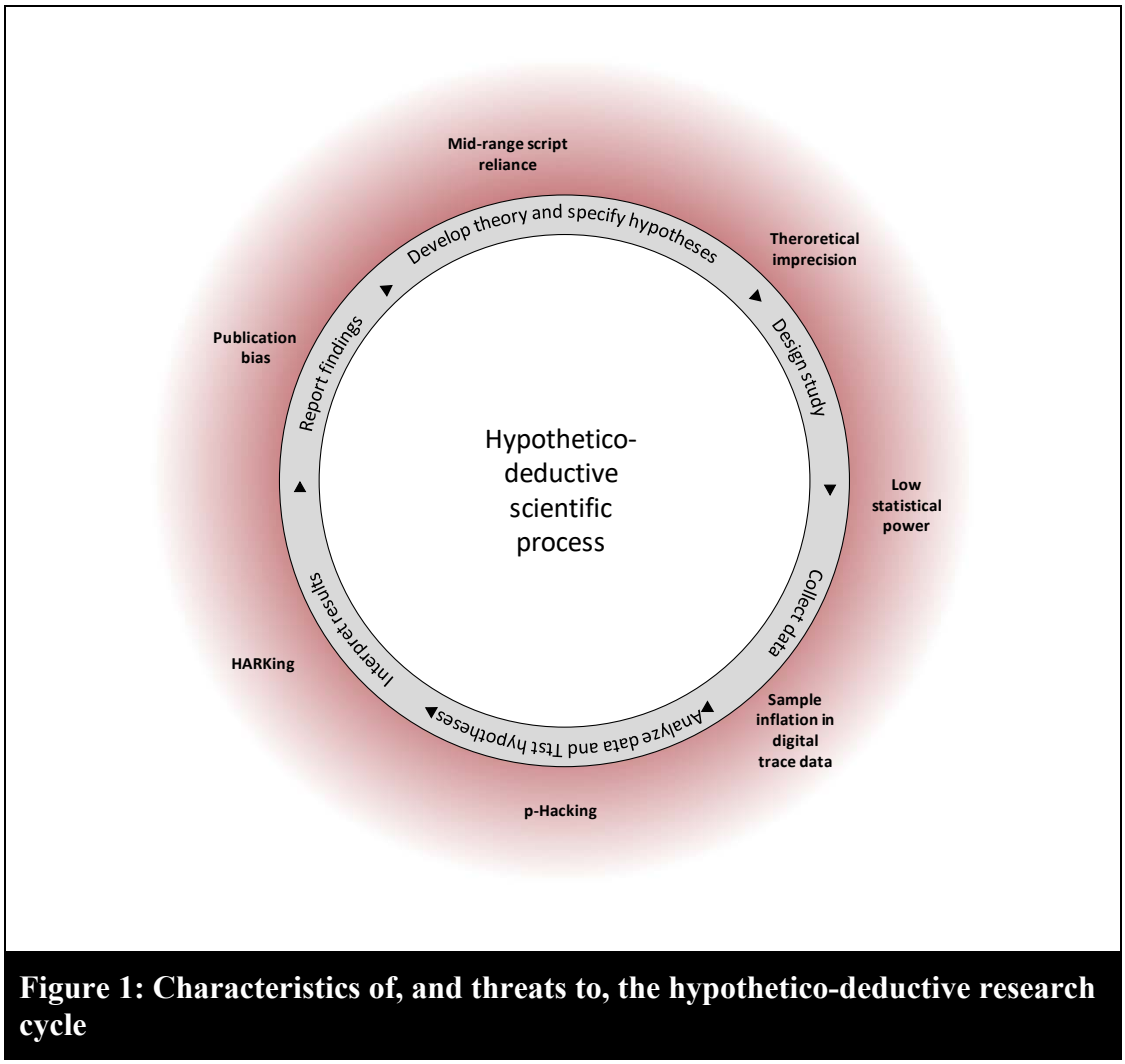

Figure 1: Characteristics of, and threats to, the hypothetico-deductive research cycle

Studies on basis of the hypothetico-deductive model to science typically proceed in six stages:

1) Researchers posit a new theory in the form of one or more hypotheses (e.g., people with small hands type faster).
2) They design an empirical study to obtain data (e.g., measures of typing speed and hand size).
3) They collect the data from a sample (e.g., a group of students).
4) They attempt to corroborate the hypotheses, by analyzing the gathered data and calculating some test statistic (e.g., a *t*-test comparing typing speed of those with large hands to those with small hands). They calculate a probability, the p-value, under the specified statistical model, that a particular test statistic (e.g., the average typing speed)

would be equal to or more extreme than its observed value, while assuming that some logical rival hypothesis is true in the population (e.g., people with small and large hands type *at the same speed*). This rival hypothesis is referred to as the null hypothesis, because it typically assumes the *absence* of an effect (e.g., no difference in typing speed). The p-value—the probability of finding the difference in typing speed that we found in our sample, or a larger difference, if we assume that there is no difference in the population—is then usually compared to certain thresholds (typically 0.05 or 0.01).

5) They interpret the results from the statistical tests. If the null hypothesis is rejected, researchers typically construe this result as denoting "accept" or "support" for the hypothesis stated at the beginning (e.g., people with small hands indeed type faster).
6) Finally, they submit a report detailing theory, study design and outcomes to a scientific peer-reviewed journal for publication.

The use of practices associated with NHST is deeply engrained in this scientific model. Not only is NHST the dominant approach to statistical data analysis as described above (Gigerenzer, 2004; Hubbard, 2004; Lin et al., 2013), NHST also forms the logical basis for most hypothesis development (Edwards & Berry, 2010; Lee & Hubona, 2009). Identifying samples that yield sufficient statistical power for NHST is a key component during study design (Baroudi & Orlikowski, 1989; Faul, Erdfelder, Lang, & Axel, 2007; Goodhue, Lewis, & Thompson, 2007), and data collection procedures involve several techniques for increasing statistical properties relevant for NHST, such as sample size (Sivo, Saunders, Chang, & Jiang, 2006). Finally, result interpretation and reporting also commonly follow recommendations that relate to NHST, either in the form of validation guidelines (Gefen, Rigdon, & Straub, 2011; Straub, 1989; Straub, Boudreau, & Gefen, 2004) or in the form of entire scripts, i.e., institutionalized patterns for knowledge creation and dissemination (Grover & Lyytinen, 2015; Tams & Straub, 2010).

The story goes that this way of using NHST within the hypothetico-deductive process was based on an intellectual debate, a misunderstanding of that debate, and a matter of convenience (Branch, 2014; Gigerenzer, 2004; Greenland et al., 2016; Lehmann, 1993). The debate mainly took place in the first half of the 20th century between Fisher (e.g., 1935a, 1935b; 1955) on the one hand, and Neyman and Pearson (e.g., 1928, 1933) on the other hand. Fisher introduced the idea of significance testing involving the probability *p* to quantify the chance of a certain event or state occurring, while Neyman and Pearson introduced the idea of accepting a hypothesis based on critical rejection regions. Fisher's idea is essentially an approach based on proof by contradiction (Christensen, 2005; Pernet, 2016): we pose a null model and test if our data conforms to it. This computation yields the probability of observing a result at least as extreme as a test statistic (e.g. a *t* value), assuming the null hypothesis of the null model (no effect) is true. This probability reflects the conditional, cumulative probability of achieving the observed outcome or larger: $p(Obs \geq t|H_0)$. Neyman and Pearson's idea was a framework of two hypotheses: the null hypothesis of no effect and the alternative hypothesis of an effect, together with controlling the probabilities of making errors. This idea introduced the notions of control of error rates, and of critical intervals. Together, these notions allow distinguishing type-I (rejecting H_0 when there is no effect) and type-II errors (not rejecting H_0 when there is an effect).

While both parties disagreed with each other's approach, a blend between both emerged as the now dominant approach to testing hypotheses (Lehmann, 1993). It is said that this occurred because scientists were in need of clear heuristics and were likely confused by the ongoing debate, and created a usable blend (Field, 2013; Reinhart, 2015). It is this "blend" of practices that emerged in the application of NHST, more so than properties of NHST itself, which is at the core of concerns in several disciplines, and should also be critically reflected upon in IS.

It is important here to declare that we do not mean to discredit the hypothetico-deductive model *per se*. In fact, like many of our colleagues, we have ourselves followed this model many times and benefitted from the advantages it provides:

- a strong foundation for building a cumulative knowledge tradition,
- means for both novel theory generation and incremental theoretical advance through intension and extension (Burton-Jones, Recker, Indulska, Green, & Weber, 2017; Kaplan, 1998/1964),
- means for comparison and reproduction of study results across different settings and samples,
- a shared language that is common to scientists across many fields, and
- cognitive advantages for both authors and readers in creating and assessing knowledge creation and the scripts we produce.

Yet, we believe that it is healthy to constantly revisit our scholarship procedures and ask whether normed habits and practices remain effective and efficient vehicles in light of new theory, empirics and ongoing changes to knowledge transfer mechanisms. Therefore, the analysis that follows focuses on what practices exist in using NHST in this model and the threats for knowledge creation efficiency, validity and robustness that flow from these practices.

Threats stemming from the application of NHST in the hypothetico-deductive research cycle

NHST has been controversial since its inception (e.g., Branch, 2014; Gigerenzer, 2004; Greenland et al., 2016) but recent developments have amplified some of the traditional concerns and saw new concerns emerge. We start by first reviewing *traditional threats* to research that stem from the application of NHST that have persisted over time, before then discussing *emergent threats* that have come to the forefront only or particularly in recent years. We discuss

both types of threats and the potential risks associated with them in some detail, noting that even broader accounts of these threats are available in the literature (Amrhein et al., 2019; Baker, 2016; Branch, 2014; Christensen, 2005; Dixon, 2003; Gelman & Stern, 2006; Gigerenzer, 2004; Greenland et al., 2016; McShane & Gal, 2017; Meehl, 1978; Munafò et al., 2017; Nickerson, 2000; Reinhart, 2015; Schwab, Abrahamson, Starbuck, & Fidler, 2011; Szucs & Ioannidis, 2017; Wasserstein & Lazar, 2016).[5]

Traditional Threat 1: NHST is difficult to understand and often misinterpreted. NHST builds on the p-value measure, which is arguably a sophisticated statistic because it provides an approach to summarizing the incompatibility between a particular set of data and a proposed model for the data. The most common context for applying NHST is a model describing hypotheses, constructed under a set of assumptions, together with the null hypothesis. However, applying NHST in this way typically involves construing double negatives and null hypotheses that are by design meant to be obviously false. Key terms such as "statistical significance" and "p-value" are demonstrably often misconstrued (Amrhein et al., 2019; Cohen, 1994; Greenland et al., 2016; Haller & Kraus, 2002; McShane & Gal, 2017; Reinhart, 2015). Several misinterpretations are particularly common: The p-value *is not an indication of the strength or magnitude of an effect* (Haller & Kraus, 2002). Any interpretation of the p-value in relation to the effect under study (strength, reliability, probability) is wrong, since p-values speak only about the null hypothesis. In addition, while p-values are randomly distributed (if all the assumptions of the test are met) when there is no effect, their distribution depends on both the population effect size and the number of participants, making it impossible to infer the strength of an effect from them. Similarly, 1-p *is not the probability of replicating an effect* (Cohen, 1994). Often, a small p-value is considered to indicate a strong likelihood of getting the same results on another try, but again this cannot be obtained because the p-value is not informative about the effect itself (Miller,

[5] To illustrate the magnitude of the conversation: In June 2019, *The American Statistician* published a special issue on null hypothesis significance testing that contains 43 articles on the topic (Wasserstein, Schirm, & Lazar, 2019).

2009). Because the p-value depends on the number of subjects, it can only be used in high-powered studies to interpret results. In low powered studies, the p-value has a large variance across repeated samples. A p-value also *is not an indication favoring a given hypothesis* (Szucs & Ioannidis, 2017). Because a low p-value only indicates a misfit of the null hypothesis to the data, it cannot be taken as evidence in favor of a specific alternative hypothesis more than any other possible alternatives such as measurement error and selection bias (Gelman, 2013). In fact, it is likely that the proportion of false positive findings in NHST-based studies is much greater than assumed (Nuzzo, 2014; Szucs & Ioannidis, 2017). The p-value also does *not describe the probability of the null hypothesis* $p(H_0)$ *being true* (Schwab et al., 2011). This common misconception arises from a confusion between the probability of an observation given the null $p(Obs \geq t|H_0)$ and the probability of the null given an observation $p(H_0|Obs \geq t)$ that is then taken as an indication for $p(H_0)$.

The only correct interpretation is that a p-value indicates the probability of obtaining the observed result or anything more extreme than that actually observed in the available sample data, assuming that (1) the null-hypothesis holds true in the population (by design largely an invalid assumption) and (2) all underlying model and test assumptions are met (e.g., random sampling, independence of sampled units, normality of distributions) (McShane & Gal, 2017).

The possible risk associated with incorrectly interpreting NHST is that researchers may either disregard evidence that fails to attain statistical significance or undervalue it relative to evidence that purportedly attains it, in turn leading to ill-informed judgments based on the evaluation of evidence (McShane & Gal, 2017). Interventions or treatments designed based on incorrectly interpreted evidence can lack effectiveness or even be harmful. Also, spurious findings may be published leading to diffusion of unsubstantiated theoretical claims.

Traditional Threat 2: NHST is sensitive to sampling strategy and sample size. The logic of NHST builds strongly on an appropriate sampling strategy. NHST logic demands random sampling because results from statistical analyses conducted on a sample are used to draw conclusions about the population. If samples are not drawn independently from measured variables and either selected randomly or selected to represent the population precisely, the conclusions drawn from NHST are not valid because it is impossible to correct for sampling bias, which statistical significance testing assumes is non-existent (Leahey, 2005). Yet, it is common practice to forego this requirement (Leahey, 2005; Starbuck, 2013).

With large enough sample sizes, a statistically significant rejection of a null hypothesis can be highly probable even if the underlying discrepancy in the examined statistics (e.g., the differences in means) is substantively trivial (Smith, Fahey, & Smucny, 2014). Sample size sensitivity occurs in NHST with so-called point-null hypotheses (Edwards & Berry, 2010), i.e., predictions expressed as point values. While such types of hypotheses are desired in the natural sciences (Szucs & Ioannidis, 2017, pp. 10-11), in social sciences, such as management, psychology or information systems, they lead to the paradox of stronger research designs yielding weaker tests because most hypotheses are specified as directional statements (such as a positive or negative relationship between two variables), where the point-null hypothesis describes the absence of a correlation, mean or variance difference (Schwab et al., 2011). A researcher that gathers a large enough sample can then reject basically any point-null hypothesis because the confidence interval around the null effect becomes smaller (Lin et al., 2013).

The possible risk is that with large sample sizes, applications of NHST lead to worse inferences (Meehl, 1967). Depending on the type of sampling strategy, especially in observational studies, it can be near impossible to control for the relationship of all irrelevant variables that are correlated with the variables of interest, which can lead to the identification of many correlations that can be

mistaken as revealing true relationships (Bruns & Ioannidis, 2016) and in computing biased and inconsistent estimations of effects.

Traditional Threat 3: NHST logic is incomplete. NHST rests on the formulation of a null hypothesis and its test against a particular set of data. This tactic relies on the so-called modus tollens (denying the consequence) (Cohen, 1994), a much used logic in both positivist and interpretive research in IS (Lee & Hubona, 2009). While modus tollens is logically correct, problems arise when it neglects pre-data probabilities: An example illustrates the error: *if a person is a researcher, it is very likely she does not publish in MISQ [null hypothesis]; this person published in MISQ [observation], so she is probably not a researcher [conclusion]*. This logic is, evidently, flawed.[6] In other words, the logic that allows for the falsification of a theory loses its validity when uncertainty and/or pre-data probabilities are included in the premises, yet both uncertainty (e.g., about true population parameters) and pre-data probabilities (pre-existent correlations between any set of variables) is at the very core null hypothesis significance testing as applied in the social sciences, especially when used in single research designs (such as one survey or experiment) (Falk & Greenbaum, 1995): in social reality, no two variables are ever perfectly unrelated (Meehl, 1967).

A second manifestation of incomplete logic is that NHST neglects predictions under H_1 (Szucs & Ioannidis, 2017). A widespread misconception is that rejecting H_0 allows accepting a specific H_1 (Nickerson, 2000). But NHST does not require a specification of the data that H_1 would predict, it only computes probabilities conditional on H_0. Rejection of H_0 thus offers no insight about how well the data might fit a general or specific H_1.

The possible risk associated with incomplete NHST logic beyond conceptual confusion and generation of misleading inferences is that it entices researchers to judge theories as better or

[6] An excellent analogous example using the relationship between mammograms and likelihood of breast cancer is provided by Gigerenzer, Gaissmeyer, Kurz-Milcke, Schwartz, and Woloshin (2008) in more detail.

worse even in the absence of direct comparisons to alternative theories. It also favors vaguely defined hypotheses because these are harder to definitely assess against credible alternatives. It makes it difficult and unlikely that theories are ever conclusively falsified (Edwards & Berry, 2010).

Traditional Threat 4: NHST fosters selective threshold-based reporting. p-value thresholds such as < 0.05 or even < 0.001 were never intended to be used as a basis for making 'pass or fail' decisions (Fisher, 1955). Neither Neyman and Pearson (1933) nor Fisher (1955) intended for the p-value to become a firm basis for accepting or rejecting hypotheses—let alone the only basis. Neyman and Pearson (1933, p. 291) wrote: "no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis". Rather, they proposed that p-values could help in reducing the chance of Type I and Type II errors:
"we may look at the purpose of tests from another view-point. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behavior with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong."

Neyman and Pearson (1928, p. 205) did passingly use a probability of five percent in one of their examples and as one of multiple arguments for why the tested hypothesis may best be rejected, and Fisher (1935a) did at some point argue that results with higher than 5% or even 1% probability should not be seen as 'unexpected' and therefore simply be ignored – the original intention was merely to use the term statistical significance to indicate that a particular result warrants further inspection. Although Fisher (1955) later changed his mind again, by that time scholars had already started using these fixed thresholds, thereby gradually solidifying the cut-off and reducing the importance of other arguments. Notably, in the social sciences, the vast majority of papers by now focus on statistically significant results (Szucs & Ioannidis, 2017), often not

fully or not entirely disclosing information about results that do not meet the commonly established thresholds.

The possible risk of threshold-based reporting is that the publication of "negative" or "insignificant" results is impeded, which leads to publication bias, the systematic suppression of research findings due to small magnitude, statistical insignificance, or contradiction of prior findings or theory (Harrison, Banks, Pollack, O'Boyle, & Short, 2014).

Emergent Threat 1: NHST is susceptible to questionable research practices. Shifts in academic culture, the availability of scholarly performance metrics and regulatory moves toward measuring research impact have created ample pressures on academics to publish "significant" contributions (Starbuck, 2016) to meet expectations for promotion and tenure (Dennis, Valacich, Fuller, & Schneider, 2006) and demonstrate research impact (Lyytinen, Baskerville, Iivari, & Te'Eni, 2007). One consequence of these pressures has been the emergence of a dominant type of research design where directional hypotheses are proposed alongside null hypotheses that claim there is *no effect*. This type of research design has been referred to as the "mid-range script", which is a legitimate, popular, reasonable and safe way of constructing knowledge with good prospects of publishability (Grover & Lyytinen, 2015, p. 279), but which also limits richer theorizing, constrains the freedom in relating theory and empirics, and weakens alternative forms of knowledge construction, such as data-driven research or blue ocean theorizing (Grover & Lyytinen, 2015, p. 285).

A second consequence of the publication pressure in academic culture is the growing prevalence of so-called questionable research practices (Bedeian, Taylor, & Miller, 2010; O'Boyle Jr. et al., 2017) that skirt the line between ethical and unethical behavior. The adoption of these practices is often understated but evidence amounts that they are prevalent in academia today (Bedeian et al., 2010; Kerr, 1998; O'Boyle Jr. et al., 2017; Starbuck, 2016).

The most prominent behaviors have become known under labels such as p-Hacking (manipulating, transforming, testing, and analyzing data until some statistically significant result emerges) and HARKing (Hypothesizing After Results are Known) although others also exist (O'Boyle Jr. et al., 2017). P-Hacking involves subjecting data to many calculations or manipulations in search for an equation that yields strong patterns. HARKing means presenting a *post hoc* hypothesis in a research report as if it were an *a priori* hypothesis (e.g., in the introduction) (Kerr, 1998). HARKing treads a fine line between theory-testing and theory-generating research because there are several variations to it depending on whether hypotheses were in fact anticipated and/or plausible (Kerr, 1998).

The possible risk is that p-Hacking can turn any false hypothesis into one that has statistically significant support, i.e., that false positive results are published, which could lead to scholars spending scarce resources chasing down false leads, and organizations and institutions implementing ineffective or even harmful policies (Nelson, Simmons, & Simonsohn, 2018). HARKing invalidates the idea of *a priori* hypothesis generation and subsequent testing and can lead to distorted publications limited to ideas and findings, without a faithful representation of the scientific process through which these ideas were born, which skews the image of science to students and the public audience. HARKing also risks increasing levels of type-I errors: if one attempts (too) many post-hoc analyses on the same data, some tests will generate false positives simply by chance (Szucs & Ioannidis, 2017). This runs the risk of misconstruing hypotheses that predicted false positives as theory to account for what is effectively an illusory effect. It also risks favoring weaker theories that post hoc accommodate results rather than correctly predict them, which in turns promotes developing narrow theory at the expense of broader, richer theorizing, and inhibits the generation of plausible alternative hypotheses.

Emergent Threat 2: NHST is unfit for many studies involving big data or digital trace data. The emergence of big data (Chen, Chiang, & Storey, 2012; George, Haas, & Pentland, 2014) and

the growing prevalence of digital trace data – evidence of activities and events that is logged and stored digitally (Freelon, 2014, p. 59) – increasingly allow researchers to obtain very large amounts of data, often to the point that the data collected resembles entire populations or at least very large fractions of populations.[7] Yet, NHST originally was conceived as a small-sample statistical inference technique (Meehl, 1967). In contexts involving digital trace population-level data, statistical inferences are increasingly meaningless because parameters of the data closely or fully resemble parameters of the studied populations (Starbuck, 2013). Likewise, in contexts involving big data, samples are dramatically statistically over-powered (Szucs & Ioannidis, 2017) leading to worse inferences (Lin et al., 2013).

The possible risks associated with NHST in studies involving big data is that it can lead researchers to claim support for statistically significant results that are not practically significant (Lin et al., 2013). The risk with digital trace data is that it is often generated organically, not following an explicit research design, which increases the likelihood of undermining the robustness of findings through potential errors in algorithmic outputs and in parametric and procedural choices for data processing. The opaqueness of the generation of digital trace data also threatens construct and internal validity (Xu et al., 2019).

How pervasive is NHST in hypothetico-deductive IS research?

We wanted to ascertain whether the discussed threats stemming from the application of NHST matter to the IS community, so we decided to collect data about their prevalence in our own field. Our reasoning was that if we can demonstrate that NHST is a commonly applied technique in IS research, it is important that our field engages in critical review and debate about the threats and possible risks associated with NHST.

[7] See Lin et al. (2013) for several examples.

We proceeded as follows. We reviewed 100 top cited papers in the senior scholar basket of eight IS journals between 2013 and 2016. Appendix A provides details about our procedures. We do not mean to claim that this is an exhaustive or representative sample of research papers in IS. Still, the papers' high citation counts suggest that other authors take inspiration for their own research from these papers. The reputation of the outlets and the citation count of the papers also suggest that they are considered to be of high quality by the community. As such, we believe these papers will allow us to develop some insights into the accepted research culture in IS, the 'way we do things around here'. When we point out suboptimal practices in these papers, we do not in any way wish to incriminate the excellent scholars who produced and reviewed these papers. We use these papers to talk about our whole community.

Of the 100 papers in our sample, 39 were quantitative research articles following the hypothetico-deductive model, a further two studies employed mixed method designs that involved quantitative empirical data collection and analysis in accordance to this model. Two additional design science papers involved quantitative data in the same vain. Our final sample was thus 43 papers. Of these, 15 employed surveys, followed by experiments, text mining and panel data studies (5 each). Six studies employed multiple types of data collections: two combined survey and experiment data, one combined experiment with interviews, and three combined surveys with either text mining, interviews or digital trace data.

Both the raw data and our coded data (*doi*:*10.25912/5cede0024b1e1)*, as well as the coding protocol (*doi*:*10.17605/OSF.IO/2GKCS)*, are available online for open inspection and assessment. Appendix A1 summarizes frequency counts for selected coding categories but we urge all readers to consult the data directly. Table 1 summarizes the main observations from the coding of the 43 papers, grouped by stage of the hypothetico-deductive scientific cycle, together with our interpretations of these observations in relation to the above-discussed threats. In what follows, we will discuss the conclusions we drew from our inspection of the data.

Table 1: Main findings from the coding of 43 published IS papers between 2013-2016 that follow the hypothetico-deductive model to science.

Stage of the hypothetico-deductive cycle	**Observations**	**Our Interpretations**
1. Develop hypotheses	38 of 43 papers state a-priori hypotheses. Two papers state hypotheses only in graphical form (as part of a research model). The largest share of hypotheses (13) are formulated as directional statements, followed by comparisons (6). Of 15 papers stating multiple forms of hypotheses, 10 involve directional statements.	NHST is a frequently applied technique in IS scholarship. Indicative of emergent threat #1: our theories often involve directional predictions around a null value indicating no effect. Strong research designs potentially yield weak tests of such theories.
2. Design study	39 of the 43 papers use a research design common to the mid-range script (Grover & Lyytinen, 2015). Four papers reportedly use an exploratory study design. Three papers report research designs set up as tests of competing theoretical models.	Indicative of emergent threat #1: the dominance of the mid-range script (Grover & Lyytinen, 2015) limits alternative modes of rich, inductive theorizing. Indicative of traditional threats #1 and #3: only few papers predict alternative or competing H_1.
3. Collect data	36 of 43 papers do not discuss statistical power during study design (2 papers report post-hoc power analyses). Four studies reportedly use power analysis for sampling. 22 of 43 papers use convenience sampling, six use systematic sampling, and four random sampling. Nine studies collect entire population-level data.	Indicative of traditional threat #2: it is common practice in IS scholarship to forego sampling and sample size requirements of NHST. Indicative of emergent threat #2: The studies involving big data or digital trace data in our sample draw on organically generated data (Xu et al., 2019) and do not adjust their statistical approach.
4. Analyze data	Across all 43 papers, 82% of hypotheses are reported as supported. The only study reporting less than 50% of supported hypotheses is the single replication study in the sample (none of eight hypotheses supported). Three papers consistently report exact p-values, eight papers do so selectively, 28 use threshold-based reporting.	 Indicative of traditional threat #4: threshold-based reporting occurs in IS scholarship.

	26 of 43 papers use R^2 measures for effect size reporting. Two use standardized means difference scores. Four papers report multiple effect size measures, 10 report none. 34 of 43 papers do not report confidence intervals in their results. Three do so consistently, two selectively. 11 of 43 papers use post-hoc analyses.	Indicative of traditional threat #1: estimations of strength or magnitude of discovered effects are neither always nor consistently reported in IS scholarship.
5. Interpret results	Three papers consistently refer to "statistical significance" when reporting on p-values. Several papers explicitly interpret significance as importance or magnitude of an effect (see point 6 below). 11 of 43 papers refer to p-values to point at the absence of an effect. Six of 43 papers use abductive reasoning in their interpretation of "unexpected results".	Indicative of traditional threat #1: Erroneous use and misinterpretation of NHST occur in IS scholarship.
6. Report findings	Almost all of the 43 papers contain language that declares hypotheses as supported/accepted or rejected on the basis of p-values exceeding a certain threshold. Consider the following examples (with modifications to mask identity): *"Our results reveal that the extent of [independent variables] are significant antecedents of [dependent variable] and that [dependent variables] are all significant […] reactions to [independent variable]".* *"Table 3 shows significant effects of [independent variable] on [dependent variable] at $p < .01$ for all […] cases, leading to strong support for H2a".* *"The significance level of each path coefficient indicates that each hypothesized path is significant. This means that [independent variables] have a significant impact on [dependent variable]. In addition, [independent variables] are significant influencing factors for [dependent variables]."* Of 20 papers in which some of the hypotheses are *not* supported by the data, seven papers refer to the statistically insignificant results as the basis for drawing explicit conclusions about the absence of an effect; four papers draw this conclusion	Indicative of traditional threat #4: threshold-based reporting occurs in IS scholarship. Indicative of traditional threat #1: misinterpretations of NHST occur in IS scholarship.

	implicitly. We also found cases where a proposed hypothesis is in fact a null hypothesis, and rejecting it is interpreted as support: *"Consistent with our expectations, none of the main effects of [independent variable] on [dependent variable 1] (β = valueβ_1, t = valuet_1) and [dependent variable 2] (β = valueβ_2, t = valuet_2) were significant".* *"The results indicate that the interaction terms of [independent variable] and [independent variable] are not significantly related to [dependent variable]. Therefore, we conclude that [independent variables] do not play a moderating role in the relationship between [independent variable] and [dependent variable]."*	

We draw two main conclusions from the observations and our interpretations in Table 1: First, we believe that the data shows that NHST is a well-established technique in hypothetico-deductive IS research. Second, we believe the data shows signs that the threats associated with NHST also have at least some level of occurrence in IS scholarship. Most of the hypothetico-deductive IS papers in our sample follow the common mid-range script (Grover & Lyytinen, 2015) and explicitly state *a priori hypotheses*, designed with *binary decisions* (accept vs. reject) and the *absence of no effect* in mind (Edwards & Berry, 2010). We also note that in much of the hypothesis development in IS papers, directional statements dominate and more precise non-point value or non-directional alternatives are scarce, which may indicate a lack of theoretical precision.

During study design and data collection, we note a lack of attention to statistical requirements of NHST such as random sampling – used in less than ten percent of articles in our sample. We also note that the large-sample studies within our sample used NHST for inference testing without making adjustments such as finite sample-size correction or avoiding inference statistics (such as p-values) altogether. Moreover, considerations related to statistical power, by and large, were not an explicit criterion in the study reports in our sample.

During data analysis and interpretation, we believe there is evidence to suggest that threshold-based reporting is prevalent in IS scholarship. We also found instances where the usage of statistical significance and p-values confuse statistical and practical significance (Lee, Mohajeri, & Hubona, 2017). For example, if we find that the unstandardized regression coefficient for the effect of perceived ease of use on perceived usefulness is .116, this means that someone scoring one point closer to the 'strongly agree' side of a 7-point Likert-type response scale for perceived ease of use, scores .116 points closer to 7 for perceived usefulness. This effect may be "statistically significant", but it is hardly practically meaningful.

We also find that analyses often rely largely on p-values alone. As can be seen in our coded data (https://doi.org/10.25912/5cede0024b1e1), only few reports in our sample make use of contextualized information, such as confidence intervals, effect sizes, post-hoc analyses, plots and graphs, and power. Admittedly, of course, examples to the contrary also exist in our sample (e.g., Burtch, Ghose, & Wattal, 2013; Dewan & Ramaprasad, 2014; Lu, Ba, Huang, & Feng, 2013; Mithas, Tafti, & Mitchell, 2013; Rishika, Kumar, Janakiraman, & Bezawada, 2013; Zeng & Wei, 2013). We found only six instances of papers where we interpreted the language as being indicative of abductive reasoning coupled with post-hoc analyses to make sense of purportedly "unexpected" results.

Finally, we believe that result reporting in our field shows similar signs of publication bias as in other fields (e.g., Harrison et al., 2014). The average support for the hypotheses in our sample (82%) seems disproportionately high (Edwards & Berry, 2010, p. 669), especially when considering that this figure includes the one replication study that found none of its eight hypotheses supported. This situation could be seen as an indicator that our review practices are biased toward "statistically significant" results (Emerson et al., 2010) although further research is needed to examine this speculation.

Proposing A Way Forward

Whilst our motivation was to scrutinize prevalence and potential threats in the use of NHST in IS research, we are not the first to examine issues in hypothetico-deductive IS research. Many broader issues discussed in our community relate to our proposal. For example, IS scholars have called for an increased emphasis on method and data triangulation, testing assumptions, using a balanced set of metrics including measures of model fit and effect sizes, and considering the magnitude of effects as well as their significance (Gerow, Grover, Roberts, & Thatcher, 2010; MacKenzie, Podsakoff, & Podsakoff, 2011; Venkatesh, Brown, & Bala, 2013). The "mid-range script" and its typical statistically testable model has been challenged as a mode of knowledge

construction; Grover and Lyytinen (2015) call for either more theoretically or practically oriented epistemic scripts. Moreover, there has already been a push for theory testing to go beyond 'effect' and 'prediction' testing, and for equal weight to be given to statistical significance and 'practical significance' (Lee et al., 2017). Others have highlighted the danger of Type I errors ('false positives') when sample sizes are large (Lin et al., 2013) and when reviewing papers (Straub, 2008), and discussed challenges relating to measurement (Bagozzi, 2011; Burton-Jones & Lee, 2017) and generalization (Lee & Baskerville, 2003; Tsang & Williams, 2012). However, none of these or any other papers in IS that we have seen thus far have explicitly examined the validity of the practices surrounding the application of NHST and its core elements such as proposing, accepting and rejecting hypotheses on the basis of p-values.

While this debate has not yet occurred in our own field, it is certainly alive in several other disciplines. The recent attention to the p-value debate in *Science* and *Nature* (Baker, 2016; McNutt, 2016; Nuzzo, 2014; Open Science Collaboration, 2015) show that it is a timely issue, and one that has the potential to endanger cumulative knowledge traditions (Johnson, Payne, Wang, Asher, & Mandal, 2017; Wasserstein et al., 2019).

We believe it is important for IS scholars to join the debate and help push toward new solutions. To identify an entry point into this debate, we collected and inspected proposals made in scientific disciplines that also follow the hypothetico-deductive model, such as psychology (Johnson et al., 2017; Trafimow & Marks, 2015; Tryon, Patelis, Chajewski, & Lewis, 2017), biology (Madden, Shah, & Esker, 2015; Nakagawa & Cuthill, 2007), epidemiology (Greenland et al., 2016), biomedicine (Twa, 2016), strategic and operations management (Bettis, Ethiraj, Gambardella, Helfat, & Mitchell, 2016; Guide Jr. & Ketokivi, 2015), organization science (Schwab et al., 2011), management science (O'Boyle Jr. et al., 2017; Xu et al., 2019), and statistical science (Wasserstein & Lazar, 2016).

When inspecting the proposals made in these fields, we noticed that there was both consensus and substantial variance in the proposals made. For example, in strategic and operations management, the *Strategic Management Journal* made two moves in 2016: first, it started welcoming replications and non-results as a primary type of contribution in top-level journals; and second, it no longer accepted papers for publication that refer to cut-off levels of statistical significance (Bettis et al., 2016). We suggest these two proposals differ in terms of *level of programming*: A **weakly programmed** proposal (e.g., welcoming replications a contribution befitting top level journals) is a move that incentivizes and encourages particularly desirable behaviors. A **strongly programmed** proposal (e.g., rejecting papers that use statistical significance thresholds in their argumentation) penalizes particularly undesirable behavior.

This distinction of weakly and strongly programmed proposals was also evident elsewhere. For example, in psychology, the journal *Basic and Applied Social Psychology* banned the use and reporting of p-values altogether (Trafimow & Marks, 2015) – a strongly programmed proposal. In organization science, Schwab et al. (2011) encourage researchers to include measures of uncertainty in their reporting, such as likelihood ratios, posterior probability distributions, or entire distributions of inferences – a weakly programmed proposal.

A second distinction we found useful is the *implementation timeframe* (short-term to long-term). Some proposals to better NHST-related practices can readily be implemented in the **short-term** by making adjustments, for example, to statistical reporting standards, or by implementing confirmatory signoffs during paper submissions to journals; whilst other proposals require **long-term** investments or cultural/institutional changes, such as the inclusion of alternative types of contributions welcomed by top level journals, the provision of independent methodological support or coaching in statistical methods, or the change in review modes to include results-blind reviewing (Locascio, 2019).

With these two distinctions, we developed an overview of the range of possible actions that the IS community could pursue in moving forward (Table 2). We explain each possible course of action by detailing the **change proposal** it entails, the likely **outcomes and implications** from its adoption with regard to the discussed NHST-associated threats, and the **primary stakeholder group** implicated (i.e., authors, reviewers, publishers and/or policy makers).

The proposals in Table 2 are practical and implementable. They are also backed with an increasing amount of evidence about their effectiveness (e.g., Munafò et al., 2017; Starbuck, 2016) that has been accrued through meta-research that examines scientific practices, develops and tests alternatives (e.g., Ioannidis, Fanelli, Drunne, & Goodman, 2015).

We do not suggest implementing all these proposals, let alone all at once. Our intent is merely to show that there is a range of options available to our field in moving forward, depending on our aptitude for penalties or incentives, and our willingness to move fast or slow. But we take the position that change we must. Formed habits are difficult to break at the best of times, and defective practices hard to stop. Change will also have to be implemented and accepted at all levels of our scholarly ecosystem. Changes in designing studies, analyzing data, writing or reviewing papers alone will not have the desired effect if they are not accompanied with an ecosystem-wide understanding of what qualifies as "good" research. This is why, by explicating choices for the different stakeholder groups, our proposal gives options to *authors* to engage in practices they find laudable to adopt (e.g., using pre-registrations). Our proposal also gives impetus for *journals* and *publishers* to strongly program certain behaviors. For example, proposals such as reporting number of statistical tests, confirming independent methodological oversight or declaring development of hypotheses truly a priori, could all be implemented in journal manuscript management systems (e.g., by configuring ScholarOne) during paper submission, much in the same vein we require authors to confirm ethical conduct. Finally, other proposals we include in Table 2 (such as developing standardized reporting checklists or running

a special issue on alternatives to statistical data analysis) are likely food for thought and discussion amongst groups including *authors* and *journals*, and will require individuals with an interest to take up the challenge to design such proposals.

Table 2: Change Proposals by Stage of the Hypothetico-Deductive Model to Science, differentiated by Level of Programming and Implementation Timeframe

Stage of the hypothetico-deductive research cycle	**Proposal**	**Implementation timeframe**	**Level of programming**	**Implicated outcome**	**Implicated stakeholders[8]**
1. Develop hypotheses	Encourage different epistemic script as alternatives to hypothetico-deductive research.	Short-term	Weakly programmed	Mitigates traditional threat #3 and emergent threat #2: it provides room for pluralistic and diverse modes of knowledge construction and theory generation (Grover & Lyytinen, 2015).	A, R, J
	Enforce pre-registration of hypotheses prior to data analysis.	Short-term	Strongly programmed	Mitigates traditional threat #4 and emergent threat #1: it minimizes risks from publication bias and HARKing (Warren, 2018).	A, R, J
2. Design study	Change the top journals' contribution model to embrace replications of prior hypotheses as desired contributions.	Long-term	Weakly programmed	Mitigates traditional threat #4 and emergent threat #1: it provides a stronger incentive for scholars to pursue reproducibility.	J, P
	Pre-register replication studies.	Long-term	Strongly programmed	Mitigates traditional threat #2 and #4: it maintains leeway for scientific creativity in original studies whilst enforcing strict rigor in replication studies (Gelman, 2015).	J, P
	Encourage sequential testing designs.	Short-term	Weakly programmed	Mitigates traditional threat #2 and #3: it promotes using multiple samples to test hypotheses (against alternatives where available) and implements a stage-gate model that stops when results do not continuously appear promising (Johnson et al., 2017).	A, J

[8] A = Authors, R = Reviewers/Editors, J = Journals/Publishers, P = Policy makers/Regulators

3. Collect data	Implement a results-blind review stage in journals prior to data collection.	Long-term	Strongly programmed	Mitigates traditional threat #2, #4 and emergent threat #1: it minimizes risks from publication bias and p-Hacking and allows focusing the review on theory development and study design (Greve, Bröder, & Erdfelder, 2013).	A, R, J
	Promote sharing of datasets in open repositories.	Short-term	Weakly programmed	Mitigates traditional threat #2, #4 and emergent threat #1: it fosters replication, independent inspection, and data re-use.	A, R, J
	Require authors to conduct multi-site data collections.	Short-term	Strongly programmed	Mitigates traditional threat #2, #3 and emergent threats #1 and #2: it allows distinguishing data-independent confirmatory research for testing hypotheses from data-contingent exploratory research for generating hypotheses.	J, P
4. Analyze data	Run special issues on alternative quantitative analyses for theory testing research in IS.	Long-term	Weakly programmed	Mitigates traditional threat #2, #3 and emergent threat #2: it fosters the development of novel inferential approaches that can be used in complementary or substitutive fashion with NHST, thereby adding value whilst eliminating the most egregious features (Matthews, 2019).	R, J
	Require authors to confirm independent methodological quality assurance.	Short-term	Strongly programmed	Mitigates traditional threat #1 and emergent threat #1: it protects against methodological shortcomings and encourages team science.	A, J
	Eliminate NHST as an approach to data analysis.	Short-term	Strongly programmed	Mitigates traditional threats #1, #2, #4 and emergent threat #2: it removes all vestiges of NHST, such as p-values, significance cut-offs, statements of "significance" and so forth, until new, widely accepted ways of data analysis have been developed (Trafimow & Marks, 2015).	R, J, P
5. Interpret results	Develop reporting checklists.	Long-term	Weakly programmed	Mitigates traditional threat #4 and emergent threat #1: Improves completeness and quality of reporting, ensures comparability across studies, and enables meta-analytic reviews (e.g., Shaw & Ertug, 2017).	A, R, J

	Eliminate language around "statistical significance" in papers.	Short-term	Strongly programmed	Mitigates traditional threats #1 and #4: it minimizes the risk for misinterpretation of NHST concepts and fosters more mindful interpretation of statistical results (Wasserstein et al., 2019).	A, R, J
6. Report findings	Reward transparent, open and reproducible reporting (e.g., through open research badges)[9].	Short-term	Weakly programmed	Mitigates traditional threats #2, #4 and emergent threat #1: it provides recognition to authors and makes open science practices desirable.	J, P
	Require authors to report the number of statistical tests conducted upon submission to journal.	Short-term	Strongly programmed	Mitigates traditional threat #4 and emergent threat #1: Makes scholars more mindful of their own practices and allow readers to better assess the veracity and power of reported results (Goldfarb & King, 2016).	J, P
	Build digital twins of entire research processes, decisions and outcomes.	Long-term	Weakly programmed	Mitigates traditional threat #4 and emergent threat #1: it provides a more accurate, timely and complete description of the research process than the ex post crafting of a paper.	A, P
	Encourage post-publication reviews.	Short-term	Weakly programmed	Mitigates traditional threats #3, #4 and emergent threat #1: it diversifies and extends peer review.	R, J

[9] See https://osf.io/tvyxz/wiki/home/ for more information on available types of open research badges.

Putting the Foot Down: Two readily implementable Proposals

On the individual stakeholder level: New guidelines for authors working on hypothetico-deductive IS research

We now describe measures that one core stakeholder group, namely researchers/authors, can adopt today. Table 3 details new guidelines for IS scholars, consisting of three sets of recommendations, two to encourage ("Should do" and "Could do") and one to discourage ("Must not do") certain NHST-relevant practices. The combination of "Should, Could and Must not" forms a balanced checklist that helps researchers throughout all stages of the research cycle to protect themselves against cognitive biases (e.g., by pre-registering protocols or hypotheses), improve statistical mastery where possible (e.g., through consulting independent methodological advice) and become modest, humble, contextualized and transparent (Wasserstein et al., 2019) wherever possible (e.g., by following open science reporting guidelines and cross-checking terminology and argumentation).

We make the distinction between "should do" and "could do" for two reasons. One, because some of the recommendations that scholars could opt to follow may not be applicable in all scenarios. For example, in research settings involving emergent technology or new, unexplored phenomena, directional hypotheses may be an appropriate way of developing new theory and sufficient information for alternative, more precisely formulated theories may not yet be available. Likewise, declaring a quantitative paper as theory-generating hinges on academic journals' aptitude to consider such work as a welcomed mode of contribution. Two, several of the "could do" recommendations draw on emergent science practices that have not yet widely been implemented or tested. For example, a conclusive verdict is at this point not yet available about how effective pre-registration is or how it can best be integrated into the constructive, developmental reviewing practices that many our discipline's journals adhere to (e.g., Saunders, 2005; Saunders et al., 2017).

Several of the “Must not” guidelines already exist in the form of educational materials or software solutions. For example, statistical power analysis can be performed using standalone tools (e.g., G*Power 3, Faul et al., 2007), and several tools exist to cross-check against p-Hacking and reporting bias (e.g., Schönbrodt, 2018). Also in the “Should do” and “Could do” categories are several options already available (e.g., for pre-registration and reporting), such as the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Statement (Moher, Liberati, Tetzlaff, Altman, & The PRISMA Group, 2009) relevant to correlational, observational studies or the Consolidated Standards of Reporting Trials (CONSORT) Statement (Begg et al., 1996) for experiments. It feels paradoxically both needless and important to point out that neither these nor other initiatives have yet found any substantial uptake in our community. For example, the Association for Information Systems has already launched a dedicated journal for replication (Saunders et al., 2017) and also is currently running a replication project (Dennis et al., 2018). Yet, the sheer existence of such outlets and initiatives that point out that particular scientific processes are “the right thing to do” is certainly necessary, but not sufficient.

Readers following the debate about NHST over the years may also feel that several variants of the guidelines in Table 3 have been suggested before. Yet, as our analysis showed, these practices have *not* widely diffused into our own community routines, which is why we decided to present a very clear, instructive checklist of what always to do, what possibly to do, and what most certainly not to do.

Table 3: New Guidelines for Hypothetico-Deductive IS Researchers.

Stage of the hypothetico-deductive cycle	**Should do**	**Could do**	**Must not do**
1. Develop hypotheses	• Specify hypotheses that test competing theories (Burton-Jones et al., 2017; Gray & Cooper, 2010) or compare data to naïve models rather than null hypotheses (Schwab et al., 2011, p. 1114).	• Specify hypotheses in non-point value or non-directional alternative forms (Edwards & Berry, 2010). • Pre-register hypotheses.	• Declare hypotheses as a priori if they were conceived post data collection and/or analysis.
2. Design study	• Explicate sampling strategy. • Pre-register protocols and instruments.	• Design studies for replications.	• Use inference study designs for population-level data collections. • Under- or over-power your data collection.
3. Collect data	• Always run and report *a priori* statistical power analyses.	• Share raw data in open data repositories.	• Collect data without written approval from ethics review boards.
4. Analyze data	• Always conduct effect size analyses. • Always report test statistics together with standard errors and confidence intervals.	• When using p-values, report them as continuous, descriptive quantities.	• Engage in creative data analysis and p-value polishing. • Dichotomize results as statistically significant or not depending on whether the p-value is below or above the size α of the hypothesis test. • Use statistical significance to measure the size of an effect.
5. Interpret results	• Eliminate language around "statistical significance" (Gelman & Stern, 2006). • Report effect sizes. • Translate effect sizes back to real-world phenomena/measures to demonstrate practical significance.	• Consult statisticians for independent methodological oversight and involve practitioners to cross-check practical relevance of results.	• Base your conclusions solely on whether an association or effect was found to be "statistically significant" without considering effect sizes. • Conclude anything about scientific or practical importance based on statistical significance or lack thereof (Lee et al., 2017).
6. Report findings	• Distinguish between a priori expectations and post-hoc inferences. • Use reporting checklists (McNutt, 2016).	• Declare your paper as theory-generating when hypothesizing after the results were known. • Follow open science reporting guidelines (Nosek et al., 2015).	• Hide, downplay or exclude unexpected or "non-significant" results, or non-results.

On the collective (institutional) level: Diversifying the peer review and publication process

A second set of practical and achievable steps we wish to explicate is for the entire IS research community to embrace the open science culture (Nosek et al., 2015), which recognizes *transparency*, *openness*, and *reproducibility* as vital values of scientific endeavor. Our stake is that IS research should not be a laggard in embracing open science ideas, it should be a leader and early adopter: the open science movement has since 2014 strongly embraced the possibilities offered by digital, networked platforms and readily available online infrastructure to implement ideas that go back hundreds of years (David, 2004). Open science promotes openness across the entire hypothetico-deductive cycle through (a) design standards that increase transparency about the research process and reduce vague or incomplete reporting, (b) open standards for sharing research materials, (c) data sharing standards that incentivize authors to make data available in trusted repositories.

This movement has made it possible, and over recent years also technologically feasible, to decouple two functions that have long been confounded: dissemination and evaluation of research (Munafò et al., 2017). Dissemination and evaluation has traditionally been a joint function of academic journals, however, dissemination can now be controlled independently from evaluation, or they could be loosely and temporally coupled at various stages and in various formats. For example, preprint services allow for dissemination of information to the research community at any stage of the research process. Online journals allow temporally decoupling peer reviews from the dissemination lifecycle (e.g., by substituting or complementing pre-publication peer review with post-publication peer reviews).

This move is not intended as a way to publish any research at any time. Instead, it opens possibilities for the harvesting of feedback from peers during the construction of a study or its

reporting. It also frees journals to trial alternative review models, such as results-free review (Button, Bal, Clark, & Shipley, 2016) or post-publication reviews (publons, 2017).

Table 4 demonstrates how the peer review and publication process across the stages of the hypothetico-deductive research cycle could be decoupled and expanded. It also lists new digital services that are already available to IS researchers but which, to the best of our knowledge, are not widely used. For purposes of illustration, therefore, we have taken steps wherever possible to use these services as they apply to this paper (see rightmost column). During that process, we immediately noticed several notable changes: first, being open by design undermined the double-blind peer review process (which is why we consulted with the senior editor prior to making these moves). Second, we were surprised about the sophisticated ways in which the open repositories, through standardized reporting protocols and interfaces, disseminated the various knowledge elements (data, protocols, paper versions) across different platforms (e.g., from open science registries to ResearchGate, ORCID and other platforms), and also how readily these moves found their way into academic conversations– we received platform, email and twitter inquires about this paper already during the review process just 24h after we posted a pre-print version on an open science server.[10] We are not ideologists and we are aware that there are both upsides and risks to being open during (not after) the reporting and peer review process, but we note that there is already evidence that open peer review improves the quality of reviews (Walsh, Rooney, Appleby, & Wilkinson, 2000) and that studies using pre-registered protocols markedly report more null findings (Warren, 2018). Both are laudable outcomes, in our view, that justify experimenting with these ideas.

[10] To illustrate, consider this tweet from June 3, 2019: *"Discussion on the #statisticalSignificance has reached ISR. "Null hypothesis significance testing in quantitative IS research: a call to reconsider our practices [submission to a second AIS Senior Scholar Basket of 8 Journal, received Major Revisions]" a new paper by @janrecker"* (https://twitter.com/AgloAnivel/status/1135466967354290176).

Table 4: A diversified model of the peer review and publication process, by stage of the scientific process, with examples.

Stage of the hypothetico-deductive scientific process	**Form of reporting**	**Suitable outlets**	**Type of review**	**Examples**	**Form of reporting perused for this paper**
Pre-data collection	Publication of pre-data collection theory and research design	IS conferences	Conference-level peer review	Main IS conferences such as ICIS ("short papers"), ECIS and PACIS ("research-in-progress papers), or AMCIS ("emergent research forum papers") accept pre-data collection research-in-progress reports as a type of submission.[11]	Not used.
Pre-data collection	Pre-registration of protocols	Open protocol repository	Moderation, no peer review	• Clinical trial protocols in medicine, e.g., (Lenzer, Hoffman, Furberg, & Ioannidis, 2013) • Open Science Foundation Registries https://osf.io/registries • Center for Open Science Preregistration https://cos.io/prereg/	The literature coding scheme is uploaded at http://dx.doi.org/10.17605/OSF.IO/2GKCS.
	Disclosure of ethics approval	Ethics approval database	Ethics board review	E.g., the Research Ethics Application Database https://tread.tghn.org/	Not applicable.
Pre-data analysis	Publication of raw data	Open data repositories	None	• http://datadryad.org/[12] (cross-disciplinary curated not-for-profit membership organization)	Both raw and coded data are uploaded at https://doi.org/10.25912/5cede00

[11] Note that the reviewing of these submission types at IS conferences is not strongly programmed, in the sense that these conferences accept both pre- and post-data collection papers (as well as other types of reports) to be submitted and reviewed.

[12] Note that presently DataDryad does not have IS journals as registered outlets.

				• Research Data Finder (institution-level data service provided by Queensland University of Technology)	24b1e1.
	Curation of research-in-progress papers	Web databases, galleries	Conference-level peer review	Could be automatically harvested from galleries such as https://icis2018postergallery.weebly.com/	Not applicable.
Pre-interpretation	Registration of "minimum replicable datasets".	Open data repositories	Through independent methodology-only reviewing	• http://datadryad.org/ (cross-disciplinary curated not-for-profit membership organization) • Research Data Finder (institution-level data service provided by Queensland University of Technology)	Both raw and coded data are uploaded at https://doi.org/10.25912/5cede0024b1e1.
Post-interpretation	Pre-review prints	Repositories for electronic pre-prints	Moderation but no peer review	• arXiv, https://arXiv.org/ • SocArXiv, https://osf.io/preprints/socarxiv • Social Science Research Network (SSRN) https://www.ssrn.com/[13]	The complete manuscript version history (eight versions) is uploaded on SocArXiv at https://doi.org/10.31235/osf.io/5qr7v.
	Pre-publication peer-review	Traditional academic journals	Editorial and peer review	Any mainstream IS journal.	
Post-review	Post-review pre-publication	Academic social networking sites	Peer-level adoption and comments	E.g., www.researchGate.net.	ResearchGate automatically imported the manuscript (versions) from SocArXiv. We updated the metadata throughout the process.
	Post-publication	Publication meta data	None	E.g., PubMed in medicine (a free resource developed and maintained by the National Center	Not applicable.

[13] Note that SSRN has been bought by the publisher Elsevier in July 2016.

	registration	repositories		for Biotechnology Information at the U.S. National Library of Medicine): https://www.ncbi.nlm.nih.gov/pubmed/	
	Post-publication review	Online academic journals	Peer review	• E.g., the *Australasian Journal in Information Systems* (e.g., Burmeister, 2016) • E.g., PubMed Commons in medicine[14] • E.g., Publons (publons, 2017)	Not (yet) available at the time of writing.

[14] Note that PubMed Commons has been discontinued because of the low level of participation, with comments submitted on only 6,000 of the 28 million articles indexed in PubMed (NCBI Insights, 2018).

Finally, by demonstrating in Table 4 how readily available this way of diversifying our dissemination and reviewing practices is to our community, we also wish to call out how the growing advent of the open science movement itself is entirely a digitally enabled and embodied phenomenon: open science processes and outcomes build on digital platforms, digital referencing, open interfaces, data exchange standards, and large-scale online databases. We ask, why are we not pushing the further development of these platforms and the practices they afford, why are we not studying these developments in much greater detail and volume, and why are we not yet broad adopters?

Conclusion

In this paper we developed new guidelines for the application of NHST in hypothetico-deductive IS research. We are not idealists. We know that breaking or changing routinized practices is difficult. We also know that like the other papers in the broader conversation of this proud IS research tradition, ours may raise more questions than it answers. For example, one of the most fundamental questions is whether the changes we propose will ultimately improve the robustness, validity and efficiency of our research. We tried to be forward looking and balanced in proposing several courses of actions, distinguished by level of programming and implementation timeframe. This way, it will be up to us as a community to decide whether we want to change directions by incentivizing ways we deem promising, or by implementing safeguards against ways of working we deem no longer acceptable. We do not believe either way is correct on its own. But if we can agree on experimenting on the right balance between encouragement and discouragement, we can allow our proud research tradition to continue to prosper.

We also tried to be assertive. Many of the issues we discussed have been discussed before but what is new is that there are now more pervasive elements of the threats, their implications, and generally a sense of what is not working, so the time is opportune to renew the call – and change the tone. We developed two sets of practical and achievable steps ready to be adopted

straightaway. But even if these suggestions will only lead to counter-proposals being made and perhaps implemented, then we see value in our proposal toward the ultimate aim of ensuring that IS research remains unbiased, rigorous, meaningful and relevant.

We are also of course ourselves "guilty as we charge". Personally, we have also in the past employed the same institutionalized practices like others in our community and we are very mindful that our own practices of NHST and the reporting in our own papers are just as susceptible to threats such as those we identified in our sample. At the same time, we are also adamant in our own stance to change the situation to the better. We are astutely aware of the mantra "walk our own talk". We have in the past organized seminars to educate students and researchers on the correct use of NHST. We have written a textbook on this topic (Mertens, Pugliese, & Recker, 2017). Where possible, we have already implemented several of the proposals we make in this paper, including sharing of datasets, pre-registering study protocols, disclosing the history of research and publication process changes and so forth, not only in this paper but also in others we were involved with in recent times.

Finally, we were not meaning to write an overly critical contribution. We do not mean to revive the IS anxiety debate (Grover, Straub, & Galluch, 2009). Where possible, we partake in the development of our own field. One matter that is dear to us in this context is that with our analysis of IS scholarship in this paper we do not mean to criticize our colleagues for how they constructed their articles. Science is a social endeavor and published articles are a poor representation of this complex process that involves negotiations between authors, reviewers and editors, which means that if there are potentially harmful habits that formed in this process, we as the entire ecosystem of IS scholars must work together to achieve change. We hope our proposal will help us engage in healthy periodical reviewing, constant self-reflection, critical self-assessment, and continuous improvement so that IS research can continue blending rigorous conduct, brilliant hypothesizing and the necessary quantity of good luck to continue to prosper.

Acknowledgments

We are indebted to the senior editor, Allen Lee, and two anonymous reviewers, for constructive and developmental feedback that helped us improve the paper. We thank participants at seminars at Queensland University of Technology and University of Cologne for providing feedback on our work. We are grateful to Christian Hovestadt for help in coding papers. All faults remain ours.

References

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature, 567*, 305-307.

Bagozzi, R. P. (2011). Measurement and meaning in information systems and organizational research: Methodological and philosophical foundations. *MIS Quarterly, 35*(2), 261-292.

Baker, M. (2016). Statisticians issue warning over misuse of p values. *Nature, 531*(7593), 151-151.

Baroudi, J. J., & Orlikowski, W. J. (1989). The problem of statistical power in mis research. *MIS Quarterly, 13*(1), 87-106.

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.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., . . . Stroup, D. F. S. (1996). Improving the quality of reporting of randomized controlled trials: The consort statement. *Journal of the American Medical Association, 276*(8), 637-639.

Berente, N., Seidel, S., & Safadi, H. (2019). Data-driven computationally-intensive theory development. *Information Systems Research, 30*(1), 50-64.

Bettis, R. A. (2012). The search for asterisks: Compromised statistical tests and flawed theories. *Strategic Management Journal, 33*(1), 108-113.

Bettis, R. A., Ethiraj, S., Gambardella, A., Helfat, C., & Mitchell, W. (2016). Creating repeatable cumulative knowledge in strategic management. *Strategic Management Journal, 37*(2), 257-261.

Branch, M. (2014). Malignant side effects of null-hypothesis significance testing. *Theory & Psychology, 24*(2), 256-277.

Bruns, S. B., & Ioannidis, J. P. A. (2016). P-curve and p-hacking in observational research. *PLOS ONE, 11*(2), e0149144.

Burmeister, O. K. (2016). A post publication review of "a review and comparative analysis of security risks and safety measures of mobile health apps". *Australasian Journal of Information Systems, 20*, 1-4.

Burtch, G., Ghose, A., & Wattal, S. (2013). An empirical examination of the antecedents and consequences of contribution patterns in crowd-funded markets. *Information Systems Research, 24*(3), 499-519.

Burton-Jones, A., & Lee, A. S. (2017). Thinking about measures and measurement in positivist research: A proposal for refocusing on fundamentals. *Information Systems Research, 28*(3), 451-467.
Burton-Jones, A., Recker, J., Indulska, M., Green, P., & Weber, R. (2017). Assessing representation theory with a framework for pursuing success and failure. *MIS Quarterly, 41*(4), 1307-1333.
Button, K. S., Bal, L., Clark, A., & Shipley, T. (2016). Preventing the ends from justifying the means: Withholding results to address publication bias in peer-review. *BMC Psychology, 4*, 59.
Chen, H., Chiang, R., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impacts. *MIS Quarterly, 36*(4), 1165-1188.
Christensen, R. (2005). Testing fisher, neyman, pearson, and bayes. *The American Statistician, 59*(2), 121-126.
Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997-1003.
Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oaks, California: Sage Publications.
David, P. A. (2004). Understanding the emergence of 'open science' institutions: Functionalist economics in historical context. *Industrial and Corporate Change, 13*(4), 571-589.
Dennis, A. R., Brown, S. A., Wells, T., & Rai, A. (2018). Information systems replication project Retrieved June 12, 2019, from https://aisel.aisnet.org/trr/aimsandscope.html
Dennis, A. R., & Valacich, J. S. (2015). A replication manifesto. *AIS Transactions on Replication Research, 1*(1), 1-4.
Dennis, A. R., Valacich, J. S., Fuller, M. A., & Schneider, C. (2006). Research standards for promotion and tenure in information systems. *MIS Quarterly, 30*(1), 1-12.
Dewan, S., & Ramaprasad, J. (2014). Social media, traditional media, and music sales. *MIS Quarterly, 38*(1), 101-121.
Dixon, P. (2003). *The p-value fallacy and how to avoid it.* 57 doi:10.1037/h0087425, Canadian Psychological Association, Canada.
Edwards, J. R., & Berry, J. W. (2010). The presence of something or the absence of nothing: Increasing theoretical precision in management research. *Organizational Research Methods, 13*(4), 668-689.
Emerson, G. B., Warme, W. J., Wolf, F. M., Heckman, J. D., Brand, R. A., & Leopold, S. S. (2010). Testing for the presence of positive-outcome bias in peer review: A randomized controlled trial. *Archives of Internal Medicine, 170*(21), 1934-1939.
Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology, 5*(1), 75-98.
Faul, F., Erdfelder, E., Lang, A.-G., & Axel, B. (2007). G*power 3: A flexible statistical power analysis for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191.
Field, A. (2013). *Discovering statistics using ibm spss statistics*: Sage.
Fisher, R. A. (1935a). *The design of experiments*. Edinburgh, Scotland: Oliver and Boyd.
Fisher, R. A. (1935b). The logic of inductive inference. *Journal of the Royal Statistical Society, 98*(1), 39-82.
Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society. Series B (Methodological), 17*(1), 69-78.
Freelon, D. (2014). On the interpretation of digital trace data in communication and social computing research. *Journal of Broadcasting & Electronic Media, 58*(1), 59-75.
Gefen, D., Rigdon, E. E., & Straub, D. W. (2011). An update and extension to sem guidelines for administrative and social science research. *MIS Quarterly, 35*(2), iii-xiv.
Gelman, A. (2013). P values and statistical practice. *Epidemiology, 24*(1), 69-72.
Gelman, A. (2015). Statistics and research integrity. *European Science Editing, 41*, 13-14.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician, 60*(4), 328-331.
George, G., Haas, M. R., & Pentland, A. (2014). From the editors: Big data and management. *Academy of Management Journal, 57*(2), 321-326.
Gerow, J. E., Grover, V., Roberts, N., & Thatcher, J. B. (2010). The diffusion of second-generation statistical techniques in information systems research from 1990-2008. *JITTA: Journal of Information Technology Theory and Application, 11*(4), 5.
Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics, 33*(5), 587-606.
Gigerenzer, G., Gaissmeyer, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2008). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest, 8*(2), 53-96.
Godfrey-Smith, P. (2003). *Theory and reality: An introduction to the philosophy of science*. Chicago, Illinois: University of Chicago Press.
Goldfarb, B., & King, A. A. (2016). Scientific apophenia in strategic management research: Significance tests & mistaken inference. *Strategic Management Journal, 37*(1), 167-176.
Goodhue, D. L., Lewis, W., & Thompson, R. L. (2007). Statistical power in analyzing interaction effects: Questioning the advantage of pls with product indicators. *Information Systems Research, 18*(2), 211-227.
Gray, P. H., & Cooper, W. H. (2010). Pursuing failure. *Organizational Research Methods, 13*(4), 620-643.
Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology, 31*(4), 337-350.
Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly, 30*(3), 611-642.
Gregor, S., & Klein, G. (2014). Eight obstacles to overcome in the theory testing genre. *Journal of the Association for Information Systems, 15*(11), i-xix.
Greve, W., Bröder, A., & Erdfelder, E. (2013). Result-blind peer reviews and editorial decisions: A missing pillar of scientific culture. *European Psychologist, 18*(4), 286-294.
Grover, V., & Lyytinen, K. (2015). New state of play in information systems research: The push to the edges. *MIS Quarterly, 39*(2), 271-296.
Grover, V., Straub, D. W., & Galluch, P. (2009). Editor's comments: Turning the corner: The influence of positive thinking on the information systems field. *MIS Quarterly, 33*(1), iii-viii.
Guide Jr., V. D. R., & Ketokivi, M. (2015). Notes from the editors: Redefining some methodological criteria for the journal. *Journal of Operations Management, 37*, v-viii.
Hair, J. F., Sarstedt, M., Ringle, C. M., & Mena, J. A. (2012). An assessment of the use of partial least squares structural equation modeling in marketing research. *Journal of the Academy of Marketing Science, 40*(3), 414-433.
Haller, H., & Kraus, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7*(1), 1-20.
Harrison, J. S., Banks, G. C., Pollack, J. M., O'Boyle, E. H., & Short, J. (2014). Publication bias in strategic management research. *Journal of Management, 43*(2), 400-425. doi: 10.1177/0149206314535438
Harzing, A.-W. (2010). *The publish or perish book: Your guide to effective and responsible citation analysis*. Melbourne, Australia: Tarma Software Research Pty Limited.
Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems, 12*(12), 767-797.
Hubbard, R. (2004). Alphabet soup. Blurring the distinctions between p's and a's in psychological research. *Theory & Psychology, 14*(3), 295-327.

Ioannidis, J. P. A., Fanelli, D., Drunne, D. D., & Goodman, S. N. (2015). Meta-research: Evaluation and improvement of research methods and practices. *PLoS Biology, 13*(10), e1002264.

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., & Mandal, S. (2017). On the reproducibility of psychological science. *Journal of the American Statistical Association, 112*(517), 1-10. doi: 10.1080/01621459.2016.1240079

Kaplan, A. (1998/1964). *The conduct of inquiry: Methodology for behavioral science*. Piscataway, New Jersey: Transaction Publishers.

Kerr, N. L. (1998). Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196-217.

Lang, J. M., Rothman, K. J., & Cann, C. I. (1998). That confounded p-value. *Epidemiology, 9*(1), 7-8.

Lazer, D., Pentland, A. P., Adamic, L. A., Aral, S., Barabási, A.-L., Brewer, D., . . . Van Alstyne, M. (2009). Computational social science. *Science, 323*(5915), 721-723.

Leahey, E. (2005). Alphas and asterisks: The development of statistical significance testing standards in sociology. *Social Forces, 84*(1), 1-24.

Lee, A. S., & Baskerville, R. (2003). Generalizing generalizability in information systems research. *Information Systems Research, 14*(3), 221-243.

Lee, A. S., & Hubona, G. S. (2009). A scientific basis for rigor in information systems research. *MIS Quarterly, 33*(2), 237-262.

Lee, A. S., Mohajeri, K., & Hubona, G. S. (2017). *Three roles for statistical significance and the validity frontier in theory testing*. Paper presented at the 50th Hawaii International Conference on System Sciences, Waikoloa Village, Hawaii.

Lehmann, E. L. (1993). The fisher, neyman-pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association, 88*(424), 1242-1249. doi: 10.2307/2291263

Lenzer, J., Hoffman, J. R., Furberg, C. D., & Ioannidis, J. P. A. (2013). Ensuring the integrity of clinical practice guidelines: A tool for protecting patients. *British Medical Journal, 347*, f5535.

Levy, M., & Germonprez, M. (2017). The potential for citizen science in information systems research. *Communications of the Association for Information Systems, 40*(2), 22-39.

Lin, M., Lucas Jr., H. C., & Shmueli, G. (2013). Too big to fail: Large samples and the p-value problem. *Information Systems Research, 24*(4), 906-917.

Locascio, J. J. (2019). The impact of results blind science publishing on statistical consultation and collaboration. *The American Statistician, 73*(sup1), 346-351.

Lu, X., Ba, S., Huang, L., & Feng, Y. (2013). Promotional marketing or word-of-mouth? Evidence from online restaurant reviews. *Information Systems Research, 24*(3), 596-612.

Lukyanenko, R., Parsons, J., Wiersma, Y. F., & Maddah, M. (2019). Expecting the unexpected: Effects of data collection design choices on the quality of crowdsourced user-generated content. *MIS Quarterly, 43*(2), 623-647.

Lyytinen, K., Baskerville, R., Iivari, J., & Te'Eni, D. (2007). Why the old world cannot publish? Overcoming challenges in publishing high-impact is research. *European Journal of Information Systems, 16*(4), 317-326.

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in mis and behavioral research: Integrating new and existing techniques. *MIS Quarterly, 35*(2), 293-334.

Madden, L. V., Shah, D. A., & Esker, P. D. (2015). Does the p value have a future in plant pathology? *Phytopathology, 105*(11), 1400-1407. doi: 10.1094/PHYTO-07-15-0165-LE

Matthews, R. A. J. (2019). Moving towards the post $p < 0.05$ era via the analysis of credibility. *The American Statistician, 73*(sup1), 202-212.

McNutt, M. (2016). Taking up top. *Science, 352*(6290), 1147.

McShane, B. B., & Gal, D. (2017). Blinding us to the obvious? The effect of statistical training on the evaluation of evidence. *Management Science, 62*(6), 1707-1718.
Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*(2), 103-115.
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*, 806-834.
Mertens, W., Pugliese, A., & Recker, J. (2017). *Quantitative data analysis: A companion for accounting and information systems research*. Cham, Switzerland: Springer.
Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review, 16*(4), 617-640.
Mithas, S., Tafti, A., & Mitchell, W. (2013). How a firm's competitive environment and digital strategic posture influence digital business strategy. *MIS Quarterly, 37*(2).
Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The prisma statement. *PLOS Medicine, 6*(7), e1000100.
Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., . . . Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*(0021), 1-9.
Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews, 82*(4), 591-605. doi: 10.1111/j.1469-185X.2007.00027.x
NCBI Insights. (2018). Pubmed commons to be discontinued Retrieved June 11, 2019, from https://ncbiinsights.ncbi.nlm.nih.gov/2018/02/01/pubmed-commons-to-be-discontinued/
Nelson, L. D., Simmons, J. P., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*, 511-534.
Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. *Biometrika, 20A*(1/2), 175-240. doi: 10.2307/2331945
Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231*, 289-337.
Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241-301.
Nielsen, M. (2011). *Reinventing discovery: The new era of networked science*. Princeton, New Jersey: Princeton University Press.
Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., . . . Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*(6242), 1422-1425.
Nosek, B. A., Ebersole, C. R., C., D. A., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences 115*(11), 2600-2606.
Nuzzo, R. (2014). Statistical errors: P values, the “gold standard” of statistical validity, are not as reliable as many scientists assume. *Nature, 506*(150), 150-152.
O'Boyle Jr., 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.
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), 943.
Pernet, C. (2016). Null hypothesis significance testing: A guide to commonly misunderstood concepts and recommendations for good practice [version 5; peer review: 2 approved, 2 not approved]. *F1000Research, 4*(621). doi: 10.12688/f1000research.6963.5

publons. (2017). 5 steps to writing a winning post-publication peer review Retrieved June 11, 2019, from https://publons.com/blog/5-steps-to-writing-a-winning-post-publication-peer-review/
Reinhart, A. (2015). *Statistics done wrong: The woefully complete guide*. San Francisco, California: No Starch Press.
Ringle, C. M., Sarstedt, M., & Straub, D. W. (2012). Editor's comments: A critical look at the use of pls-sem in mis quarterly. *MIS Quarterly, 36*(1), iii-xiv.
Rishika, R., Kumar, A., Janakiraman, R., & Bezawada, R. (2013). The effect of customers' social media participation on customer visit frequency and profitability: An empirical investigation. *Information systems research, 24*(1), 108-127.
Rönkkö, M., & Evermann, J. (2013). A critical examination of common beliefs about partial least squares path modeling. *Organizational Research Methods, 16*(3), 425-448.
Rönkkö, M., McIntosh, C. N., Antonakis, J., & Edwards, J. R. (2016). Partial least squares path modeling: Time for some serious second thoughts. *Journal of Operations Management, 47-48*, 9-27.
Saunders, C. (2005). Editor's comments: Looking for diamond cutters. *MIS Quarterly, 29*(1), iii-viii.
Saunders, C., Brown, S. A., Bygstad, B., Dennis, A. R., Ferran, C., Galletta, D. F., . . . Sarker, S. (2017). Goals, values, and expectations of the ais family of journals. *Journal of the Association for Information Systems, 18*(9), 633-647.
Schönbrodt, F. D. (2018). P-checker: One-for-all p-value analyzer Retrieved June 11, 2019, from http://shinyapps.org/apps/p-checker/
Schwab, A., Abrahamson, E., Starbuck, W. H., & Fidler, F. (2011). Perspective—researchers should make thoughtful assessments instead of null-hypothesis significance tests. *Organization Science, 22*(4), 1105-1120.
Shaw, J. D., & Ertug, G. (2017). From the editors: The suitability of simulations and meta-analyses for submissions to academy of management journal. *Academy of Management Journal, 60*(6), 2045-2049.
Siegfried, T. (2014). To make science better, watch out for statistical flaws. *ScienceNews Context Blog, 2019*.
Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534-547.
Sivo, S. A., Saunders, C., Chang, Q., & Jiang, J. J. (2006). How low should you go? Low response rates and the validity of inference in is questionnaire research. *Journal of the Association for Information Systems, 7*(6), 351-414.
Smith, S. M., Fahey, T., & Smucny, J. (2014). Antibiotics for acute bronchitis. *Journal of the American Medical Association, 312*(24), 2678-2679.
Starbuck, W. H. (2013). Why and where do academics publish? *M@n@gement, 16*(5), 707-718.
Starbuck, W. H. (2016). 60th anniversary essay: How journals could improve research practices in social science. *Administrative Science Quarterly, 61*(2), 165-183.
Straub, D. W. (1989). Validating instruments in mis research. *MIS Quarterly, 13*(2), 147-169.
Straub, D. W. (2008). Editor's comments: Type ii reviewing errors and the search for exciting papers. *MIS Quarterly, 32*(2), v-x.
Straub, D. W., Boudreau, M.-C., & Gefen, D. (2004). Validation guidelines for is positivist research. *Communications of the Association for Information Systems, 13*(24), 380-427.
Szucs, D., & Ioannidis, J. P. A. (2017). When null hypothesis significance testing is unsuitable for research: A reassessment. *Frontiers in Human Neuroscience, 11*(390), 1-21. doi: 10.3389/fnhum.2017.00390
Tams, S., & Straub, D. W. (2010). The effect of an is article's structure on its impact. *Communications of the Association for Information Systems, 27*(10), 149-172.

The Economist. (2013). Trouble at the lab. *The Economist* Retrieved April 7, 2017, from http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble
Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*(1), 1-2.
Tryon, W. W., Patelis, T., Chajewski, M., & Lewis, C. (2017). Theory construction and data analysis. *Theory & Psychology*, 0959354316684043. doi: 10.1177/0959354316684043
Tsang, E. W. K., & Williams, J. N. (2012). Generalization and induction: Misconceptions, clarifications, and a classification of induction. *MIS Quarterly, 36*(3), 729-748.
Twa, M. D. (2016). Transparency in biomedical research: An argument against tests of statistical significance. *Optometry & Vision Science, 93*(5), 457-458.
Venkatesh, V., Brown, S. A., & Bala, H. (2013). Bridging the qualitative-quantitative divide: Guidelines for conducting mixed methods research in information systems. *MIS Quarterly, 37*(1), 21-54.
Vodanovich, S., Sundaram, D., & Myers, M. D. (2010). Research commentary—digital natives and ubiquitous information systems. *Information Systems Research, 21*(4), 711-723.
Walsh, E., Rooney, M., Appleby, L., & Wilkinson, G. (2000). Open peer review: A randomised controlled trial. *The British Journal of Psychiatry, 176*(1), 47-51.
Warren, M. (2018). First analysis of 'pre-registered' studies shows sharp rise in null findings. *Nature, 24 October 2018*, d41586-41018-07118-41581.
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.
Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p < 0.05". *The American Statistician, 73*(sup1), 1-19.
Xu, H., Zhang, N., & Zhou, L. (2019). Validity concerns in research using organic data. *Journal of Management*, In Press.
Yong, E. (2012). Nobel laureate challenges psychologists to clean up their act. *Nature News* Retrieved April 7, 2017, from doi:10.1038/nature.2012.11535
Yoo, Y. (2010). Computing in everyday life: A call for research on experiential computing. *MIS Quarterly, 34*(2), 213-231.
Zeng, X., & Wei, L. (2013). Social ties and user content generation: Evidence from flickr. *Information Systems Research, 24*(1), 71-87.

Appendix A: Literature Review Procedures

Identification of Papers

In our intention to demonstrate "open science" practices (Locascio, 2019; Nosek, Ebersole, C., & Mellor, 2018; Warren, 2018) we preregistered our research procedures using the Open Science Framework "Registries" (*doi*:*10.17605/OSF.IO/2GKCS*).

We proceeded as follows: We identified the 100 top cited papers (per year) between 2013 and 2016 in the senior scholar basket of eight IS journals using Harzing's Publish or Perish version 6 (Harzing, 2010). We ran the queries separately on 7 February 2017 and then aggregated the results to identify the 100 most cited papers (based on citations per year) across the basket of eight journals.[15] The raw data (together with the coded data) is available at an open data repository hosted by Queensland University of Technology (*doi*:*10.25912/5cede0024b1e1)*.

We identified from this set of papers those that followed the hypothetico-deductive model. First, we excluded 48 papers that did not involve empirical data: 31 papers that offered purely theoretical contributions, 11 that were commentaries in the form of forewords, introductions to special issues or editorials, five methodological essays, and one design science paper. Second, we identified from these 52 papers those that reported on collection and analysis of *quantitative* data. We found 46 such papers; of these, 39 were traditional quantitative research articles, three were essays on methodological aspects of quantitative research, two studies employed mixed method designs involving quantitative empirical data and two design science papers that involved quantitative data. Third, we eliminated from this set the three methodological essays as the focus of these papers was not on developing and testing new theory to explain and predict IS

.[15] Our query terms were: [Management Information Systems Quarterly OR MIS Quarterly OR MISQ], [European Journal of Information Systems OR EJIS], [Information Systems Journal OR IS Journal OR ISJ], [Information Systems Research OR ISR], [Journal of the Association for Information Systems OR Journal of the AIS OR JAIS], [Journal of Information Technology OR Journal of IT OR JIT], [Journal of Management Information Systems OR Journal of MIS OR JMIS], [Journal of Strategic Information Systems OR Journal of SIS OR JSIS]. We checked for and excluded inaccurate results, such as papers from MISQ Executive, EJIS European Journal of Interdisciplinary Studies, etc.

phenomena. This resulted in a final sample of 43 papers, including two design science and two mixed method studies.

Coding of Papers

We developed a coding scheme in an excel repository to code the studies. The repository is available in our OSF registry. We used the following criteria. Where applicable, we refer to literature that defined the variables we used during coding.

- What is the main method of data collection and analysis (e.g., experiment, meta-analysis, panel, social network analysis, survey, text mining, economic modeling, multiple)?
- Are testable hypotheses or propositions proposed (yes/in graphical form only/no)?
- How precisely are the hypotheses formulated (using the classification of Edwards & Berry, 2010)?
- Is null-hypothesis significance testing used (yes/no)?
- Are exact p-values reported (yes/all/some/not at all)?
- Are effect sizes reported and, if so, which ones primarily (e.g., R^2, standardized means difference scores, f^2, partial eta^2)?
- Are results declared as "statistically significant" (yes/sometimes/not at all)?
- How many hypotheses are reported as supported (%)?
- Are p-values used to argue the absence of an effect (yes/no)?
- Are confidence intervals for test statistics reported (yes/selectively/no)?
- What sampling method is used (i.e., convenient/random/systematic sampling, entire population)?[16]

[16] We used the definitions by Creswell (2009, p. 148): *random sampling* means each unit in the population has an equal probability of being selected; *systematic sampling* means that specific characteristics are used to stratify the sample such that the true proportion of units in the studied population is reflected; and *convenience sampling* means that a nonprobability sample of available or accessible units is used.

- Is statistical power discussed and if so, where and how (e.g., sample size estimation, ex-post power analysis)?
- Are competing theories tested explicitly (Gray & Cooper, 2010)?
- Are corrections made to adjust for multiple hypothesis testing, where applicable (e.g., Bonferroni, alpha-inflation, variance inflation)?
- Are post-hoc analyses reported for unexpected results?

We also extracted quotes that in our interpretation illuminated the view taken on NHST in the paper. This was important for us to demonstrate the imbuement of practices in our research routines and the language used in using key NHST phrases such as “statistical significance” or “p-value” (Gelman & Stern, 2006).

To be as unbiased as possible, we hired a research assistant to perform the coding of papers. Before he commenced coding, we explained the coding scheme to him during several meetings. We then conducted a pilot test to evaluate the quality of his coding: the research assistant coded five random papers from the set of papers and we met to review the coding by comparing our different individual understandings of the papers. Where inconsistencies arose, we clarified the coding scheme with him until we were confident he understood it thoroughly. During the coding, the research assistant highlighted particular problematic or ambiguous coding elements and we met and resolved these ambiguities to arrive at a shared agreement. The coding process took three months to complete. The results of our coding are openly accessible at *doi*:*10.25912/5cede0024b1e1*. Appendix A1 provides some summary statistics about our sample.

Appendix A1. Selected descriptive statistics on 43 often cited IS papers from 2013-2016

Main method for data collection and analysis	Experiment	5
	Meta-analysis	2
	Panel	5
	Social network analysis	4
	Survey	15
	Text mining	5
	Economic modeling	1
	Multiple	6
Empirical data	Newly collected or analyzed primary data	40
	Re-analyzed or secondary data	3
Hypotheses	Testable hypotheses or propositions proposed	38
	No testable hypotheses or propositions proposed	5
	Average percentage of hypotheses per study that were supported by the data	82 %
Statement of hypotheses	As relations	0
	As upper/lower limits	0
	As directions	13
	In non-nil form	0
	In functional form	0
	In contingent form	2
	As comparisons	6
	In multiple ways	15
	Not formulated	2
	Not applicable	5
NHST	Uses NHST techniques or terminology	42
	Does not use NHST techniques or terminology	1
Exact p-values	Reports exact p-values	3
	Reports exact p-values selectively	8
	Reports indicators for different levels of statistical significance	28
	Does not report p-values	3
Inverse use of p-values	Uses p-values to point at absence of an effect or accept the null hypothesis	11
	Does not use p-values to point at absence of effect or accept null hypothesis	29
	Not applicable	3
‘Statistical’ significance	Does not explicitly refer to ‘statistical significance’	23
	Consistently refers to ‘statistical significance’	3
	Selectively refers to ‘statistical significance’	16
	Not applicable	1
Effect sizes	Reports R^2 measures	26

	Reports mean difference score measures	2
	Reports multiple effect size measures	4
	Does not report effect size measures	10
	Not applicable	1
Confidence intervals	Reports confidence intervals consistently	3
	Reports confidence intervals selectively	2
	Reports confidence intervals for bootstrapping results (no p-value available)	3
	Does not report confidence intervals	34
	Not applicable	1
Sampling	Convenient	22
	Systematic	6
	Random	4
	Entire population	8
	Not applicable	3
Competing Theories	Tested explicitly	7
	Not tested	35
	Not applicable	1
A posteriori analyses	Provided	11
	Not provided	31
	Not applicable	1