Abandon Statistical Significance

Blakeley B. McShane*, Northwestern University.
and
David Gal, University of Illinois at Chicago
and
Andrew Gelman, Columbia University
and
Christian Robert, Université Paris-Dauphine
and
Jennifer L. Tackett, Northwestern University

April 9, 2018

Abstract

We discuss problems the null hypothesis significance testing (NHST) paradigm poses for replication and more broadly in the biomedical and social sciences as well as how these problems remain unresolved by proposals involving modified thresholds, confidence intervals, and Bayes factors. We then discuss our own proposal, which is to abandon statistical significance. We recommend dropping the NHST paradigm—and the p-value thresholds intrinsic to it—as the default statistical paradigm for research, publication, and discovery in the biomedical and social sciences. Specifically, we propose that the p-value be demoted from its threshold screening role and instead, treated continuously, be considered along with currently neglected factors (e.g., prior and related evidence, plausibility of mechanism, study design and data quality, real world costs and benefits, novelty of finding, and other factors that vary by research domain) as just one among many pieces of evidence. We have no desire to "ban" p-values or other purely statistical measures. Rather, we believe that such measures should not be thresholded and that, thresholded or not, they should not take priority over the neglected factors. Instead, we offer recommendations for how our proposal can be implemented in the scientific publication process as well as in statistical decision making more broadly.

Keywords: null hypothesis significance testing; statistical significance; p-value; sociology of science; replication

---

*Correspondence concerning this article should be addressed to Blakeley B. McShane, Marketing Department, Kellogg School of Management, Northwestern University, 2211 Campus Drive, Evanston, IL 60208. E-mail: b-mcshane@kellogg.northwestern.edu. We thank the National Science Foundation, the Institute for Education Sciences, and the Office of Naval Research for partial support of this work.

1 The Status Quo and Two Alternatives

The biomedical and social sciences are facing a widespread crisis, with published findings failing to replicate at an alarming rate. Often, such failures to replicate are associated with claims of huge effects from subtle, sometimes even preposterous, interventions (or experimental manipulations). Further, the primary evidence adduced for these claims is one or more comparisons that are anointed "statistically significant"—typically defined as comparisons with p-values less than the conventional 0.05 threshold relative to a sharp point null hypothesis of zero effect and zero systematic error.

Indeed, the *status quo* is that $p < 0.05$ is deemed as strong evidence in favor of a scientific theory and is required not only for a result to be published but even for it to be taken seriously. Specifically, statistical significance serves as a lexicographic decision rule whereby any result is first required to have a p-value that attains the 0.05 threshold and only then is consideration—often scant—given to such factors as prior and related evidence, plausibility of mechanism, study design and data quality, real world costs and benefits, novelty of finding, and other factors that vary by research domain (we hereafter refer to these collectively as the *neglected factors* for want of a better term).

Traditionally, the $p < 0.05$ rule has been considered to be a safeguard against noise-chasing and thus a guarantor of replicability. However, in recent years, a series of well-publicized examples such as Carney *et al.* (2010) and Bem (2011), coupled with theoretical work has made it clear that statistical significance can easily be obtained from pure noise. Consequently, low replication rates are to be expected given existing scientific practices (Ioannidis, 2005; Smaldino and McElreath, 2016), and calls for reform, which are not new (see, for example, Meehl (1978)), have become insistent.

One alternative, suggested by Daniel Benjamin and seventy-one coauthors including distinguished scholars from a wide variety of fields, is to *redefine* statistical significance, "to change the default p-value threshold for statistical significance for claims of new discoveries from 0.05 to 0.005" (Benjamin *et al.*, 2018). While, as they note, "changing the p-value threshold is simple, aligns with the training undertaken by many researchers, and might quickly achieve broad acceptance," we believe this "quick fix," this "dam to contain the flood" in the words of a prominent member of the seventy-two (Resnick, 2017), is insufficient to overcome current difficulties with replication. Instead, we believe it opportune to proceed immediately with other measures, perhaps more radical and more difficult but likely also more principled and permanent.

In particular, we propose to *abandon* statistical significance, to drop the null hypothesis significance testing (NHST) paradigm—and the p-value thresholds intrinsic to it—as the default statistical paradigm for research, publication, and discovery in the biomedical and social sciences. Specifically, rather than allowing statistical significance as determined by $p < 0.05$ (or some other threshold whether based on p-values, confidence intervals, Bayes factors, or some other purely statistical measure) to serve as a lexicographic decision rule in scientific publication and statistical decision making more broadly, we propose that the p-value be demoted from its threshold screening role and instead, treated continuously, be considered along with the neglected factors as just one among many pieces of evidence.

We make this recommendation for three broad reasons. First, in the biomedical and social sciences, the sharp point null hypothesis of zero effect and zero systematic error used in the overwhelming majority of applications is generally not of interest because it is generally implausible. Second, the p-value thresholds intrinsic to NHST are not only problematic in and of themselves but they also routinely result in erroneous scientific reasoning even by experienced scientists and statisticians; for example, the standard use of NHST—to take the rejection of the straw man sharp point null hypothesis of zero effect and zero systematic error as positive or even definitive evidence in favor of some preferred alternative hypothesis—is a logical fallacy. Third, p-value and other statistical thresholds encourage researchers to analyze and report single comparisons rather than focusing on the totality of their data and relevant results.

To be clear, we have no desire to "ban" p-values or other purely statistical measures. Rather, we believe that such measures should not be thresholded and that, thresholded or not, they should not take priority over the neglected factors.

While our proposal to abandon statistical significance may seem on the surface quite radical, at least one aspect of it—to treat p-values or other purely statistical measures continuously rather than in a thresholded manner—most certainly is not. Indeed, this was advocated by R. A. Fisher himself (Fisher, 1956; Greenland and Poole, 2013) as well as by other early and eminent statisticians including Karl Pearson (Hurlbert and Lombardi, 2009), David Cox (Cox, 1977, 1982), and Erich Lehmann (Lehmann, 1993; Senn, 2001). It has also been advocated outside of statistics over the decades (see, for example, Boring (1919), Eysenck (1960), and Skipper *et al.* (1967)) and recently (see, for example, Drummond (2015), Lemoine *et al.* (2016), Amrhein *et al.* (2017), and Greenland (2017)).

This aspect of our proposal is also fully consistent with the recent American Statistical Association (ASA) Statement on Statistical Significance and p-values ("Principle 3: Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold;" Wasserstein and Lazar (2016)) as well as Valentin Amrhein and Sander Greenland's related proposal to *remove* statistical significance and treat p-values continuously (Amrhein and Greenland, 2018).

This aspect also stands in contrast to an alternative proposal which may perhaps at first pass sound similar but ultimately is quite distinct, namely the proposal of Daniel Lakens and eighty-three coauthors to *customize* statistical significance, to "justify [the] choice for an alpha level [i.e., statistical significance threshold] before collecting the data." (Lakens *et al.*, 2018). While this proposal is closer to ours than the status quo and the Benjamin *et al.* (2018) proposal in that it opposes a fixed and uniform statistical significance threshold—whether 0.05 or 0.005 or otherwise—it nonetheless rests upon NHST and the p-value thresholds intrinsic to it.

In sum, our proposal is part of a long literature both inside and outside of statistics over the decades that advocates treating p-values or other purely statistical measures continuously and thus stands in direct opposition to the threshold-based status quo and proposal of Benjamin *et al.* (2018) (and, for that matter, of Lakens *et al.* (2018)). Where we may perhaps differ from this literature is in two ways. First, we suggest that p-values or other purely statistical measures, thresholded or not, should not take priority over the neglected factors—noting of course that others have emphasized this including the recent ASA Statement which advises that "researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis" and cautions that "no single index should substitute for scientific reasoning" (Wasserstein and Lazar, 2016). Second, we offer recommendations for authors as well as editors and reviewers for how our proposal to abandon statistical significance can be implemented in practice.

Before elaborating on our recommendations, we discuss general problems with NHST that remain unresolved by the Benjamin *et al.* (2018) proposal as well as problems specific to the proposal. We then proceed to our recommendations for how, in practice, the p-value can be demoted from its threshold screening role and instead, treated continuously, be considered along with the neglected factors as just one among many pieces of evidence in the scientific publication process as well as in statistical decision making more broadly.

2 Problems General to Null Hypothesis Significance Testing

As noted, the NHST paradigm upon which the status quo and the Benjamin *et al.* (2018) proposal rest upon is the default statistical paradigm for research, publication, and discovery in the biomedical and social sciences (Morrison and Henkel, 1970; Gigerenzer, 1987; Sawyer and Peter, 1983; McCloskey and Ziliak, 1996; Gill, 1999; Anderson *et al.*, 2000; Gigerenzer, 2004; Hubbard, 2004). Despite this, it has been roundly criticized both inside and outside of statistics over the decades (Rozenboom, 1960; Bakan, 1966; Meehl, 1978; Serlin and Lapsley, 1993; Cohen, 1994; McCloskey and Ziliak, 1996; Schmidt, 1996; Hunter, 1997; Gill, 1999; Gigerenzer, 2004; Gigerenzer *et al.*, 2004; Briggs, 2016; McShane and Gal, 2016). Indeed, the breadth of literature on this topic across time and fields makes a complete review intractable. Consequently, we focus on what we view as among the most important criticisms of NHST for the biomedical and social sciences.

First, in the biomedical and social sciences, effects are typically small and vary considerably across people and contexts. In addition, measurements can be highly variable and are often only indirectly related to underlying constructs of interest; thus, even when sample sizes are large, the possibilities of systematic bias and variation results in the equivalent of small or unrepresentative samples. Consequently, estimates from any single study are themselves generally noisy. However, the single study is typically the fundamental unit of analysis.

These facts pose at least two problems for the NHST paradigm. The first results from the null hypothesis employed in the overwhelming majority of applications, namely the sharp point null hypothesis of zero effect—that is, no difference among two or more treatments—and zero systematic error—which encompasses both the adequacy of the statistical model used to compute the p-value (e.g., in terms of functional form and distributional assumptions) as well as any and all forms of systematic / non-sampling error which vary by field but include measurement error; problems with reliability and validity; biased samples; non-random treatment assignment; missingness; non-response; failure of double-blinding; non-compliance; and confounding. Simply put, this null hypothesis is itself implausible (Berkson, 1938; Edwards *et al.*, 1963; Bakan, 1966; Tukey, 1991; Cohen, 1994; Gelman *et al.*, 2014; McShane and Böckenholt, 2014; Gelman, 2015). Specifically, given that effects are generally small and variable, the assumption of zero effect is false. Further, given that measurements are generally noisy and systematically so, even were an effect truly zero, it would not be in any study designed to test it.

The second results from the lexicographic nature of the publication process under the status quo and the Benjamin *et al.* (2018) proposal. Specifically, noisy estimates that attain statistical significance and are thus eligible for publication are biased upwards (potentially to a large degree) and often of the wrong sign (Gelman and Carlin, 2014). Further, the lexicographic nature of the publication process at least indirectly encourages researchers to conduct many smaller, less resource-intensive, noisier studies as opposed to fewer larger, more resource-intensive, better studies as the former are more likely to yield—or can be made more likely to yield (Simmons *et al.*, 2011)—one or more statistically significant results. All of these issues are further compounded when researchers engage in multiple comparisons—whether actual or potential (the "garden of forking paths"; Gelman and Loken (2014)).

In sum, various features of contemporary biomedical and social sciences—for example, small and variable effects, systematic error, noisy measurements, the lexicographic nature of the publication process, and research practices—make NHST and in particular the sharp point null hypothesis of zero effect and zero systematic error particularly poorly suited for these domains.

Second, NHST is associated with a number of problems related to the dichotomization of evidence into the different categories "statistically significant" and "not statistically significant" (or, sometimes, trichotomization with "marginally significant" as an intermediate category) depending upon where the p-value stands relative to certain conventional thresholds. Indeed, one well-known criticism of the NHST paradigm is that the conventional 0.05 threshold—or for that matter any one—is entirely arbitrary (Fisher, 1926; Yule and Kendall, 1950; Cramer, 1955; Cochran, 1976; Cowles and Davis, 1982). A related line of criticism, discussed at length in the prior section, suggests that the problem is with having a threshold in the first place: the dichotomization (or trichotomization) of evidence into different categories of statistical significance itself has "no ontological basis" (Rosnow and Rosenthal, 1989).

We would go even further and say that it seldom makes sense to calibrate scientific evidence as a function of the p-value, given that this statistic is, in the overwhelming majority of applications, defined relative to the generally uninteresting and implausible null hypothesis of zero effect and zero systematic error. Recall, NHST examines the assumption that one or more model parameters equal the tested values—but only given all other model assumptions! These other assumptions (in particular zero systematic error) seldom hold or are at least far from given in the biomedical and social sciences. Consequently, "a small p-value only signals that there may be a problem with at least one assumption, without saying which one. Asymmetrically, a large p-value only means that this particular test did not detect a problem—perhaps because there is none, or because the test is insensitive to the problems, or because biases and random errors largely canceled each other out" (Greenland, 2017).

Third, dichotomization (or trichotomization) of evidence also results in erroneous scientific reasoning. For instance, researchers often confuse statistical significance and practical importance. Further, they often make scientific conclusions largely if not entirely based on whether or not a p-value crosses the 0.05 threshold instead of taking a more holistic view of the evidence that includes the consideration of the neglected factors. Finally, because the assignment of evidence to different categories is a strong inducement to the conclusion that the items thusly assigned are categorically different, they engage in dichotomous thinking; specifically, they interpret evidence that reaches the conventionally defined threshold for statistical significance as a demonstration of a difference and in contrast they interpret evidence that fails to reach this threshold as a demonstration of no difference.

An example of erroneous reasoning resulting from dichotomous thinking is provided by Gelman and Stern (2006) who show that applied researchers often fail to appreciate that "the difference between 'significant' and 'not significant' is not itself statistically significant." Additional examples are provided by McShane and Gal (2016) who show that researchers across a wide variety of fields including medicine, epidemiology, cognitive science, psychology, and economics (i) interpret p-values dichotomously rather than continuously, focusing solely on whether or not the p-value is below 0.05 rather than the magnitude of the p-value; (ii) fixate on p-values even when they are irrelevant, for example when asked about descriptive statistics; and (iii) ignore other evidence, for example the magnitude of treatment differences. McShane and Gal (2017) show that even statisticians are susceptible to these errors.

3 Problems Specific to the Benjamin *et al.* (2018) Proposal

Beyond concerns about the NHST paradigm upon which the status quo and the Benjamin *et al.* (2018) proposal rest, there are additional problems specific to the latter proposal. First, Benjamin *et al.* (2018) propose the 0.005 threshold because it (i) "corresponds to Bayes factors between approximately 14 and 26" in favor of the alternative hypothesis and (ii) "would reduce the false positive rate to levels we judge to be reasonable." However, little to no justification is provided for either of these choices of levels.

Second, Benjamin *et al.* (2018) "restrict [their] recommendation to claims of discovery of new effects" which is problematic for at least two reasons. First, because they fail to define what constitutes a new effect, their proposal is rendered entirely impractical; this is especially so in domains where research is believed to be incremental and cumulative. Second, the proposed policy also would lead to incoherence when applied to replication—

the very issue their proposal is meant to address. In particular, the order in which two independent studies of a common phenomenon are conducted is irrelevant in Bayesian updating. However, given two studies with $p < 0.005$ and $p \in (0.005, 0.05)$, it would matter crucially which study was conducted first (and was thus "new") under the definition of replication employed in practice (i.e., a subsequent study is considered to successfully replicate a prior study if either both fail to attain statistical significance or both attain statistical significance and are directionally consistent): the second (replication) study would be deemed a success under the Benjamin *et al.* (2018) proposal if the first study was the $p < 0.005$ study but a failure otherwise.

Third, the fact that uncorrected multiple comparisons—both actual and potential—are the norm in applied research strictly speaking invalidates all p-values outside those from studies with preregistered protocols and data analysis procedures (and even with preregistration p-values can be invalidated if the underlying model that generated the p-value is misspecified in an important manner). This concern is acknowledged by Benjamin *et al.* (2018).

Fourth, the mathematical justification underlying the Benjamin *et al.* (2018) proposal has come under no small amount of criticism. Specifically, the uniformly most powerful Bayesian tests (UMPBTs) that underlie the proposal were introduced and defended by Johnson (2013b) in parallel with his call in Johnson (2013a)—and now repeated in Benjamin *et al.* (2018)—to use 0.005 as the new threshold. We see a number of concerns with UMPBTs that we discuss in Appendix A; perhaps most relevant for the biomedical and social sciences, the UMPBT approach is deeply entrenched in the century-old Neyman-Pearson formalism of binary decisions and 0-1 loss functions which does not in general map, even in an approximate way, to processes of scientific learning or costs and benefits.

More speculatively, we are not convinced the more stringent 0.005 threshold for statistical significance would be helpful. In the short term, it could reduce the flow of low quality work that is currently polluting even top journals. In the medium term, it could motivate researchers to perform higher-quality work that is more likely to crack the 0.005 barrier. On the other hand, it could lead to even more overconfidence in results that do get published as well as a concomitant greater exaggeration of the effect sizes associated with such results. It could also lead to the discounting of important findings that happen not to reach the more stringent threshold. In sum, we have no idea whether implementation of the proposed 0.005 threshold would improve or degrade the state of science as we can envision both positive and negative outcomes resulting from it. Ultimately, while this question may be interesting if difficult to answer, we view it as outside our purview because we believe that thresholds whether based on p-values or other purely statistical measures are a bad idea in general.

Perhaps curiously, we do not necessarily expect that Benjamin *et al.* (2018) would disagree with our criticism that their proposal is insufficient to overcome current difficulties with replication (or perhaps even with our own proposal to abandon statistical significance). After all, they "restrict [their] recommendation to claims of discovery of new effects" and recognize that "the choice of any particular threshold is arbitrary" and "should depend on the prior odds that the null hypothesis is true, the number of hypotheses tested, the study design, the relative cost of Type I versus Type II errors, and other factors that vary by research topic." Indeed, "many of [the authors] agree that there are better approaches to statistical analyses than null hypothesis significance testing."

4 Abandoning Statistical Significance

4.1 Summation and Recommendations

What can be done? Statistics is hard, especially when effects are small and variable and measurements are noisy as in the biomedical and social sciences. There are no quick fixes. Proposals such as changing the default p-value threshold for statistical significance, employing confidence intervals with a focus on whether or not they contain zero, or employing Bayes factors along with conventional classifications for evaluating the strength of evidence suffer from the same or similar issues as the current use of p-values with the 0.05 threshold. In particular, each implicitly or explicitly categorizes evidence based on thresholds relative to the generally uninteresting and implausible null hypothesis of zero effect and zero systematic error. Further, each is a purely statistical measure that fails to take a more holistic view of the evidence that includes the consideration of the traditionally neglected factors, that is, prior and related evidence, plausibility of mechanism, study design and data quality, real world costs and benefits, novelty of finding, and other factors that vary by research domain.

In brief, each is a form of statistical alchemy that falsely promises to transmute randomness into certainty, an "uncertainty laundering" (Gelman, 2016) that begins with data and concludes with dichotomous declarations of truth or falsity—binary statements about there being "an effect" or "no effect"—based on some p-value or other statistical threshold being attained. A critical first step forward is to begin accepting uncertainty and embracing variation in effects (Carlin, 2016; Gelman, 2016) and recognizing that we can learn much (indeed, more) about the world by forsaking the false promise of certainty offered by such dichotomization.

As noted, we have no desire to "ban" p-values or other purely statistical measures. Rather, we believe that such measures should not be thresholded and that, thresholded or not, they should not take priority over the neglected factors.

Instead, we offer recommendations for how, in practice, the p-value can be demoted from its threshold screening role and instead be considered as just one among many pieces of evidence. First, we recommend authors use the neglected factors to motivate their data collection, statistical analysis, interpretation of results, writing, and related matters; we also recommend they analyze and report all of their data and relevant results. Second, we recommend editors and reviewers explicitly evaluate papers with regard to not only purely statistical measures but also the neglected factors.

As a highly inter-disciplinary research team with representation from statistics, political science, psychology, and consumer behavior, we are acutely aware that the implementation of our broad recommendations will and ought to vary tremendously across—and even within—a given domain or subdomain of the biomedical and social sciences. We are not so supercilious to believe that we, by ourselves, are capable of providing concrete and specific guidance on the application of these recommendations across all or perhaps even any of these domains and subdomains. Indeed, we do not believe a "template" for our recommendations is possible or desirable. In fact, such a template could even be dangerous in that it might result in a rote and recipe-like application of our recommendations that would not be entirely dissimilar to, even if perhaps less harmful than, the current practice of rote and recipe-like application of NHST. To those who might argue that, without such a template, our recommendations are unrealistic or unlikely to be adopted in practice, we reiterate that statistics is hard especially in the biomedical and social sciences and a formulaic approach to statistics is a principal cause of the current replication crisis. It is for these reasons we advocate this more radical and more difficult but also more principled and permanent approach. Nonetheless, we below suggest some broad principles that show how our recommendations might be applied. We also provide a case study in Appendix B.

4.2 For Authors

We recommend authors use the neglected factors to motivate their data collection, statistical analysis, interpretation of results, writing, and related matters; we also recommend they analyze and report all of their data and relevant results rather than focusing on single comparisons that attain some p-value or other statistical threshold.

One specific operationalization of this recommendation might be to include in their manuscripts a section that directly addresses how each of the neglected factors motivated their various decisions regarding data collection, statistical analysis, interpretation of results, and writing in the context of the totality of the data and results. Such a section could, for example, discuss study design in the context of subject-matter knowledge and expectations of effect sizes, as discussed by Gelman and Carlin (2014). It could also discuss the plausibility of the mechanism by (i) formalizing the hypothesized mechanism for the effect in question and explicating the various components of it, (ii) clarifying which components were measured and analyzed in the study, and (iii) discussing aspects of the results that support the proposed mechanism as well as those that are in conflict with it.

One might think that this recommendation—analyzing and reporting all of the data and relevant results—is such a fundamental principle of science that it need hardly be mentioned. However, this is not the case! As discussed above, the status quo in scientific publication is a lexicographic decision rule whereby $p < 0.05$ is virtually always required for a result to be published and, while there are some exceptions, standard practice is to focus on such results and to not report all relevant findings.

Given the current state of practice, authors may not have a sense for how they might go about this. Rather than attempt to provide broad guidance, we direct the reader to illustrations in clinical psychology (Tackett *et al.*, 2014), epidemiology (Gelman and Auerbach, 2016a,b), political science (Trangucci *et al.*, 2018), program evaluation (Mitchell *et al.*, 2018; forthcoming), and social psychology and consumer behavior (McShane and Böckenholt, 2017) .

4.3 For Editors and Reviewers

We recommend editors and reviewers explicitly evaluate papers with regard to not only purely statistical measures but also the neglected factors; this should be far superior to the status quo, namely giving consideration, often scant, to the neglected factors only once some p-value or other statistical threshold has been reached.

One specific operationalization of this recommendation might be to incorporate consideration of these factors into various stages of the review process. For example, editors could require reviewers to provide quantitative evaluations of each factor—including domain-specific factors determined by the editor—as well as an overall quantitative evaluation of the strength of the evidence as a supplement to the current open-ended, qualitative evaluations. These could then be weighted by the editors' publicly-disclosed (or even reviewers' own) importance rating of each factor. Additionally, editors could discuss and address the evaluation and importance of each factor in decision letters, thereby providing a more holistic view of the evidence.

One might object here and call our position naive: do not editors and reviewers require some bright-line threshold to decide whether the data supporting a claim is far enough from pure noise to support publication? Do not statistical thresholds provide objective standards for what constitutes evidence and does this not in turn provide a valuable brake on the subjectivity and personal biases of editors and reviewers? Against these, we would argue that even were such a threshold needed, it does not make sense to set it based on the p-value, given that this statistic is, in the overwhelming majority of applications, defined relative to the generally uninteresting and implausible null hypothesis of zero effect and zero systematic error and given that the costs and benefits of publishing noisy results varies by field. Additionally, the p-value is not a purely objective standard: different model specifications and statistical tests for the same data and null hypothesis yield different p-values; to complicate matters further, many subjective decisions regarding data protocols and analysis procedures such as coding and exclusion are required in practice and these often strongly impact the p-value ultimately reported. Finally, we fail to see why such a threshold screening rule is needed: editors and reviewers already make publication decisions one at a time based on qualitative factors, and this could continue to happen if the p-value were demoted from a default screening rule to merely one piece of evidence. Indeed, no single number—whether it be a p-value, Bayes factor, or some other statistical or non-statistical measure—is capable of eliminating subjectivity and personal biases.

Instead, we believe it is entirely acceptable to publish an article featuring a result with, say, a p-value of 0.2 or a 90% confidence interval that includes zero provided it is relevant to a theoretical or applied question of interest and the interpretation is sufficiently accurate. It should also be possible to publish a result with, say, a p-value of 0.001 without this being taken to imply the truth of some favored alternative hypothesis.

The p-value is relevant to the question of how easily a result could be explained by a particular null model, but there is no reason this should be the crucial factor in publication. A result can be consistent with a null model but still be relevant to science or policy debates, and a result can reject a null model without offering anything of science or policy interest.

In sum, editors and reviewers can and should feel free to accept papers and present readers with the relevant evidence. We would much rather see a paper that, for example, states that there is weak evidence for an interesting finding but that existing data remain consistent with null effects, than for the publication process to screen out such findings or encourage authors to cheat to obtain statistical significance.

4.4 Abandoning Statistical Significance Outside Scientific Publishing

While our focus has been on statistical significance thresholds in scientific publication, similar issues arise in other areas of statistical decision making, including for example neuroimaging where researchers use voxelwise NHSTs to decide which results to report or take seriously; medicine where regulatory agencies such as the Food and Drug Administration use NHSTs to decide whether or not to approve new drugs; policy analysis where non-governmental and other organizations use NHSTs to determine whether interventions are beneficial or not; and business where managers use NHSTs to make binary decisions via A/B testing. In addition, thresholds arise not just around scientific publication but also within research projects, when researchers use NHSTs to decide which avenues to pursue further based on preliminary findings.

While considerations around taking a more holistic view of the evidence and consequences of decisions are rather different across each of these settings and different from those in scientific publication, we nonetheless believe our proposal to demote the p-value from its threshold screening role and emphasize the neglected factors applies in these settings. For example, in neuroimaging, the voxelwise NHST approach misses the point in that there are typically no true zeros and changes are generally happening at all brain locations at all times. Graphing images of estimates and uncertainties makes sense to us, but we see no advantage in using a threshold. For regulatory, policy, and business decisions, cost-benefit calculations seem clearly superior to acontextual statistical thresholds.

In these settings, we acknowledge thresholds—of a non-statistical variety—may sometimes be useful. For example, consider a firm contemplating sending a costly offer to customers. Suppose the firm has customer-level model of the revenue expected in response to the offer. In this setting, it could make sense for the firm to send the offer only to customers that yield an expected profit greater than some threshold, say zero.

Even in pure research scenarios where there is no obvious cost-benefit calculation—for example a comparison of the underlying mechanisms, as opposed to the efficacy, of two drugs used to treat some disease condition—we see no value in p-value or other statistical thresholds. Instead, we'd like to see researchers simply report results: estimates, standard errors, confidence intervals, etc., with statistically inconclusive results being relevant for motivating future research.

While we see the intuitive appeal of using p-value or other statistical thresholds as a screening device to decide what avenues (e.g., ideas, drugs, or genes) to pursue further, this approach fundamentally does not make efficient use of data. As we noted, it seldom makes sense to calibrate scientific evidence as a function of the p-value, given that it is often defined relative to the generally uninteresting and implausible null hypothesis. However, even when the null hypothesis is plausible (e.g., studying the association between a large number of genes and some outcome), p-values and other purely statistical measures are and ought to be treated as continuous measures of the strength of evidence; put differently, even when the truth about some underlying phenomenon is binary or discrete, our degree of knowledge of and uncertainty in the phenomenon is generally continuous. Finally, and perhaps most importantly, there is no general connection between a p-value—a probability based on a particular null model—and either the potential gains from pursuing a potential research lead or the predictive probability that the lead in question will ultimately be successful.

Instead, to the extent that decisions do need to be made about which lines of research to pursue further, we recommend making such decisions using a model of the distribution of effect sizes and variation, thus working directly with hypotheses of interest rather than reasoning indirectly from a null model. We'd also like to see—when possible in these and other settings—more precise individual-level measurements, a greater use of within-person or longitudinal designs, and increased consideration of models that use informative priors, that feature varying treatment effects, and that are multilevel or meta-analytic in nature (Gelman, 2015, 2017; McShane and Böckenholt, 2017, 2018).

Our recommendations will not themselves resolve the replication crisis in science, but we believe they will have the salutary effect of pushing researchers away from the pursuit of irrelevant statistical targets and toward understanding of theory, mechanism, and measurement. We also hope they will push them to move beyond the paradigm of routine "discovery," and binary statements about there being "an effect" or "no effect," to one of continuous and inevitably flawed learning that is accepting of uncertainty and variation.

4.5 Wrapping It Up

In this paper, we have proposed to abandon statistical significance and offered recommendations for how this can be implemented in the scientific publication process as well as in statistical decision making more broadly Our recommendations will not themselves resolve the replication crisis in science, but we believe they will have the salutary effect of pushing researchers away from the pursuit of irrelevant statistical targets and toward understanding of theory, mechanism, and measurement. We also hope they will push them to move beyond the paradigm of routine "discovery," and binary statements about there being "an effect" or "no effect," to one of continuous and inevitably flawed learning that is accepting of uncertainty and variation.

But, how do we get from here—NHST, deterministic summaries, overconfidence in results, and statistical analysis focused on reporting just some of the data—to there—statistical analysis and interpretation of results that accepts uncertainty and embraces variation and features full reporting of results rather than focusing on whatever happens to exceed some statistical threshold?

We have offered the recommendations that we believe will serve researchers best. However, we recognize that research takes place within an institutional structure that often encourages behavior that is counter to these recommendations. Researchers respond to the expectations of funding agencies in study design and editors and reviewers in writing. Conversely, funding agencies must choose among the submissions they receive and editors can only publish papers that are submitted to them. A careful research proposal that openly grapples with uncertainty may unfortunately lose out in the funding competition to a more traditional proposal that blithely promises 80% power based on selected and overestimated effect sizes. Similarly, a paper that presents all the data without making inappropriate claims of certainty may not get published in a journal that is also receiving submissions in which statistically significant results are presented at face value.

These institutional problems are difficult and we do not propose solutions to them. We imagine improvement will come in fits and starts, in several parallel tracks, all of which we and others have tried to contribute to in our applied and methodological research: improved statistical methods that move beyond the NHST paradigm and include multilevel modeling, machine learning, statistical graphics, and other tools for analyzing and visualizing large amounts of data; applied examples using these improved methods, thereby demonstrating that it is possible to perform a successful statistical analysis without aiming for deterministic results; theoretical work on the statistical effects of selection based on statistical significance and other decision criteria; and criticism of published work with gross overestimates of effect sizes or inappropriate claims of certainty. While we recognize change will likely require institutional reform including major modifications of current practices of funding agencies and editors and reviewers, we are also optimistic that some combination of recognition of error and the awareness of alternatives can also motivate change.

A Problems with Uniformly Most Powerful Bayesian Tests

The mathematical justification underlying the Benjamin *et al.* (2018) proposal has come under no small amount of criticism. Specifically, the uniformly most powerful Bayesian tests (UMPBTs) that underlie the proposal were introduced and defended by Johnson (2013b) in parallel with his call in Johnson (2013a)—and now repeated in Benjamin *et al.* (2018)—to use 0.005 as the new threshold. We see a number of concerns with UMPBTs.

First, there is no reason for non-Bayesians to adopt UMPBTs when they can instead rely on the standard Neyman-Pearson approach to uniformly most powerful (non-Bayesian) tests.

Second, and perhaps most relevant for the biomedical and social sciences, the UMPBT approach is deeply entrenched in the century-old Neyman-Pearson formalism of binary decisions and 0-1 loss functions. As Pericchi *et al.* (2014) note, even in settings where the NHST paradigm is reasonable, "the essence of the problem of classical testing of significance lies in its goal of minimizing Type II error (false negative) for a fixed Type I error (false positive)." While this formalism allows for mathematical optimization under some restricted collection of distributions and testing problems, it is quite rudimentary from a decision-theoretic point of view, even to the extent of failing most purposes of running a sharp point null hypothesis test.

More specifically, the 0-1 loss function implicit in the NHST paradigm does not in general map, even in an approximate way, to processes of scientific learning or costs and benefits. In particular, even if the proposal to move to a lower p-value threshold is good advice in certain application areas, the fact remains that the logic underlying it avoids firmly confronting the nature of the issue: any such rule implicitly expresses a particular tradeoff between Type I and Type II error, but in reality this tradeoff should depend on the costs, benefits, and probabilities of all outcomes (Gelman and Robert, 2014) which depend on the problem at hand and vary tremendously across studies in the biomedical and social sciences. Instead, the UMPBT is based on a minimax prior that does not correspond to any distribution of effect sizes but rather represents a worst case scenario under a set of mathematical assumptions.

Third, defining the dependence of the procedure over a threshold (γ in the notation of Johnson (2013b)) replicates the fundamental difficulty with the century-old Fisherian answer to hypothesis testing. To further seek a full agreement with the classical rejection region as advocated by Johnson (2013b) is to simply negate the appeal of a truly Bayesian approach to this issue; moreover, this agreement is impossible to achieve for realistic statistical models.

Fourth, the construction of a UMPBT relies on the assumption of a "true" prior, which can be criticized in a vast majority of cases and which in any case moves one away from the Bayesian paradigm: with a single and "true" prior, the Bayesian model becomes an errors-in-variables model.

Fifth, the argument to maximize a probability for the Bayes factor to exceed a certain threshold also moves one away from the Bayesian paradigm because: (i) it ignores the motives for running the NHST and the subsequent steps taken in decision making or inference; (ii) it further negates any prior modeling of the alternative hypothesis aiming at separating the parameter space into regions of different (prior) likelihood; (iii) it does not condition upon the actual observation but instead integrates over the observation space and hence may fall afoul of the likelihood principle; (iv) it posits a single and fixed threshold γ for rejecting the null when there is no reason for γ not to depend on the observed data, as also argued above; (v) the maximization step eliminates the role of the prior distribution, as also argued above; (vi) in the rare one-dimensional settings where the maximization step can be conducted in closed form, the solution is a distribution with finite support; (vii) in the event the null hypothesis is rejected, the uniformly most powerful prior (or alt-prior) corresponding to the alternative cannot be used as such in subsequent inference but must instead be replaced with a regular prior over the whole parameter space, which is a strong violation of Bayesian coherence.

Sixth, speaking more generally, the concept of uniformly most powerful priors (and tests) does not easily extend to multivariate settings and even less to realistic cases that involve complex null hypotheses that contain nuisance parameters. The first solution proposed in Johnson (2013b), to integrate out the nuisance parameters in the null hypothesis using a specific prior distribution, falls short of solving the issue of "objective Bayesian tests." The second solution, namely to replace the unknown nuisance parameters with standard estimates, stands even farther from a Bayesian perspective.

Indeed, the Bayes factor itself is a consequence of the rudimentary Neyman-Pearson formalism, which as such caters to the issue of statistical significance. A discussion of the difficulties with this from a Bayesian perspective is provided in Kamary *et al.* (2014), with a proposal of setting the hypothesis problem as one of mixture estimation.

Seventh, Johnson (2013b) contains very little support for the asymptotic relevance of the approach, beyond the limiting normal distribution of the uniformly most powerful log Bayes factor and the convergence of the support to the "true" value of the parameter.

In closing, we note that many of our criticisms of the Johnson (2013b) approach relate to the fact that it falls short of being truly Bayesian. However, we do not mean to say that hypothesis testing must be done in a Bayesian manner. Rather, we emphasize this because, to the extent that the Johnson (2013b) approach loses its Bayesian connection, it also loses the Bayesian justification of the 0.005 rule. Consequently, 0.005 becomes just another arbitrary threshold, justified by some implicit tradeoff between false positives and negatives which we think does not make sense in any absolute and acontextual way.

B Case Study

In the context of a hypothetical case study on the effects of sodium on blood pressure, we discuss how authors as well as editors and reviewers might follow our recommendation to demote the p-value from its threshold screening role and instead, treated continuously, be considered along with the neglected factors—prior and related evidence, plausibility of mechanism, study design and data quality, real world costs and benefits, novelty of finding, and other factors that vary by research domain—as just one among many pieces of evidence.

We recommend authors use the neglected factors to motivate their data collection, statistical analysis, interpretation of results, writing, and related matters. In this example, the authors might consider *prior and related evidence* that indicates the importance of blood pressure as a marker for healthy arteries, suggests the role of sodium in hemodynamics, and so forth. This evidence might also reveal a *plausible mechanism*, namely to excrete excess sodium the body must increase blood pressure.

In terms of *study design and data quality*, the authors might consider various possibilities for data collection. How should they recruit subjects? Should they randomize them to a low-sodium versus high-sodium diet? Or should they track them longitudinally, say via routine annual checkups over the course of years? Or is such data already available from some prior study? When and how often should sodium and blood pressure be measured? And how? The authors might measure sodium through a relatively dietary recall questionnaire (noisy), through asking participants to maintain a food diary (somewhat less noisy), or through collection of urine to measure urinary sodium excretion (precise but restricted to a limited time point). Likewise, for blood pressure, they might rely on measurements conducted by someone convenient like friends or family members of the subjects who likely do not possess formal clinical training (noisy) or by paid clinicians instructed on the proper protocol for blood pressure measurement (precise).

Suppose the authors have a hypothesis of an association between sodium consumption and high blood pressure. For the moment, let's assume their statistical analysis is restricted to the current practice of NHST and they obtain a p-value of 0.0001. How should this impact their interpretation of results, writing, and statistical decision making more broadly? Certainly, they have gained support for their hypothesis. However, can they conclude sodium is associated with—or even causes—high blood pressure more broadly?

Well, it would depend on the context and limitations of the *study design and data quality*. For example, supposing the study took place in Japan, perhaps the association exists in the Japanese subject population studied but does not in European populations whether because of some genetic differences between the two populations or because of some dietary differences (e.g., dietary sodium levels are much higher among Japanese so the association might not hold over levels typical among Europeans).

In terms of a causal interpretation, this would depend on *prior and related evidence*, *plausibility of mechanism*, and *study design and data quality*. If prior studies show consistent and strong associations between sodium consumption and blood pressure, if evidence from physiological studies and animal models are consistent with an effect of sodium consumption on blood pressure, or if sodium levels were randomized, this increases the support for a causal role of sodium in increasing sodium.

Given, say, that the causal interpretation holds and holds broadly, the authors could then consider clinical significance or *real world costs and benefits*. This depends not at all on a p-value but the estimates of magnitude of the effect—not only on blood pressure but on downstream outcomes such as cardiovascular disease. It also depends on costs of potential interventions such as lower sodium diets and drugs. They could also discuss *novelty of finding* in light of all of the above.

Now, suppose they had instead obtained a p-value of 0.2. Can they conclude sodium is not associated with high blood pressure more broadly? Again, this would depend on all the factors discussed above. For example, perhaps the association does not exist in the Japanese population but does in European ones and so on.

However, we do not believe the authors' statistical analysis should be restricted to the current practice of NHST. Instead, we recommend authors report all of their data and relevant results rather than focusing on single comparisons that attain some p-value or other statistical threshold. In this context, this might involve modeling the association between sodium and blood pressure as a function of additional health and dietary variables, demographic variables, and geography using a multilevel model. Such a model would not yield one single p-value thereby encouraging dichotomous declarations of truth or falsity—binary statements about there being "an effect" or "no effect." Instead, it would many estimates that vary based on, for example, health and dietary variables, demographic variables, and geography, as well as the uncertainty in these estimates. Indeed, by accepting uncertainty and embracing variation in effects, the authors would discover and present a much richer and more nuanced story about the association between sodium and blood pressure.

Turning to editors and reviewers, we recommend they explicitly evaluate papers with regard to not only purely statistical measures but also the neglected factors. How might this work? We envision it would be rather similar to the above but in reverse. Specifically, editors and reviewers evaluating the authors' paper on sodium and blood pressure would systemically assess, and possibly even indicate the weight they assign, to each of the following: How does the paper fit in with and build upon *prior and related evidence*? Is the *mechanism plausible*? Are the *study design and data quality* sufficient to justify the conclusions? What are the implications in terms of *real world costs and benefits*? How *novel* are the findings? And, of course, how appropriate are the statistical analyses employed and how strong is the statistical support, whether in the form of a p-value or some other measure, resulting from these analyses?

In this more holistic view of the evidence, statistical measures are just one among many pieces of evidence considered by editors and reviewers and do not take priority. Of course, this does not mean that they cannot or will not strongly impact or alter their evaluation decisions. For example, in the context of the authors' paper on sodium and blood pressure, strong statistical support, whether in the form of a low p-value or otherwise, for a finding that sodium consumption is associated with low blood pressure—the direction opposite of that indicated by prior evidence—in the context of a high quality study design featuring large samples and precise measurements might be deemed more novel and worthy of publication than if the statistical support had been weaker or the finding was in the same direction as that indicated by prior evidence.

In sum, authors as well as reviewers and editors need not use statistical significance as a lexicographic decision rule. Results need not first have a p-value or some other purely statistical measure that attains some threshold before consideration is given to the neglected factors. Instead, treated continuously, statistical measures should be considered along with the neglected factors as just one among many pieces of evidence and should not take priority thereby yielding a more holistic view of the evidence. This should be far superior to the status quo.

References

Amrhein, V. and Greenland, S. (2018). Remove, rather than redefine, statistical significance. *Nature Human Behaviour* **2**, 4.

Amrhein, V., Korner-Nievergelt, F., and Roth, T. (2017). The earth is flat (p¿ 0.05): significance thresholds and the crisis of unreplicable research. *PeerJ* **5**, e3544.

Anderson, D. R., Burnham, K. P., and Thompson, W. L. (2000). Null hypothesis testing: problems, prevalence, and an alternative. *Journal of Wildlife Management* **64**, 912–923.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin* **66**, 6, 423–437.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology* **100**, 407–425.

Benjamin, D. J., Berger, J. O., Johannesson, M., *et al.* (2018). Redefine statistical significance. *Nature Human Behaviour* **2**, 6–10.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association* **33**, 203, 526–536.

Boring, E. G. (1919). Mathematical vs. scientific significance. *Psychological Bulletin* **16**, 10, 335–338.

Briggs, W. M. (2016). *Uncertainty: The Soul of Modeling, Probability and Statistics.* Springer, New York, NY.

Carlin, J. B. (2016). Is reform possible without a paradigm shift? *The American Statistician, supplemental material to the ASA statement on p-values and statistical significance* **10**.

Carney, D. R., Cuddy, A. J., and Yap, A. J. (2010). Power posing brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science* **21**, 10, 1363–1368.

Cochran, W. G. (1976). *On the history of statistics and probability*, chap. Early development of techniques in comparative experimentation. Dekker, New York.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist* **49**, 997–1003.

Cowles, M. and Davis, C. (1982). On the origins of the .05 level of significance. *American Psychologist* **44**, 1276–1284.

Cox, D. R. (1977). The role of significance tests. *Scandinavian Journal of Statistics* **4**, 2, 49–70.

Cox, D. R. (1982). Statistical significance tests. *British Journal of Clinical Pharmacology* **14**, 325–331.

Cramer, H. (1955). *The Elements of Probability Theory.* Wiley, New York.

Drummond, G. (2015). Most of the time, p is an unreliable marker, so we need no exact cut-off. *British journal of anaesthesia* **116**, 6, 894–894.

Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological review* **70**, 3, 193.

Eysenck, H. J. (1960). The concept of statistical significance and the controversy about one-tailed tests. *Psychological Review* **67**, 4, 269.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture* **33**, 503–513.

Fisher, R. A. (1956). *Statistical methods and scientific inference.* Hafner Publishing Co.

Gelman, A. (2015). The connection between varying treatment effects and the crisis of unreplicable research: A bayesian perspective. *Journal of Management* **41**, 2, 632–643.

Gelman, A. (2016). The problems with p-values are not just with p-values. *The American Statistician, supplemental material to the ASA statement on p-values and statistical significance* **10**.

Gelman, A. (2017). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin* **44**, September.

Gelman, A. and Auerbach, J. (2016a). Age-aggregation bias in mortality trends. *Proceedings of the National Academy of Sciences* **113**, 7, E816–E817.

Gelman, A. and Auerbach, J. (2016b). Mortality trends by race/ethnicity, sex, age and state. Tech. rep., Columbia University.

Gelman, A. and Carlin, J. (2014). Beyond power calculations assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science* **9**, 6, 641–651.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL, 3rd edn.

Gelman, A. and Loken, E. (2014). The statistical crisis in science. *American Scientist* **102**, 6, 460–465.

Gelman, A. and Robert, C. P. (2014). Revised evidence for statistical standards. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 19, E1933–E1933.

Gelman, A. and Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician* **60**, 4, 328–331.

Gigerenzer, G. (1987). *The Probabilistic Revolution. Vol. II: Ideas in the Sciences*, vol. II. MIT Press, Cambridge, MA.

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics* **33**, 587–606.

Gigerenzer, G., Krauss, S., and Vitouch, O. (2004). *Handbook on Quantitative Methods in the Social Sciences*, chap. The null ritual: What you always wanted to know about null hypothesis testing but were afraid to ask., 389–406. Sage Publications, Inc, Thousand Oaks, CA.

Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly* **52**, 3, 647–674.

Greenland, S. (2017). Invited commentary: The need for cognitive science in methodology. *American Journal of Epidemiology* **186**, 6, 639–646.

Greenland, S. and Poole, C. (2013). Living with statistics in observational research. *Epidemiology* **24**, 1, 73–78.

Hubbard, R. (2004). Alphabet soup: Blurring the distinctions between p's and α's in psychological research. *Theory and Psychology* **14**, 295–327.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science* **8**, 3–7.

Hurlbert, S. H. and Lombardi, C. M. (2009). Final collapse of the neyman-pearson decision theoretic framework and rise of the neofisherian. *Annales Zoologici Fennici* **46**, 5, 311–349.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine* **2**, 8, e124.

Johnson, V. E. (2013a). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences* **110**, 48, 19313–19317.

Johnson, V. E. (2013b). Uniformly most powerful bayesian tests. *Annals of Statistics* **41**, 4, 1716–1741.

Kamary, K., Mengersen, K., Robert, C., and Rousseau, J. (2014). Testing hypotheses as a mixture estimation model. Tech. rep., `https://arxiv.org/pdf/1214.4436.pdf`.

Lakens, D., Adolfi, F. G., Albers, C. J., *et al.* (2018). Justify your alpha. *Nature Human Behaviour* **2**, 3, 168–171.

Lehmann, E. L. (1993). *Testing Statistical Hypotheses*. Chapman and Hall, New York.

Lemoine, N. P., Hoffman, A., Felton, A. J., Baur, L., Chaves, F., Gray, J., Yu, Q., and Smith, M. D. (2016). Underappreciated problems of low replication in ecological field studies. *Ecology* **97**, 10, 2554–2561.

McCloskey, D. N. and Ziliak, S. (1996). The standard error of regression. *Journal of Economic Literature* **34**, 97–114.

McShane, B. B. and Böckenholt, U. (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science* **9**, 6, 612–625.

McShane, B. B. and Böckenholt, U. (2017). Single paper meta-analysis: Benefits for study summary, theory-testing, and replicability. *Journal of Consumer Research* **43**, 6, 1048–1063.

McShane, B. B. and Böckenholt, U. (2018). Multilevel multivariate meta-analysis with application to choice overload. *Psychometrika* **83**, 1, 255–271.

McShane, B. B. and Gal, D. (2016). Blinding us to the obvious? the effect of statistical training on the evaluation of evidence. *Management Science* **62**, 6, 1707–1718.

McShane, B. B. and Gal, D. (2017). Statistical significance and the dichotomization of evidence. *Journal of the American Statistical Association* .

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir karl, sir ronald, and the slow progress of soft psychology. *Journal of Counseling and Clinical Psychology* **46**, 806–834.

Mitchell, S., Gelman, A., Ross, R., Chen, J., Bari, S., Huynh, U. K., Harris, M. W., Sachs, S. E., Stuart, E. A., Feller, A., *et al.* (2018; forthcoming). The millennium villages project: a retrospective, observational, endline evaluation. *The Lancet* .

Morrison, D. E. and Henkel, R. E. (1970). *The Significance Test Controversy*. Aldine, Chicago.

Pericchi, L., Pereira, C. A., and Pérez, M.-E. (2014). Adaptive revised standards for statistical evidence. *Proceedings of the National Academy of Sciences* **111**, 19, E1935–E1935.

Resnick, B. (2017). What a nerdy debate about p-values shows about science — and how to fix it. Tech. rep.

Rosnow, R. L. and Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist* **44**, 10, 1276–1284.

Rozenboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin* **57**, 416–428.

Sawyer, A. G. and Peter, J. P. (1983). The significance of statistical significance tests in marketing research. *Journal of Marketing Research* **20**, 2, 122–133.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods* **1**, 115–129.

Senn, S. S. (2001). Two cheers for p-values? *Journal of Epidemiology and Biostatistics* **6**, 2, 193–204.

Serlin, R. C. and Lapsley, D. K. (1993). *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, chap. Rational Appraisal Psychological Research and the Good Enough Principle. Lawrence Erlbaum Associates, Hillsdale, NJ.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology : Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* **22**, 11, 1359–1366.

Skipper, J. K., Guenther, A. L., and Nass, G. (1967). The sacredness of. 05: A note concerning the uses of statistical levels of significance in social science. *The American Sociologist* **5**, 16–18.

Smaldino, P. E. and McElreath, R. (2016). The natural selection of bad science. Tech. rep., `https://arxiv.org/pdf/1605.09511v1.pdf`.

Tackett, J. L., Kushner, S. C., Herzhoff, K., Smack, A. J., and Reardon, K. W. (2014). Viewing relational aggression through multiple lenses: Temperament, personality, and personality pathology. *Development and psychopathology* **26**, 3, 863–877.

Trangucci, R., Ali, I., Gelman, A., and Rivers, D. (2018). Voting patterns in 2016: Exploration using multilevel regression and poststratification (mrp) on pre-election polls. *arXiv preprint arXiv:1802.00842* .

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science* **6**, 100–116.

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

Yule, G. U. and Kendall, M. G. (1950). *An introduction to the theory of statistics*. Griffin, London, 14th edn.