Abandon Statistical Significance*

Blakeley B. McShane[1], David Gal[2], Andrew Gelman[3],
Christian Robert[4], and Jennifer L. Tackett[1]

[1]Northwestern University, [2]University of Illinois at Chicago,
[3]Columbia University, [4]Université Paris-Dauphine

21 Sep 2017

Abstract

In science publishing and many areas of research, the status quo is a lexicographic decision rule in which any result is first required to have a p-value that surpasses 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. There have been recent proposals to change the p-value threshold, but instead we recommend abandoning the null hypothesis significance testing paradigm entirely, leaving p-values as just one of many pieces of information with no privileged role in scientific publication and decision making. We argue that this radical approach is both practical and sensible.

1. Introduction: 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 tiny, sometimes preposterous, interventions. 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. Statistical significance serves as a lexicographic decision rule whereby any result is first required to have a p-value that surpasses 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 (in the sequel, we refer to these collectively as the neglected factors).

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

*We thank Monya Baker and two anonymous reviewers for helpful comments. We also thank the National Science Foundation, Institute for Education Sciences, Office of Naval Research, and Defense Advanced Research Projects Agency for partial support of this work.

work has made it clear that so-called “researcher degrees of freedom” [Simmons et al., 2011] are abundant enough for statistical significance to 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 change the default p-value threshold for statistical significance for claims of new discoveries from 0.05 to 0.005” [Benjamin et al., 2017]. We believe this proposal is insufficient to overcome current difficulties with replication—and, perhaps curiously, we expect those authors would agree given that 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 [NHST].”

In particular, we disagree with their emphasis on a particular immediate action—“changing the p-value threshold is simple, aligns with the training undertaken by many researchers, and might quickly achieve broad acceptance.” We are not convinced this step would be helpful. In the short term, a more stringent threshold 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, a steeper cutoff could lead to even more overconfidence in results that do get published as well as 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 it. 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 p-value thresholds (as well as those based on other statistical measures) are a bad idea in general.

Instead, and consequently, we propose *another alternative*, which is to abandon statistical significance. In particular, rather than propose a “quick fix,” “a dam to contain the flood until we make sure we have the more permanent fixes” in the words of a prominent member of the seventy-two [Resnick, 2017], we recommend dropping the NHST paradigm—and the p-value thresholds associated with 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 statistical threshold) to serve as a lexicographic decision rule in scientific publication and statistical decision making more broadly as per the status quo, 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 standard use of NHST—to take the rejection of this straw man sharp point null hypothesis as positive or even definitive evidence in favor of some preferred alternative hypothesis—is a logical fallacy that routinely results in erroneous scientific reasoning even by experienced scientists and statisticians. Third, *p*-value and other statistical thresholds encourage researchers to study and report single comparisons rather than focusing on the totality of their data and results.

Before elaborating on our own suggestions for improving replicability, we discuss general problems with NHST that remain unresolved by the Benjamin et al. [2017] proposal as well as problems specific to the proposal. We then discuss the implications of abandoning statistical significance for the scientific publication process as well as for statistical decision making more broadly.

2. Problems with null hypothesis significance testing

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, so that even when sample sizes are large, the possibilities of systematic variation and bias results in the equivalent of small or unrepresentative samples. Consequently, estimates from any single study are themselves generally noisy. This, in combination with the fact that the single study is typically the fundamental unit of analysis, poses problems for the NHST paradigm.

Given that effects are small and variable and measurements are noisy, the sharp point null hypothesis of zero effect and zero systematic error used in the overwhelming majority of applications 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]. Consequently, Cohen [1994] has derided this null hypothesis as the "nil hypothesis" and lampoons it as "always false," and Tukey [1991] notes that two treatments are "always different." Indeed, even were an effect truly zero, experimental realities dictate that the effect would not be exactly zero in any study designed to test it.

In addition, noisy estimates in combination with a publication process that screens for statistical significance results in published estimates that are biased upwards (potentially to a large degree) and often of the wrong sign [Gelman and Carlin, 2014]. Indeed, the screening of estimates for statistical significance by the publication process provides an indirect incentive for researchers to conduct many small noisy studies, resulting in estimates that can be made to yield one or more statistically significant results [Simmons et al., 2011]. 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—small and variable effects, noisy measurements, a publication process that screens for statistical significance, 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.

More broadly, 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. Given this, one well-known criticism of the NHST paradigm is that the conventional 0.05 threshold—or for that matter any cutoff—is entirely arbitrary [Fisher, 1926, Yule and Kendall, 1950, Cramer, 1955, Cochran, 1976, Cowles and Davis, 1982].

A related line of criticism 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 further and say that it does not in general make sense to calibrate scientific evidence as a function of the p-value, given that this statistic is defined relative to the generally uninteresting and implausible null hypothesis of zero effect and zero systematic error.

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.

Issues related to the dichotomization of evidence intrinsic to the NHST paradigm have received a great deal of attention of late due at least in part to the recent American Statistical Association Statement on Statistical Significance and p-values [Wasserstein and Lazar, 2016], which indeed explicitly cautions against proposals like that of Benjamin et al. [2017] in its third of six principles (“scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold”).

3. Specific issues with the proposed $p < 0.005$ threshold

Beyond concerns about the use of NHST in the biomedical and social sciences and more generally, there are specific problems with the Benjamin et al. [2017] proposal to change the default p-value threshold for statistical significance. First, Benjamin et al. [2017] 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, their restriction of their proposed rule to new effects is problematic in and of itself. Because they fail to define what constitutes a new effect, their recommendation is rendered entirely impractical; this is especially so in domains where research is believed to be incremental and cumulative. The proposed policy also would lead to incoherence when applied to replication—the very issue the recommendation 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 (in which 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. [2017] 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 there 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. [2017].

Fourth, the mathematical justification underlying the Benjamin et al. [2017] 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 [2013a] in parallel with his call in Johnson [2013b]—and now repeated in Benjamin et al. [2017]—to use 0.005 as the new threshold. We see a number of concerns with UMPBTs.

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.

Defining the dependence of the procedure over a threshold (γ in the notation of Johnson [2013a]) 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 [2013a] 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.

Speaking more generally, the notion 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 [2013a], 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.

We do not mean to say that hypothesis testing must be done in a Bayesian manner. However, to the extent that the Johnson [2013a] 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.

4. Moving forward by abandoning the privileged role of statistical significance

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 surpassed. 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.

We have no desire to "ban" p-values. Instead, we offer two concrete recommendations—one for editors and reviewers and one for authors—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, for editors and reviewers, insofar as goals of correctness and novelty are in conflict, we believe a major improvement on the status quo would be to take an explicit and systematic stance with regard to the importance of the neglected factors. One specific operationalization of this recommendation might be to incorporate consideration of these factors into various stages of the review process. For example, journal portals could solicit feedback from reviewers on each factor—including domain-specific factors determined by the editor—as well as the overall strength of the evidence in addition to allowing open-ended justification for these evaluations; such evaluations could then be weighed by the editors' publically-disclosed (or even the reviewers' own) importance ratings for each factor. Additionally, editors could discuss and address the importance and evaluation of each factor in decision letters, thereby ensuring a more holistic view of the evidence.

One might object here and call our position naive: do not journals 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 which in turn provide a valuable brake on the subjectivity and personal biases of journal 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 the costs and benefits of publishing noisy results varies so much based on context. 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 journals need such a threshold screening rule. Journals 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 measure—is capable of eliminating subjectivity and personal biases.

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 theory 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, journal editors 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 by exploiting their researcher degrees of freedom to obtain statistical significance.

Second, for authors, we recommend studying and reporting the totality of their data and relevant results rather than focusing on single comparisons that surpass some p-value or other statistical threshold. In doing so, we recommend that authors use the neglected factors to motivate their statistical analyses and writing. For example, they might include in their manuscripts a section that directly addresses each in turn in the context of the totality of their data and results. For example, this section could discuss the study design in the context of subject-matter knowledge and expectations of effect sizes, for example as discussed by Gelman and Carlin [2014]. As another example, this section could 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 data results that support the proposed mechanism as well as those (in the full data) that are in conflict with it.

One might think that this recommendation—studying and reporting on the totality of the data and 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. This clearly impacts authors' studying and reporting of their data and results.

A potential objection to our recommendation is that it is ideally suited to preregistered studies where decisions regarding protocols, analysis procedures, and comparisons of interest are made in advance of data collection—and thus are independent of the particular sample of data that happened to be collected. While we agree the recommendation is cleanest in this context, even in the absence of preregistration one can derive some idea of the multiplicity of protocols, analysis procedures, and comparisons possible based on the data that happened to be collected using, for example, multiverse analysis [Steegen et al., 2016].

Our focus has been on statistical significance thresholds in scientific publication, but the same 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 decide which avenues are worth pursuing based on preliminary findings.

Our proposal to demote the p-value from its threshold screening role and emphasize the neglected factors applies to all of 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.

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 our hypothetical mechanism researcher to simply report the 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-values or other statistical thresholds as a screening device to decide what lines of research—for example, ideas, drugs, or genes—to pursue further, fundamentally this approach does not make efficient use of data: there is no general connection between a null hypothesis-based probability and the potential gains from pursuing a potential research lead or even 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 this and all 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, McShane and Böckenholt, 2017a,b, Gelman, 2017].

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.

References

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

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

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, et al. Redefine statistical significance. *Nature Human Behaviour*, 2017.

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

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

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

William G. Cochran. Early development of techniques in comparative experimentation. In D. B. Owen, editor, *On the History of Statistics and Probability*, pages 1–26. Marcel Dekker, New York, 1976.

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

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

Harald Cramer. *The Elements of Probability Theory.* Wiley, New York, 1955.

Ward Edwards, Harold Lindman, and Leonard J. Savage. Bayesian statistical inference for psychological research. *Psychological Review*, 70(3):193, 1963.

Ronald A. Fisher. The arrangement of field experiments. *Journal of the Ministry of Agriculture*, 33:503–513, 1926.

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

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

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

Andrew Gelman and John Carlin. Beyond power calculations assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651, 2014.

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

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

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

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, London, 3rd edition, 2014.

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

Valen E. Johnson. Uniformly most powerful Bayesian tests. *Annals of Statistics*, 41(4): 1716–1741, 2013a.

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

Kaniav Kamary, Kerrie L. Mengersen, Christian P. Robert, and Judith Rousseau. Testing hypotheses as a mixture estimation model. Technical report, https://arxiv.org/abs/1412.2044, 2014.

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

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

Blakeley B. McShane and Ulf Böckenholt. Multilevel multivariate meta-analysis with application to choice overload. *Psychometrika*, 2017b.

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

Blakeley B. McShane and David Gal. Statistical significance and the dichotomization of evidence. *Journal of the American Statistical Association*, 2017.

Paul E. Meehl. 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, 1978.

Luis Pericchi, Carlos A. B. Pereira, and María-Eglée Pérez. Adaptive revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, 111(19):E1935–E1935, 2014.

Brian Resnick. What a nerdy debate about p-values shows about science—and how to fix it. Technical report, `https://www.vox.com/science-and-health/2017/7/31/ 16021654/p-values-statistical-significance-redefine-0005`, 2017.

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

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

Paul E. Smaldino and Richard McElreath. The natural selection of bad science. Technical report, `https://arxiv.org/pdf/1605.09511v1.pdf`, 2016.

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11:702–712, 2016.

John W. Tukey. The philosophy of multiple comparisons. *Statistical Science*, 6:100–116, 1991.

Ronald L. Wasserstein and Nicole A. Lazar. The ASA's statement on p-values: Context, process, and purpose. *American Statistician*, 70(2):129–133, 2016.

George U. Yule and Maurice G. Kendall. *An Introduction to the Theory of Statistics*. Griffin, London, 14th edition, 1950.