

# “Statistical Significance” and Statistical Reporting: Moving Beyond Binary

Journal of Marketing  
 2024, Vol. 88(3) 1-19  
 © American Marketing Association 2024  
 Article reuse guidelines:  
[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)  
 DOI: 10.1177/00222429231216910  
[journals.sagepub.com/home/jmx](https://journals.sagepub.com/home/jmx)



Blakeley B. McShane, Eric T. Bradlow, John G. Lynch Jr. ,  
 and Robert J. Meyer

## Abstract

Null hypothesis significance testing (NHST) is the default approach to statistical analysis and reporting in marketing and the biomedical and social sciences more broadly. Despite its default role, NHST has long been criticized by both statisticians and applied researchers, including those within marketing. Therefore, the authors propose a major transition in statistical analysis and reporting. Specifically, they propose moving beyond binary: abandoning NHST as the default approach to statistical analysis and reporting. To facilitate this, they briefly review some of the principal problems associated with NHST. They next discuss some principles that they believe should underlie statistical analysis and reporting. They then use these principles to motivate some guidelines for statistical analysis and reporting. They next provide some examples that illustrate statistical analysis and reporting that adheres to their principles and guidelines. They conclude with a brief discussion.

## Keywords

null hypothesis significance testing, *P*-value, replication, sociology of science, statistical significance

## Introduction

### Background

Null hypothesis significance testing (NHST) is the default approach to statistical analysis and reporting in marketing and the biomedical and social sciences more broadly. As practiced, NHST centers on where the *P*-value<sup>1</sup>—a measure of the degree

of the compatibility of the observed data with both a target hypothesis (almost always the null hypothesis of no association or no effect) as well as countless other explicit, implicit, and overlooked background assumptions that vary by context (e.g., additivity, linearity, normality, random sampling of subjects from a population, random assignment of subjects to experimental conditions, properly functioning measurement devices, absence of data fabrication) as assessed by some test statistic (e.g., a *Z*-statistic, a  $\chi^2$ -statistic, or another quantity computed from the data)—stands relative to some threshold (almost always .05). If the *P*-value is less than the threshold, the association or effect is declared “statistically significant” and the target hypothesis is rejected, and this is deemed positive or even definitive evidence in favor of some preferred alternative hypothesis of an association or an effect. If the *P*-value is greater than the threshold, the association or effect is declared “statistically nonsignificant” and the target hypothesis is not rejected, and this is deemed positive or even definitive evidence in favor of the target hypothesis of no association or no effect.

<sup>1</sup> We capitalize *P*-value and distinguish it from its lowercase observed value *p*. The term “*P*-value” appeared by the 1920s (see, for example, Macarthur 1926; Putnam 1927; Sun 1928; the *P* is capitalized and italicized but without hyphen in these sources). According to Shafer (2020), “the term simply evolved from the use of the [capital, unitalicized] letter *P* to denote the probability that an estimated quantity or difference will fall inside or outside given limits ... in Fourier, Poisson, Gavarret, and Cournot” and especially from Karl Pearson’s use, dating from at least 1900, of capital, unitalicized *P* to denote the probability that some test statistic would be as extreme or more extreme (which in Pearson 1900 meant as large or larger) than the value of the test statistic computed from the observed data, which Pearson referred to as “the value of *P*” (Pearson 1900; Shafer 2020). Our practice is consistent with this. It is also consistent with the Neyman–Egon Pearson decision-theoretic distinction between the frequency properties of *P* and its observed value *p*, which parallels the convention of using uppercase for random variables and lowercase for their observed values. While the use of the term “*P*-value” is widespread, there is no consensus on font: we see it given in uppercase and lowercase, italicized and unitalicized, with and without hyphen. See Shafer (2020) for more on the history of the terms “*P*-value” and “statistical significance,” which notes that, despite the late nineteenth- and early twentieth-century origin of these terms, *P*-values and significance tests as statistical devices date from at least Arbuthnot (1710).

Blakeley B. McShane is Professor, Kellogg School of Management, Northwestern University, USA (email: [b-mcshane@kellogg.northwestern.edu](mailto:b-mcshane@kellogg.northwestern.edu)). Eric T. Bradlow is K.P. Chao Professor, The Wharton School, University of Pennsylvania, USA (email: [ebradlow@wharton.upenn.edu](mailto:ebradlow@wharton.upenn.edu)). John G. Lynch Jr. is University of Colorado Distinguished Professor, Leeds School of Business, University of Colorado, Boulder, USA (email: [john.g.lynch@colorado.edu](mailto:john.g.lynch@colorado.edu)). Robert J. Meyer is Frederick H. Ecker/MetLife Insurance Professor, The Wharton School, University of Pennsylvania, USA (email: [meyerr@wharton.upenn.edu](mailto:meyerr@wharton.upenn.edu))

Despite its default role, NHST has long been criticized by both statisticians and applied researchers, including those within marketing.<sup>2</sup> For example, Sawyer and Peter (1983, p. 122) noted 40 years ago in the *Journal of Marketing Research* that significance tests were being “misinterpreted and overvalued by marketing researchers.”

Several of the most prominent criticisms of NHST relate to the dichotomization of results—that is, into the categories “statistically significant” and “statistically nonsignificant” based on where the  $P$ -value stands relative to some threshold—intrinsic to it. For example, one problem with this dichotomization is that it leads researchers to wrongly interpret results that attain “statistical significance” as demonstrating an effect and those that fail to do so as demonstrating no effect. This in turn leads them, when acting as authors, editors, and reviewers, to use “statistical (non)significance” as a filter to select which results to publish, which again in turn biases the literature and encourages harmful research practices. Another problem with this dichotomization is that  $P$ -values naturally vary a great deal from study to study. Indeed, sampling variation alone can easily cause *large* differences in  $P$ -values—not only  $P$ -values that fall just barely to either side of some threshold.

### Illustration

To illustrate problems that result from the dichotomization of results intrinsic to NHST, consider two studies that are alike in every possible way except for their observed  $P$ -values. Suppose that the observed  $P$ -values in the two studies fall far from either side of the conventional .05 threshold, say,  $p = .005$  in the one study and  $p = .194$  in the other study.

We expect that many researchers would wrongly view the study with  $p = .005$  as a “success” that demonstrates some effect. We also expect that many researchers would wrongly view the study with  $p = .194$  as a “failure” that demonstrates no effect. However, it is wrong to conclude that some preferred alternative hypothesis of an effect is true on the basis of a small  $P$ -value alone. In addition, it is wrong to conclude that the target hypothesis of no effect is true on the basis of a large  $P$ -value alone.

Nonetheless, as a consequence of these incorrect beliefs, we expect that many researchers would wrongly filter which results are published. Specifically, we expect that many authors would rightly include the study with  $p = .005$  in their manuscript but wrongly exclude the study with  $p = .194$ . If both were rightly

included, we expect that many editors and reviewers would wrongly recommend excluding the study with  $p = .194$ . However, filtering which results are published biases the literature (filtering in the manner discussed here biases the literature upward in magnitude). Further, it encourages harmful research practices that yield results that pass the filter.

As a further consequence of these incorrect beliefs, we expect that many researchers would wrongly view the results of the two studies as incompatible with one another. However, the results are highly compatible:  $P$ -values naturally vary a great deal from study to study, and the observed  $P$ -value against the target hypothesis of no difference between the two studies is  $p = .289$ .<sup>3</sup>

Finally, as a consequence of the incorrect belief that the results are incompatible, we expect implications for perceptions of replication. Specifically, if the study with  $p = .005$  had been conducted prior to the study with  $p = .194$ , we expect that many researchers would wrongly view the latter to be a “failed replication” and the former to be a “false positive” despite the fact that the results are highly compatible and provide cumulative evidence.

### Proposal

Perhaps the most widespread abuse of statistics is to take where some statistical measure such as a  $P$ -value stands relative to some threshold as a basis to declare “statistical (non)significance” and to make general and certain conclusions from a single study. However, single studies are never definitive. Therefore, single studies can never demonstrate an effect or no effect. Likewise, single replication studies can never demonstrate failed replication or false positive.

Instead, the aim of studies should be to report results in an unfiltered manner so that they can later be used to make more general conclusions based on the cumulative evidence from multiple studies. Nonetheless, NHST leads researchers to wrongly make general and certain conclusions and to wrongly filter results.

Therefore, we propose a major transition in statistical analysis and reporting. Specifically, we propose moving beyond binary: abandoning NHST—and the  $P$ -value thresholds intrinsic to it—as the default approach to statistical analysis and reporting. “Statistical (non)significance” should never be used as a basis to make general and certain conclusions or as a filter to select which results to publish. Instead, all studies should be published in some form or another and reporting should focus on quantifying study results via point and interval estimates. Further, general conclusions should be made based on the cumulative evidence from multiple studies. This should be done in a manner that treats  $P$ -values continuously and as just one factor among many—including prior evidence,

<sup>2</sup> The sheer breadth of literature on this topic across time and fields makes an exhaustive review intractable. For some examples, see Rozeboom (1960), Edwards, Lindman, and Savage (1963), Bakan (1966), Morrison and Henkel (1970), Meehl (1978), Rothman (1978), Salsburg (1985), Gardner and Altman (1986), Rothman (1986), Serlin and Lapsley (1993), Cohen (1994), McCloskey and Ziliak (1996), Schmidt (1996), Hunter (1997), Gill (1999), Anderson, Burnham, and Thompson (2000), Gigerenzer (2004), Hubbard (2004), Gigerenzer, Krauss, and Vitouch (2004), Briggs (2016), McShane and Gal (2016), Wasserstein and Lazar (2016), McShane and Gal (2017), Amrhein, Greenland, and McShane (2019a), Amrhein, Trafimow, and Greenland (2019), McShane et al. (2019), and Wasserstein, Schirm, and Lazar (2019).

<sup>3</sup> The observed  $P$ -values of  $p = .005$  and  $p = .194$  in the two studies correspond to observed  $Z$ -statistics of  $z = 2.80$  and  $z = 1.30$ , respectively, and thus an observed  $Z$ -statistic of  $z = (2.80 - 1.30)/\sqrt{1^2 + 1^2} = 1.06$  and an observed  $P$ -value of  $p = .289$  against the target hypothesis of no difference between the two studies.

plausibility of mechanism, study design, data quality, and others that vary by research domain—that require joint consideration and holistic integration. It should also be done in a manner that respects the fact that such conclusions are necessarily tentative and subject to revision as new studies are conducted.

While our proposal may seem radical to some, we view it as neither controversial nor novel. Indeed, in response to long-standing criticism of NHST, similar proposals have long been made. Despite this, NHST has till recently seemed unassailable. However, due to “highly visible discussions” critical of NHST in the science press and “deep concern about issues of *reproducibility* and *replicability* of scientific conclusions,” the Board of Directors of the American Statistical Association (ASA) took the unprecedented step in 2016 of issuing a statement warning against the misuse of “statistical significance” and *P*-values (Wasserstein and Lazar 2016, p. 129). The ASA built on this statement with two major efforts devoted to improving statistical practice in science, namely the 2017 ASA Symposium on Statistical Inference and a 2019 special issue of *The American Statistician* (Wasserstein, Schirm, and Lazar 2019). Since then, researchers throughout the biomedical and social sciences have responded by publishing editorials and articles making proposals akin to ours as well as altering journal guidelines for statistical analysis and reporting so as to adhere to such proposals.<sup>4</sup> In our view, it is time for marketing to take similar steps.

Therefore, in the remainder of this article, we briefly review some of the principal problems associated with NHST. We next discuss some principles that we believe should underlie statistical analysis and reporting. We then use these principles to motivate some guidelines for statistical analysis and reporting that could be used for manuscripts under consideration for publication in the *Journal of Marketing* and journals more broadly. We next provide some examples that illustrate statistical analysis and reporting that adheres to our principles and guidelines. We conclude with a brief discussion.

In addition, we provide a brief Appendix that illustrates the degree to which (1) NHST is employed, (2) problems associated with it are fallen prey to, and (3) our guidelines for statistical analysis and reporting are adhered to in marketing. Put simply, NHST is dominant, problems associated with it are

rife, and adherence to our guidelines ranges from nil to partial in papers recently published by prominent academics in marketing. Therefore, there is great opportunity for marketing to catch up to biomedical and social sciences that have put proposals akin to ours into practice.

Finally, we provide scripts that reproduce all results and figures discussed in our examples at <https://www.blakemcshane.com/jm.statsig.zip>.

In considering this article, it is important to recognize that we believe that no single statistical approach is suitable for all research questions, and thus we advocate a “toolkit” approach that chooses the best one for the job at hand (Cox and Donnelly 2011; Efron and Hastie 2016; Gigerenzer 2004). Nonetheless, we believe that it is always inappropriate to use *P*-values and related statistical measures (such as the limits of interval estimates, likelihood ratios, posterior probabilities, and Bayes factors) in the conventional, dichotomous manner, that is, to declare “statistical (non)significance” and decide whether a result proves or disproves a scientific hypothesis based on where the value stands relative to some threshold. Again, scientific hypotheses should be evaluated based on the cumulative evidence from multiple studies in a manner that treats *P*-values and related statistical measures continuously and as just one factor among many. In doing so, the distinction between scientific hypotheses (which are *not* vague directionless or directional claims of an effect or claims of no effect) and statistical hypotheses should be firmly borne in mind (Meehl 1978, 1990).

## Problems Associated with Null Hypothesis Significance Testing

### Target Hypothesis and Background Assumptions

The target hypothesis that is employed in the overwhelming majority of applications—namely, the null hypothesis of no association or no effect—is always false in marketing and the biomedical and social sciences more broadly (Bakan 1966; Berkson 1938; Cohen 1994; Edwards, Lindman, and Savage 1963; Meehl 1967; Tukey 1991). So too are one or more of the countless other explicit, implicit, and overlooked background assumptions that vary by context (e.g., additivity, linearity, normality, random sampling of subjects from a population, random assignment of subjects to experimental conditions, properly functioning measurement devices, absence of data fabrication). Consequently, false positive rates are not relevant (i.e., because they are zero) and small *P*-values are to be expected, at least with sufficient data.

### Thresholds and Dichotomization

The conventional .05 threshold—or, for that matter, any other threshold—used to dichotomize results into the categories “statistically significant” and “statistically nonsignificant” is arbitrary (Cochran 1976; Cowles and Davis 1982; Fisher 1926, 1956; Pearson 1935). Moreover, dichotomization itself has “no ontological basis” because “there is no sharp line between

<sup>4</sup> Again, an exhaustive review is intractable. For some examples, see Amrhein, Greenland, and McShane (2019a), Bernard (2019), Bijak (2019), Bresee (2019), Curran-Everett (2019), Davidson (2019), De Koning and Noordhof (2019), Dirnagl (2019), Harrington et al. (2019), Harvey and Brinkhof (2019), Hayat et al. (2019), Lowe (2019), Marshall (2019), McShane et al. (2019), Morken (2019), Nguyen, Rivadeneira, and Civitelli (2019), Pickler (2019), O’Connor (2019), Parsons et al. (2019), Staggs (2019), Carlsson and Gönen (2020), Charlesworth and Pandit (2020), Curran-Everett (2020), Johnson et al. (2020), Knottnerus and Tugwell (2020), Marshall and Hughes (2020), Maula and Stam (2020), Michel, Murphy, and Motulsky (2020), Price, Bethune, and Massey (2020), Santibáñez, García-Rivero, and Barreiro (2020), Van Witteloostuijn (2020), Heckelei et al. (2021), Imbens (2021), Putt (2021), Robinson and Haviland (2021), Tijssen (2021), Amrhein and Greenland (2022), Butler (2022), Elkins et al. (2022), Filippini and Vinceti (2022), Greenland, Mansournia, and Joffe (2022), Bonovas and Piovani (2023), Fingerhut (2023), Hassler (2023), Montero, Hedeland, and Balgoma (2023), and Verykoui and Nakas (2023).

### Exhibit 1: *P*-Values, *S*-Values, Interval Estimates, and Compatibility

To appreciate the compatibility assessments offered by the *P*-value and forestall misinterpretations of it, it is useful to consider the (binary) *S*-value, also known as the (binary) Shannon information or surprisal (Greenland 2019; Rafi and Greenland 2020; Shannon 1948). The *S*-value is a simple cognitive device for appreciating the evidence provided by a *P*-value that is measured in bits (binary digits). Specifically, the observed *S*-value *s* corresponding to an observed *P*-value *p* is  $s = \log_2(1/p) = -\log_2(p)$ .

To illustrate the *S*-value, consider an even-money bet on tails on a single coin toss. Before placing such a bet, one would want to obtain evidence that the even-money terms are acceptable, that is, that the coin toss is not biased in favor of heads. To do so, suppose one tosses the coin *s* independent times and all tosses come up heads. If the terms are acceptable, the probability of this occurring is at most  $1/2^s$ , that is, the probability of all heads in *s* independent fair (unbiased) coin tosses. The smaller this probability, the less one should trust that the terms are acceptable.

Indeed, *s* can be thought of as a measure of evidence against acceptability. If one tossed only once ( $s = 1$ ) and the toss came up heads, this would not be surprising even if the coin toss were fair (because this has probability  $1/2$ ), and so it would provide almost no evidence against acceptability. If one tossed ten times ( $s = 10$ ) and all ten tosses came up heads, this would be very surprising if the coin tosses were fair (because this has probability  $1/2^{10} \approx .001$ ), and so it would provide considerable evidence against acceptability.

With this in mind, an observed *P*-value *p* can be converted to the number *s* of heads in a row closest to *p* by solving  $p = 1/2^s$  for *s*, which yields  $s = \log_2(1/p) = -\log_2(p)$  as the observed *S*-value.

As an example, an observed *P*-value of  $p = .05$  converts to an observed *S*-value of  $s = -\log_2(.05) = 4.32$  bits, which rounds to 4. Therefore, an observed *P*-value of  $p = .05$  is about as surprising as seeing all heads in four independent fair coin tosses (the .32 represents a negligible amount of evidence, less than that of a third of a coin toss) assuming that both the target hypothesis and the background assumptions are true and as assessed by the test statistic. Seeing all heads in four independent fair coin tosses is not all that unexpected in the intuitive sense that if one conducted many “studies,” each consisting of four independent fair coin tosses, one would not be surprised to see such studies yield four heads every now and again (because this has probability  $1/2^4 = .0625$ ). Put differently, such studies yielding four heads every now and again is reasonably compatible with our intuitive expectations about the data such studies would yield.

One can express this intuition by saying that an observed *P*-value of  $p > .05$  (which corresponds to an observed *S*-value of  $s < 4.32$  bits) indicates an at least reasonable degree of compatibility of the observed data with both the target hypothesis and the background assumptions as assessed by the test statistic.

Equivalently, because the observed *P*-value against the target hypothesis of any value inside a  $(1 - \alpha) \times 100\%$  interval estimate has  $p > \alpha$  (e.g., any value inside a 95% interval estimate has  $p > .05$ ), one can say that every value inside a 95% interval estimate is at least reasonably compatible with the data given all of the assumptions used to compute it in the following sense: for every value inside the interval, the data are about as surprising as or less surprising than seeing all heads in four independent fair coin tosses.

a ‘significant’ and a ‘nonsignificant’ difference; significance in statistics ... varies continuously between extremes” (Rosnow and Rosenthal 1989, p. 1277). This is further compounded by the fact that *P*-values naturally vary a great deal from study to study. Indeed, sampling variation alone can easily cause *large* differences in *P*-values—not only *P*-values that fall just barely to either side of some threshold (Gelman and Stern 2006; Goodman 1992; Greenland 2019; Senn 2002).

### Misinterpretation of the *P*-Value

The *P*-value is formally defined as the probability that some test statistic (e.g., a *Z*-statistic, a  $\chi^2$ -statistic, or another quantity computed from the data) would be as extreme or more extreme than the value of the test statistic computed from the observed data, assuming that both the target hypothesis and the background assumptions were true. As such, the *P*-value is a measure of the degree of the compatibility of the observed data with both the target hypothesis and the background

assumptions as assessed by the test statistic, with an observed *P*-value of  $p = 0$  indicating complete incompatibility (i.e., the observed test statistic and thus the data are impossible given the target hypothesis and background assumptions) and an observed *P*-value of  $p = 1$  indicating no *detected* incompatibility (i.e., there is no discrepancy between the observed test statistic and the target hypothesis and background assumptions; for elaboration on the compatibility assessments offered by the *P*-value, see Exhibit 1).<sup>5</sup> However, researchers commonly

<sup>5</sup> The conception of the *P*-value as a measure of compatibility is a venerable one in statistics: it is anticipated in Pearson (1900, pp. 170–71) and Fisher (1934, p. 66; 1935a, p. 207); is found in Box (1980), Bayarri and Berger (2000), Robins, Van der Vaart, and Ventura (2000), and Bayarri and Berger (2004); and is found under the term “goodness-of-fit” in Pearson (1900), “consonance” in Kempthorne and Folks (1971), Kempthorne (1976), and Folks (1981), and “consistency” in Cox and Hinkley (1974) and Cox (1977). See Greenland (2023b) for more on this and an alternative conception of the *P*-value, and see Gigerenzer (2004) for how NHST as practiced reflects neither conception.

misinterpret the  $P$ -value as, among other things, the probability that the target hypothesis is true, one minus the probability that some alternative hypothesis is true, and one minus the probability of replication (Gigerenzer 2018; Goodman 2008; Greenland et al. 2016; Oakes 1986).

### *Errors of Reasoning*

Dichotomization of results into the categories “statistically significant” and “statistically nonsignificant” leads researchers to wrongly interpret results that attain “statistical significance” as demonstrating an effect and those that fail to do so as demonstrating no effect (McShane and Gal 2016, 2017). In addition, researchers wrongly believe that “statistical significance” indicates practical importance (Boring 1919; Freeman 1993) and provides evidence that associations are causal (Holman et al. 2001). However, a small  $P$ -value suggests only that one or more of the assumptions used to compute it *may* be false without indicating which assumption(s), if any, are false (Greenland 2017). Therefore, it is wrong to conclude on this basis alone that the target (e.g., null) assumption is false or that any other specific assumption is false. Indeed, small  $P$ -values will be observed even when all of the assumptions are true. In addition, a large  $P$ -value suggests only that a false assumption was not detected—perhaps because all of the assumptions are true (untenable in marketing and the biomedical and social sciences more broadly), the  $P$ -value is insensitive to the false assumption(s), or the false assumption(s) themselves or in combination with sampling variation largely cancel one another out (Greenland 2017). Therefore, it is wrong to conclude on this basis alone that the target (e.g., null) assumption is true—an error lamented for over a century (Altman and Bland 1995; Fisher 1935b; Pearson 1906)—or that any other specific assumption is true. Indeed, large  $P$ -values may be observed even when one or more of the assumptions are false.

### *Biased Literature and Harmful Research Practices*

The incorrect beliefs that “statistical significance” demonstrates an effect and that “statistical nonsignificance” demonstrates no effect leads researchers, when acting as authors, editors, and reviewers, to use “statistical (non)significance” as a filter to select which results to publish. However, because “statistically significant” estimates are biased upward in magnitude and “statistically nonsignificant” estimates are biased downward in magnitude, this practice biases the literature (Gelman and Carlin 2014; Lane and Dunlap 1978; McShane, Böckenholt, and Hansen 2016). Further, it encourages harmful research practices that yield “statistical significance” for some desired result or “statistical nonsignificance” for some undesired result (Brodeur et al. 2016; Head et al. 2015; John, Loewenstein, and Prelec 2012; Masicampo and Lalande 2012). These issues are compounded when researchers engage in multiple comparisons—both potential and actual (Gelman and Loken 2014; see also Gelman and Loken 2013).

## **Principles for Statistical Analysis and Reporting**

### *Embrace Uncertainty*

Statistical analysis is often wrongly viewed as a kind of “alchemy” that can eliminate uncertainty and variation and thereby allow for dichotomous declarations of truth or falsity based on some  $P$ -value or related statistical threshold being crossed (Gelman 2016). Instead, the purpose of statistical analysis is to quantify uncertainty and variation (albeit in a very narrow and particular manner), and all results—even those from the most rigorous studies—are highly uncertain and variable (Amrhein, Trafimow, and Greenland 2019). Therefore, dichotomous declarations (e.g., of an effect or no effect) offer only false certainty.

### *Embrace Cumulation*

Because single studies are never definitive, the aim of studies should be to report results in an unfiltered manner so that they can later be used to make more general conclusions based on the cumulative evidence from multiple studies (Amrhein, Greenland, and McShane 2019b). Consequently, replication studies as well as meta-analyses that integrate such studies are critical. However, replication (and thus meta-analysis) is complicated in marketing and the biomedical and social sciences more broadly because studies of a given phenomenon can never be direct or exact replications of one another (Fabrigar and Wegener 2016; Greenland and O’Rourke 2008; McShane and Böckenholt 2014; Rosenthal 1990). Instead, studies differ at minimum with regard to their method factors.<sup>6</sup> Therefore, (1) studies of a given phenomenon should always be considered general (i.e., systematic or conceptual) replications and (2) the quantification of the variation (or heterogeneity) across studies as well as the identification of moderators of this variation—rather than the estimation of potentially fictitious “average” effects—are typically the most important purposes of meta-analysis (Greenland 1987; Light and Pillemer 1984; Pearson 1904).

### *Embrace Judgment*

Statistical inference is only a small part of scientific inference. Consequently,  $P$ -values and related statistical measures should not be given priority status. Instead, they should be treated as just one factor among many—including prior evidence, plausibility of mechanism, study design, data quality, and others that

<sup>6</sup> Method factors are any known or unknown factors that pertain to the implementation of a study but that are not directly related to the theory under investigation. They include factors that may seem major in some contexts, such as the operationalization of the dependent measure(s) and the operationalization of the experimental manipulation(s), as well as factors that may seem minor in some contexts, such as the subject pool and the time of day (for a comprehensive list, see Brown et al. 2014). Whether a given factor is deemed a method factor or a factor directly related to the theory under investigation will depend on the perspective of the researcher (McShane, Böckenholt, and Hansen 2022).

vary by research domain—that require joint consideration and holistic integration (McShane et al. 2019). While this requires careful thought and judgment and involves subjectivity, there is subjectivity at all stages of scientific inquiry, even if objectivity remains the ultimate goal (Lykken 1968). Indeed, *P*-values themselves are subjective in the sense that they are affected by the many necessarily subjective choices involved in study design and statistical analysis.

### Embrace Exploration

While the dividing line between exploratory and confirmatory research is seldom sharp, most research falls far to the more exploratory side. Given this, flexibility in statistical analysis is valuable and indeed necessary. While flexibility implies that the *P*-value no longer retains its exact meaning, this is, for a variety of reasons, a small price to pay. For example, the *P*-value is *never* guaranteed to retain its exact meaning—even in preregistered studies. In addition, recall that the *P*-value is a measure of the degree of the compatibility of the observed data with both the target hypothesis and the background assumptions as assessed by the test statistic; although it is a very good and useful measure of this, such a measure is quite narrow and seldom of interest in marketing and the biomedical and social sciences more broadly. Recognizing the value and pervasiveness of exploratory research promotes learning from data. So too does recognizing that preregistration and other practices that may inhibit exploratory research are unequivocally merely rearguard measures against a symptom of NHST rather than the requisite assault on it.<sup>7</sup>

### Embrace Transparency

Transparency about research practices is necessary: describe relevant statistical analyses performed, present results without regard to “statistical (non)significance,” and share materials, data, and code. Although transparency alone is not sufficient to ensure reliable research (Gelman 2017), it is difficult, if not impossible, to assess research results in the absence of transparency. Further, transparency helps calibrate expectations about the results of future studies. Therefore, editors and reviewers should stop expecting, let alone demanding, studies free of “imperfections” and be open to results-

blind review as appropriate. All studies should be published in some form or another, as a biased literature results otherwise; again, the aim of studies should be to report results in an unfiltered manner so that they can later be used to make more general conclusions based on the cumulative evidence from multiple studies (Amrhein, Greenland, and McShane 2019b).

## Guidelines for Statistical Analysis and Reporting

### Report Point and Interval Estimates

The quantification of study results is a key component of scientific inquiry. Because *P*-values and related statistical measures do not quantify study results, it is critical to report point and interval estimates which do. When possible, this should be done in meaningful units rather than unitless standardized ones (Greenland, Schlesselman, and Criqui 1986; McShane and Böckenholt 2022; Tukey 1969; Wilkinson 1999). In doing so, it is important to be mindful of five things (Amrhein, Greenland, and McShane 2019a):

1. Every value inside an interval estimate at any conventional level, such as 95%, is at least reasonably compatible (i.e., in the sense elaborated on in Exhibit 1) with the data given all of the assumptions used to compute it; therefore, it makes no sense to single out a specific value such as the null value.
2. Not every value inside an interval estimate is equally compatible: the point estimate is most compatible and values near it are more compatible than those far from it.
3. Values outside an interval estimate are not strictly incompatible (except values that are ruled out based on logic, physics, or assumptions such as temperatures below 0 K) but rather are just less compatible than those inside.
4. Not every value outside an interval estimate is equally (in)compatible: values near the limits are more compatible than those far from them and those sufficiently far may be considered highly incompatible or even for all intents and purposes strictly incompatible depending on context.
5. An interval estimate understates the true degree of uncertainty—typically woefully so—because the compatibility assessments it (and related statistical measures such as a *P*-value, likelihood ratio, posterior probability, and Bayes factor) offer depend on the correctness of all of the assumptions employed and these assumptions are typically far from given (Greenland 2023a); therefore, make them as clear as possible, check those that can be checked (e.g., by plotting data and model estimates, estimating alternative models, ensuring that randomization mechanisms followed protocol and measurement devices properly functioned), and recognize both that many cannot be checked and that many are implicit or overlooked.

<sup>7</sup> Gelman (2020) notes that “The goal should be to learn, not to test hypotheses, and the false positive probability [i.e., rate] has nothing to do with anything relevant ... hypothesis testing is not actually rigorous, it’s just a way to add noise to data. Anyway, none of this is really an issue [when] sharing [the] raw data. That’s really all the preregistration you need.” To explain, note that the truth value of a claim such as a scientific hypothesis does not depend on when the claim was made or conceived. Similarly, the reasonableness or quality of a statistical analysis and thus the results it yields does not depend on when it was performed or conceived. Therefore, because sharing data allows readers to reproduce the statistical analyses originally performed (e.g., to check assumptions, to check that they yield the results originally reported) and to perform additional ones, this is what actually matters—not whether or when someone wrote down or made public (i.e., registered) that they would perform some statistical analysis—thereby making preregistration unnecessary.

Finally, when it is deemed necessary or desirable to discuss the practical importance of study results (e.g., to argue that they necessitate further research or are “null” in the sense of being practically unimportant), discuss the practical importance of not only point estimates but also (at minimum) both the lower and upper limits of interval estimates. In doing so, be mindful that interpretations of practical importance are context specific, others may have a different interpretation, and diversity in interpretation is not problematic (Poole 1987).

### Report Interval Estimates at Multiple Levels

The 95% level is, like the .05 threshold from which it came, an arbitrary convention. Therefore, different and even multiple levels are justified in different applications. In fact, it is more accurate and complete to report interval estimates at multiple levels. Therefore, at least for focal estimands, report interval estimates at multiple levels, perhaps by plotting the interval estimate for all levels from 0% to 100% (Birnbaum 1961; Cox 1958; Poole 1987; Sullivan and Foster 1990). In doing so, it is important to be mindful that an interval estimate at a given level—like a  $P$ -value of a given value—assesses only the compatibility of the values inside it with the data given all of the assumptions used to compute it and not the probability or plausibility of those values given the data. This distinction is not subtle: compatibility is a much weaker condition than probability or plausibility. Consider, for example, that unknown errors in data collection and data management as well as intentional alteration or even complete fabrication of the data are always explanations that are compatible with the data (as has indeed occurred in some influential studies in marketing and the biomedical and social sciences more broadly), even when those explanations seem improbable or implausible (Rafi and Greenland 2020).

### When Reporting $P$ -Values, Report Them Continuously and for Relevant Nonnull Values

“ $P$ -values, with the additional information they provide, are typically more appropriate than fixed levels [i.e., thresholds] in scientific problems” (Lehmann 1986, p. 71; Lehmann and Romano 2005, p. 65). Therefore, when it is deemed necessary or desirable to report  $P$ -values, report them continuously and to sensible precision (often two and seldom more than three digits) and never as binary inequalities (e.g.,  $p < .05$  or  $p > .05$ ).<sup>8</sup> Because readers equipped with this information can use it as they see fit, avoid reporting that results are “statistically significant” or “statistically nonsignificant,” let alone that they are “marginally statistically significant” or “approaching statistical

significance” or some other such phrase (Wasserstein, Schirm, and Lazar 2019). Similarly, avoid asterisks or other adornments that signify thresholds. Further, and related to the recommendation to report interval estimates at multiple levels at least for focal estimands, when reporting the  $P$ -value against the target (null) hypothesis of no effect for such quantities, also report the  $P$ -value against at least one target hypothesis of some relevant nonnull value of the effect or even plot the  $P$ -value for a range of values of the effect (Greenland 2016; Poole 1987; Rafi and Greenland 2020). Finally, heed these recommendations *mutatis mutandis* when reporting statistical measures related to  $P$ -values such as likelihood ratios, posterior probabilities, and Bayes factors.

### Report the Rationale for the Sample Size

Editors and reviewers typically seek some rationale for the sample size of a study, often desiring one based on an a priori power calculation. Such calculations are not possible because they require knowledge of the true effect in the study (McShane and Böckenholt 2016), which is always unknown (not only before but also after the study is conducted), and which, if known, would make conducting the study unnecessary. Further, post hoc assessments of power, such as the observed power of a single study or the average power from a meta-analysis of multiple studies, are deeply problematic (e.g., they are irrelevant and typically are biased and have large sampling variation) and thus should not be calculated or reported (Gelman 2019a, b; Greenland 2012; Hoenig and Heisey 2001; McShane, Böckenholt, and Hansen 2020; Yuan and Maxwell 2005). Instead, simply report whatever was the rationale for the sample size, for example, that it was based on that used in prior studies, the size of the customer base of the firm that provided the data (or the size that the firm was willing to provide), the largest possible given resource constraints, or chosen to achieve a given level of the precision for some estimate (i.e., the accuracy in parameter estimation approach; Kelley, Maxwell, and Rausch 2003).

### Eschew Decisions

Many researchers believe both that they need to make binary decisions and that NHST provides a rigorous framework for doing so. However, decisions are seldom necessary in scientific reporting (Rozeboom 1960). Instead, they are best left for end users such as managers and clinicians. Further, many perceived binary decisions are in fact continuous (e.g., a decision about whether or not to invest is better characterized as a decision about how much to invest). Finally, when decisions (binary or otherwise) are required, they should be made using a decision analysis that integrates the costs, benefits, and probabilities of all possible consequences via a loss function (which typically varies dramatically across stakeholders)—not via arbitrary thresholds applied to statistical summaries such as  $P$ -values, which, outside of certain specialized applications such as industrial quality control, are insufficient for this purpose (McShane and Gelman 2022).

<sup>8</sup> Because researchers commonly misinterpret  $P$ -values, when reporting  $P$ -values, consider converting them to  $S$ -values and reporting  $S$ -values in addition to or instead of  $P$ -values in order to help preclude misinterpretation (for elaboration, see Exhibit 1).

## Examples of Statistical Analysis and Reporting

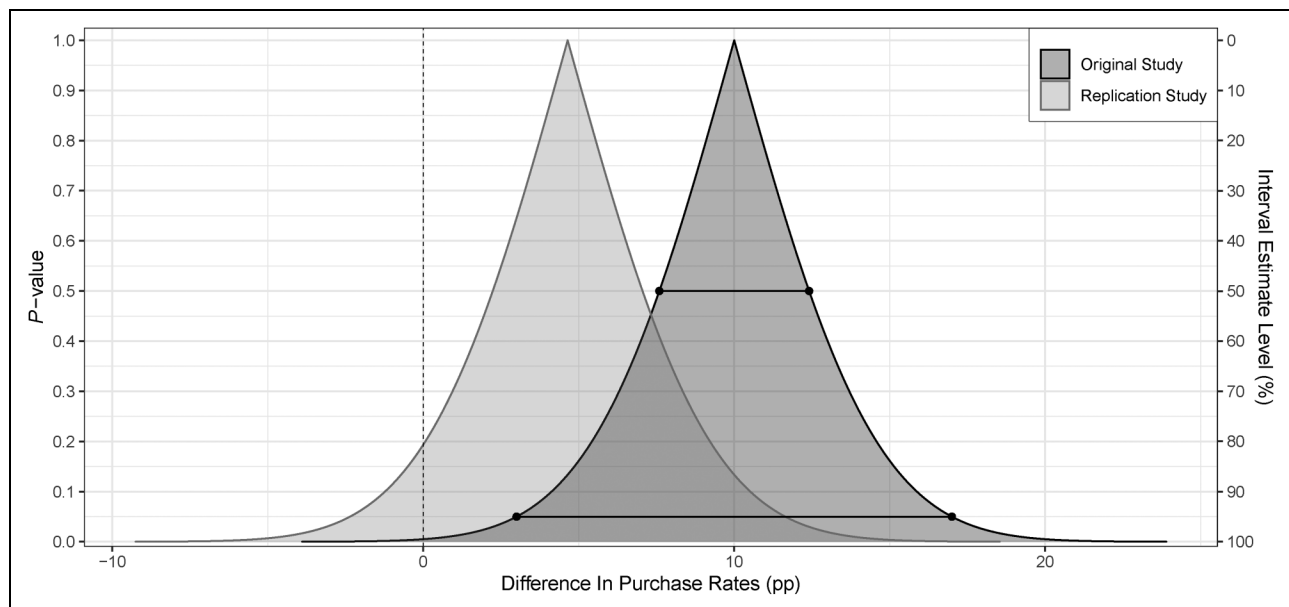
### Reporting a Single Study

Consider a hypothetical study of choice overload in which a researcher conjectures that an increase in the number of options from which to choose will result in a decrease in the likelihood of making a choice. In the study, subjects were randomly assigned to either a small or large choice set and then given the option to purchase an item. Traditionally, a researcher would compute a  $P$ -value against the target hypothesis of no difference in the purchase rates of the two choice set conditions, viewing one less than .05 as demonstrating a choice overload effect and one greater than .05 as demonstrating no choice overload effect. However, this dichotomy offers only false certainty, and the quantification of study results is a key component of scientific inquiry. Therefore, we recommend focusing on quantifying the study results by discussing the purchase rates in each choice set condition and their difference and the degree to which multiple interval estimate levels are compatible with the data, noting here that a wide range of values of the difference are more compatible than the null value of zero. For example:

In our study of choice overload, 55.0% of subjects in the small choice set condition purchased, as compared to 45.0% of subjects in the large choice set condition—a decrease of 10.0 percentage points (pp) in the purchase rate. A 95% interval estimate suggests that every value from a 3.0 pp decrease to a 17.0 pp decrease in the purchase rate is at least reasonably compatible with our data given all

of the assumptions used to compute it; a 50% interval estimate suggests that every value from a 7.6 pp decrease to a 12.4 pp decrease is highly compatible. For the interested reader, we note that these estimates correspond to an observed  $Z$ -statistic of  $z = 2.80$  and an observed  $P$ -value of  $p = .005$  against the target hypothesis of no difference in the purchase rates of the two choice set conditions using the standard two-sample  $Z$ -test for proportions.

In Figure 1, we plot in black the interval estimate for all levels from 0% to 100% using the *concurve* package (Rafi and Vigotsky 2020). The peak of the curve denotes the point estimate of a 10.0 pp decrease in the purchase rate. This may be thought of as a 0% interval estimate as indicated by the right y-axis of the figure. Equivalently, the observed  $P$ -value against the target hypothesis of a 10.0 pp decrease in the purchase rate is  $p = 1$  as indicated by the left y-axis of the figure. Each horizontal slice of the curve, such as the ones at 50% and 95% depicted in the figure, can be interpreted analogously. For example, consider the latter. The limits of a 3.0 pp decrease and a 17.0 pp decrease in the purchase rate provide a 95% interval estimate, as indicated by the right y-axis of the figure. Equivalently, the observed  $P$ -value against the target hypotheses of either a 3.0 pp decrease or a 17.0 pp decrease in the purchase rate is  $p = .05$ , as indicated by the left y-axis of the figure. The figure illustrates that it is misleading to frame the discussion of the study in terms of whether the null value is inside or outside the 95% interval estimate of the difference in the purchase rates or whether the observed  $P$ -value  $p$  is above or below .05: every value from 0 pp to a 20.0 pp decrease is more compatible with the data than the null value of zero.



**Figure 1.** Interval Estimate Curve.

Notes: The curve plots the interval estimate for each study for the level indicated by the right y-axis. Equivalently, it plots the  $P$ -value for each study against the target hypothesis of the value of the difference in purchase rates indicated by the x-axis. The 50% and 95% interval estimates for the original study discussed in the main text are depicted in the figure.



Therefore, we caution that values outside the 95% interval estimate limits of a 3.0 pp decrease and a 17.0 pp decrease are not incompatible with our data, but rather they are—as the figure illustrates—just less compatible than those inside. In closing, we note that our interval estimates are optimistically narrow because they are conditional on assumptions that almost certainly do not hold exactly. Moreover, the interval estimates indicate only the compatibility of the values inside it with the data given all of the assumptions used to compute it and not the probability or plausibility of those values given the data. Finally, we have conducted only a single study, and many more studies are needed to arrive at general conclusions.

### Reporting a Replication Study

Consider a hypothetical replication of the study of choice overload discussed in the prior example. In the replication study, the estimate of the difference in the purchase rates is smaller than in the original study (4.6 pp vs. 10.0 pp) and the observed  $P$ -value is larger ( $p = .194$  vs.  $p = .005$ ). Traditionally, a researcher would view the replication study to be a failed replication and the original study to be a false positive. However, this dichotomy offers only false certainty, and false positive rates are not relevant. Indeed,  $P$ -values naturally vary a great deal from study to study, and  $p = .005$  in the original study and  $p = .194$  in the replication study are highly compatible (i.e., because the observed  $P$ -value against the target hypothesis of no difference in the choice overload effect of the two studies is  $p = .289$ ). Therefore, we recommend integrating the results across the two studies and discussing similarities and differences between them, noting here that the difference in the purchase rates in the two studies are quite similar and more studies are needed. For example:

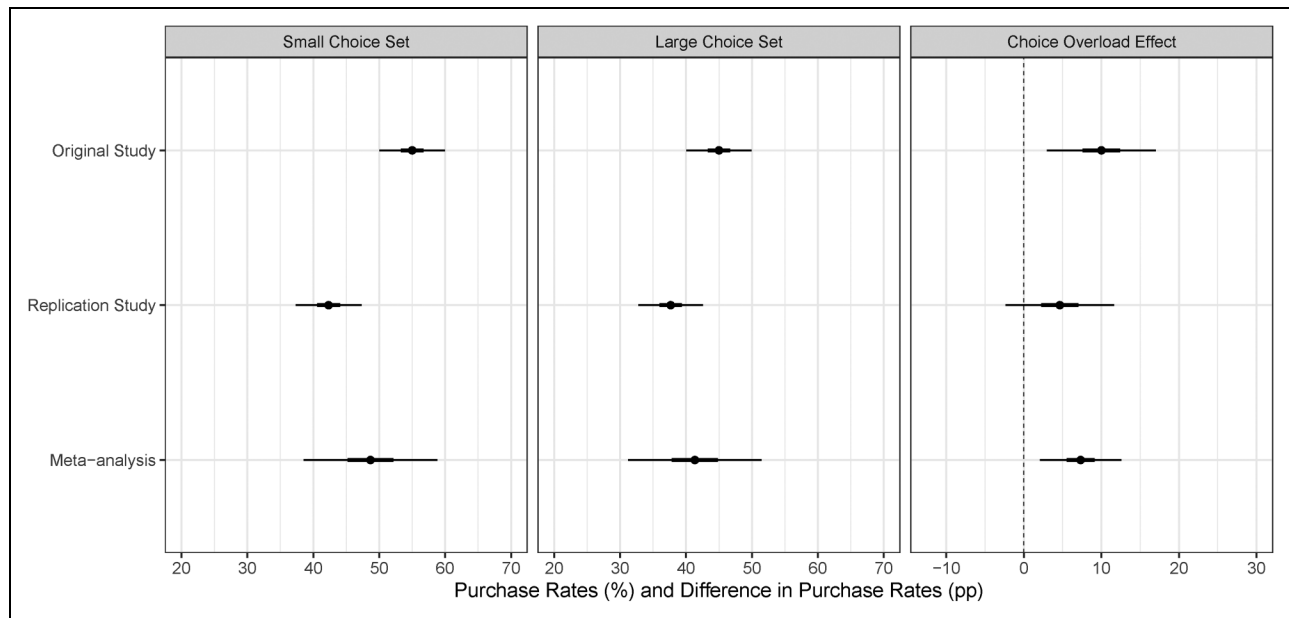
Not satisfied with our single study of choice overload, we replicated our study using the same materials but using Amazon Mechanical Turk (MTurk) workers as subjects rather than university students. In our replication study, 42.3% of subjects in the small choice set condition purchased, as compared to 37.7% of subjects in the large choice set condition—a decrease of 4.6 pp in the purchase rate. A 95% interval estimate suggests that every value from a 2.4 pp increase to a 11.6 pp decrease in the purchase rate is at least reasonably compatible with our data given all of the assumptions used to compute it. For the interested reader, we note that these estimates correspond to an observed  $Z$ -statistic of  $z = 1.30$  and an observed  $P$ -value of  $p = .194$  against the same target hypothesis and using the same test as in the original study.

In Figure 1, we plot in gray the interval estimate for all levels from 0% to 100%. The figure illustrates several important facts about this replication study considered both alone and in tandem with the original study. First, it illustrates that it is misleading to frame the discussion of the replication study in terms of whether the null value is inside or

outside the 95% interval estimate of the difference in the purchase rates or whether the observed  $P$ -value  $p$  is above or below .05: every value from 0 pp to a 9.3 pp decrease is more compatible with the data than the null value of zero. Second, it illustrates that it is misleading to frame the discussion of the pair of studies in those same dichotomous terms: the fact that the gray curve and black curve overlap to a large degree indicates that the estimates from the two studies are highly compatible with one another. Third, it illustrates that the point estimate of the difference in the purchase rates in the replication study is smaller than that in the original study: the peak of the gray curve is to the left of the peak of the black curve. Fourth, it illustrates that the two studies have similar precision: the horizontal slices of the gray curve and the black curve at each level are of similar width. We integrated the two studies using the meta-analytic methodology of McShane and Böckenholt (2017). In Figure 2, we plot point and 50% and 95% interval estimates of the purchase rates in each condition of and the choice overload effect in each study as well as the meta-analytic average. The meta-analysis yields a point estimate of the average choice overload effect of a 7.3 pp decrease and a 95% interval estimate ranging from a 2.1 pp decrease to a 12.6 pp decrease. For the interested reader, we note that these estimates correspond to an observed  $Z$ -statistic of  $z = 2.74$  and an observed  $P$ -value of  $p = .006$  against the target hypothesis of no average difference in the two choice set conditions using the Wald test.

The point estimate of the variation (or heterogeneity) in the choice overload effect across the two studies is 1.2 pp, thus reflecting the similarity of the 10.0 pp difference in purchase rates in the original study and the 4.6 pp difference in the replication study on display in the right panel of the figure. The point estimate of the variation in the level of the purchase rates across the two studies was estimated to be more sizable at 6.8 pp, thus reflecting the difference between the relatively higher 55.0% and 45.0% purchase rates in the original study and the relatively lower 42.3% and 37.7% purchase rates in the replication study on display in the left and middle panels of the figure. The 95% interval estimates of variation are, being based on only two studies, unsurprisingly quite wide, ranging from 0 pp to 29.4 pp and from 0 pp to 79.1 pp, respectively.

We caution that our meta-analysis includes only two studies and many more are needed to arrive at general conclusions. In addition, the estimates of variation are extremely limited because the same materials were used in both studies, and therefore they primarily reflect differences in sample populations, namely university students in the original study and MTurk workers in the replication study. More studies featuring greater variation in method factors are needed. Nonetheless, these two studies can and should be added to extant meta-analyses of choice overload such as Chernev, Böckenholt, and Goodman (2015) and McShane and Böckenholt (2018).



**Figure 2.** Single Study and Meta-Analytic Estimates.

Notes: Point estimates are given by the points; 50% and 95% interval estimates are given by the thick and thin lines, respectively.

### Reporting a Study with Variation in Effects

Consider a hypothetical study of an advertising intervention across 100 brands. Traditionally, a researcher would focus on the average effect across brands, computing a  $P$ -value against the target hypothesis of no average effect of the intervention and ignoring the variation (or heterogeneity) in effects. However, effects vary (in this example, across brands, but more generally as a function of method factors). Therefore, we recommend focusing on the variation in effects, noting here that the magnitude of the variation makes the results of limited meaning and interest and necessitates further research (and that this would remain the case even if the point and interval estimates of the brands were all above—or for that matter below—the null value of zero). For example:

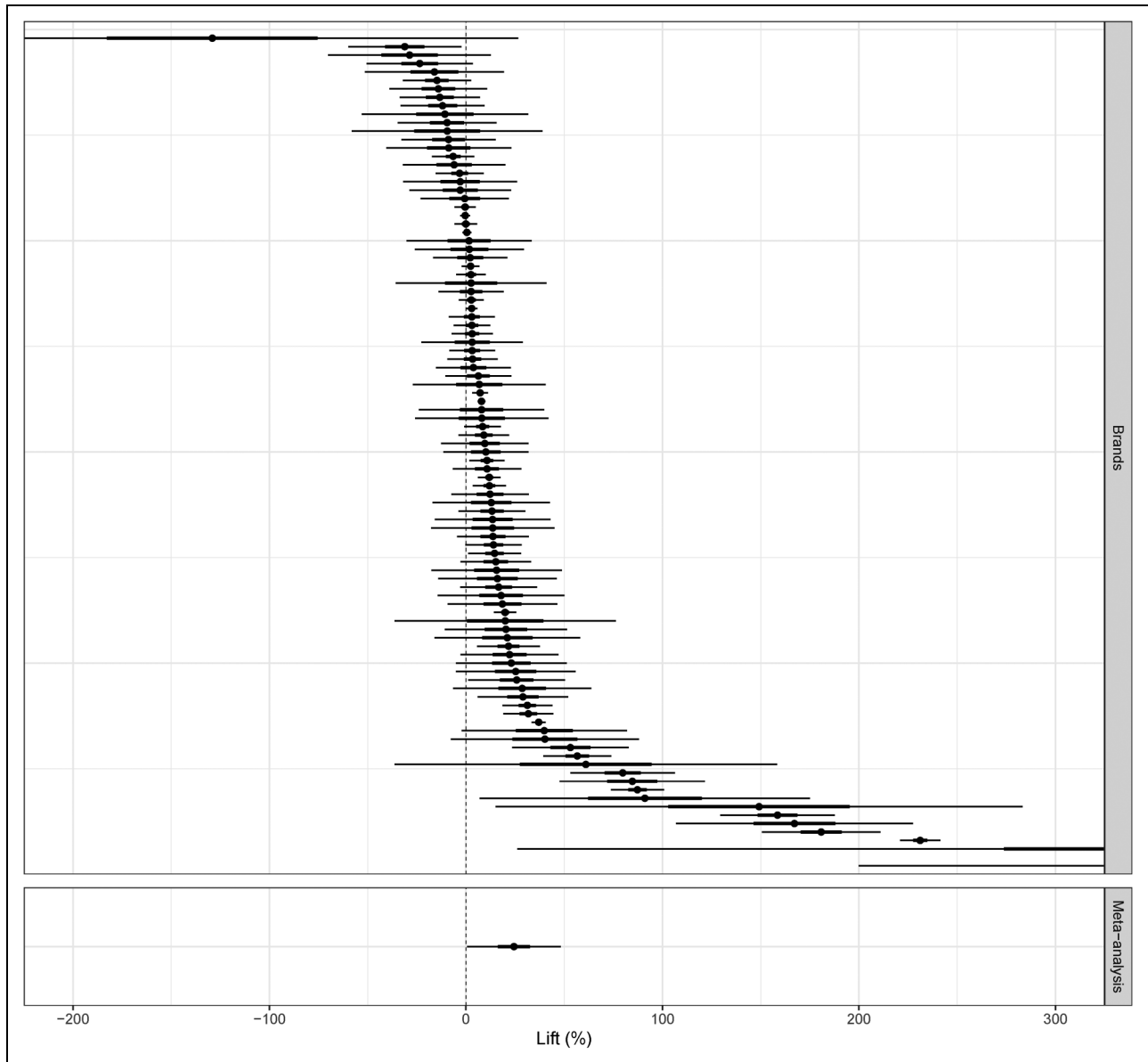
We studied the effect of a small intervention, the inclusion of a call-to-action, on the click rates of online advertisements across 100 brands. In Figure 3, we plot point and 50% and 95% interval estimates of the effect of the intervention, in particular the lift of the intervention advertisement relative to the baseline advertisement, for each brand as well as the meta-analytic average across the 100 brands using the basic random effects meta-analytic model. The figure illustrates that the point estimates varied considerably across brands, with a .25 and .75 quantile of 1.3% and 22.4%, respectively, and a .025 and .975 quantile of  $-30.0\%$  and  $321.8\%$ , respectively. The point estimate of the average lift was  $24.4\%$ , with a 95% interval estimate ranging from  $.6\%$  to  $48.2\%$ . For the interested reader, we note that these estimates correspond to an observed  $Z$ -statistic of  $z = 2.01$  and an observed  $P$ -value of  $p = .044$  against the target hypothesis of no average difference in lift using the Wald test.

Measures of central tendency like this average are not particularly meaningful or of interest given the enormous variation in the effects, which was estimated at  $119.0\%$  with a 95% interval estimate ranging from  $114.4\%$  to  $162.5\%$ . Indeed, the magnitude of the variation in effects is our central result, and it points to a need for future research to identify moderators of it. Consequently, we plan to collect detailed data on each of the 100 brands in order to do so.

### Reporting a Study with Precise “Null” Results

Consider a hypothetical large online field study of the effect of a new selling format on purchase rates. In the study, the observed  $P$ -value  $p$  against the target hypothesis of no difference in the purchase rates is above  $.05$ . Traditionally, a researcher would take this result as demonstrating no difference in the purchase rates. However, this is an error. Therefore, we recommend focusing on quantifying the study results, noting here that they are highly incompatible with a difference that is practically important (and that this would remain the case even if the observed  $P$ -value  $p$  were below  $.05$ ). For example:

In our large online field study of the effect of selling format on purchase,  $7.80\%$  of subjects assigned to the new test format purchased, as compared to  $7.66\%$  of subjects assigned to the standard format—an increase of  $.14$  pp in the purchase rate. In Figure 4, we plot the interval estimate for all levels from  $0\%$  to  $100\%$ . The figure illustrates that any difference in the purchase rate that we would view as nontrivial falls outside the interval estimate at not only any conventional level but also highly extreme levels. For example, values below a  $.23$  pp decrease and above a  $.51$



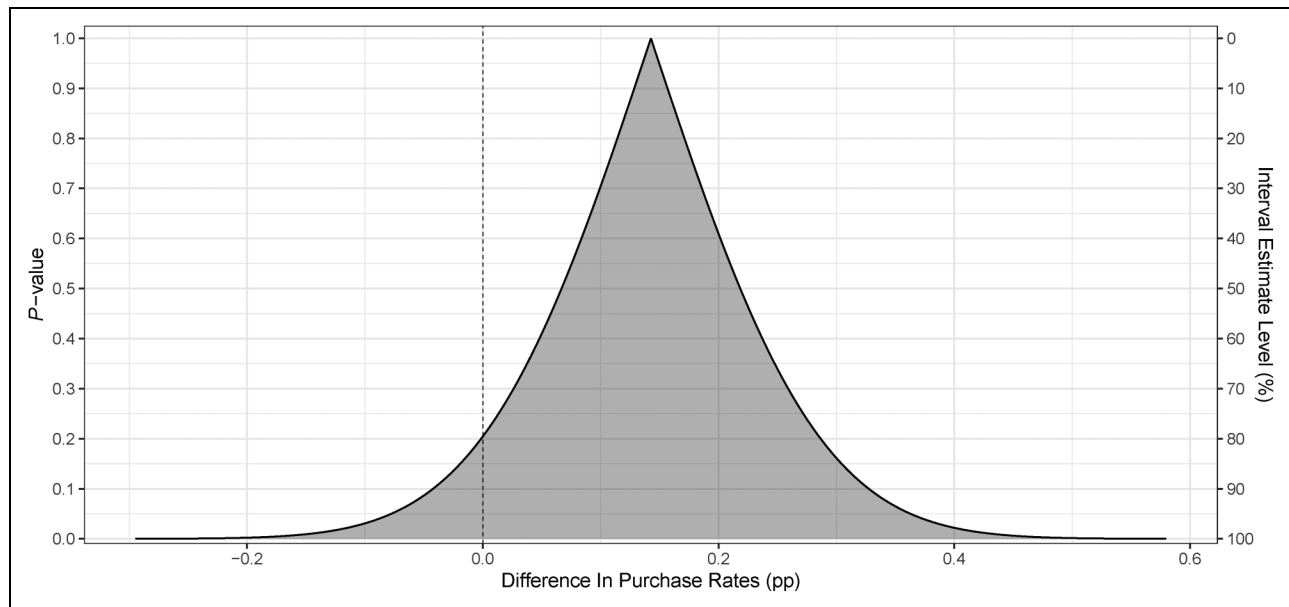
**Figure 3.** Lift Estimates.

Notes: Point estimates are given by the points; 50% and 95% interval estimates are given by the thick and thin lines, respectively. The point and interval estimates for one brand are excluded from the plot due to the lower x-axis limit; the point estimates for three brands and the interval estimate for one of these three brands are excluded from the plot due to the upper x-axis limit.

pp increase, both of which we view as trivial, fall outside the 99.9% interval estimate. Therefore, while it would be wrong to conclude that there is no difference in the purchase rates of the two selling formats, any difference that we would view as being practically important is highly incompatible with the data given all of the assumptions used to compute the interval estimates. For the interested reader, we note that these estimates correspond to an observed Z-statistic of  $z = 1.27$  and an observed P-value of  $p = .204$  against the target hypothesis of no difference in the purchase rates of the two selling formats using the standard two-sample Z-test for proportions.

### Reporting the Rationale for the Sample Size

Consider the rationale for the sample size of the hypothetical study of selling format discussed in the prior example. Traditionally, a researcher would either not report the rationale or report that it was based on an optimistic and easily gamed procedure that purports to achieve a given level of power. However, the goal of a study is not to attain “statistical significance.” Therefore, we recommend reporting the rationale for the sample size, noting here both that it was chosen to achieve a given level of precision for the estimate of the difference in the purchase rates of the two selling formats and that the sample size chosen was not obtained. For example:



**Figure 4.** Interval Estimate Curve.

Notes: The curve plots the interval estimate for the level indicated by the right y-axis. Equivalently, it plots the *P*-value against the target hypothesis of the value of the difference in purchase rates indicated by the x-axis.

In planning our large online field study of the effect of selling format on purchase rates, we chose our sample size to achieve a precise estimate of the difference in the purchase rates of the two selling formats. Specifically, we sought an estimate with precision of .50 pp as assessed by the width of the 99% interval estimate. Given that the purchase rate had historically been about 7.5% per site visit, this level of precision required a sample size of 147,295 subjects per condition, or 294,590 in total. Further, because the firm's website had historically received about 20,000 site visits per day, we decided to run our study for 15 consecutive days. In the end, we obtained a sample size somewhat below the mark, in particular, 225,805 in total.

## Discussion

We have proposed moving beyond binary: abandoning NHST—and the *P*-value thresholds intrinsic to it—as the default approach to statistical analysis and reporting. “Statistical (non)significance” should never be used as a basis to make general and certain conclusions or as a filter to select which results to publish. Instead, all studies should be published in some form or another, and reporting should focus on quantifying study results via point and interval estimates. Further, general conclusions should be made based on the cumulative evidence from multiple studies. This should be done in a manner that treats *P*-values and related statistical measures continuously and as just one factor among many that require joint consideration and holistic integration. It should also be done in a manner that respects the fact that such conclusions are necessarily tentative and subject to revision as new studies are conducted.

While we are optimistic that our proposal will lead to improved statistical analysis and reporting, we conclude by preemptively rebutting one potential criticism of our proposal, raising three considerations, and reiterating the final point made in our Introduction. Against our proposal, some may make the twofold argument that (1) researchers require a bright-line threshold to determine whether a study proffered in support of some claim provides sufficient evidence to warrant making conclusions or publication and (2) thresholds based on *P*-values and related statistical measures provide objective standards for what constitutes sufficient evidence, thereby in turn providing a valuable brake on subjectivity and personal biases.

This argument is misguided along a number of lines. First, the argument has the implicit premise that NHST has proven successful for this purpose. On the contrary, NHST leads researchers to wrongly interpret results that attain “statistical significance” as demonstrating an effect and those that fail to do so as demonstrating no effect. This in turn leads them to use “statistical (non)significance” as a filter to select which results to publish, which again in turn biases the literature and encourages harmful research practices.

Second, a bright-line threshold is not necessary: researchers already weigh evidence and make publication decisions based on various qualitative and quantitative factors, and this could continue to happen if *P*-values and related statistical measures were treated continuously and as just one factor among many.

Third, a study never warrants making conclusions, because single studies are never definitive. Instead, the aim of studies should be to report results in an unfiltered manner so that they can later be used to make more general conclusions based on the cumulative evidence from multiple studies.

Fourth, no single number is capable of eliminating subjectivity and personal biases. Further, there is subjectivity at all stages of scientific inquiry, even if objectivity remains the ultimate goal. Indeed, *P*-values and related statistical measures themselves are subjective in the sense that they are affected by the many necessarily subjective choices involved in study design and statistical analysis.

Turning to our three considerations, first, NHST has long been criticized by both statisticians and applied researchers, and in response to such criticism, proposals similar to ours have long been made. Despite this, NHST has seemed unassailable: such proposals have gone largely if not entirely ignored. We are therefore cautious about the degree to which our proposal will be heeded. That said, researchers throughout the biomedical and social sciences have recently published editorials and articles making proposals akin to ours as well as altered journal guidelines for statistical analysis and reporting so as to adhere to such proposals. Therefore, perhaps both caution and optimism are warranted.

Second, we lack evidence that our proposal will improve statistical analysis and reporting. That said, problems associated with NHST have long been known and efforts to create awareness of and ameliorate them have long been made, and so the evidence is abundant that these efforts have been to no (or at best little) avail. Therefore, perhaps proceeding despite a lack of evidence is warranted.

Third, we acknowledge that our examples of statistical analysis and reporting feature relatively simple settings and statistical analyses. That said, we note that they are realistic of settings and statistical analyses encountered by researchers in all three of marketing's subfields of consumer behavior, strategy, and quantitative modeling and therefore should prove useful to such researchers. One reason is that they apply immediately and without change to many more complex settings and statistical analyses in the sense that the reporting of point and interval estimates is largely, if not entirely, orthogonal to the setting and statistical analysis that gave rise to them. We also note that coalescing around improved statistical analysis and reporting for relatively simple settings and statistical analyses seems like a necessary if not sufficient condition for doing so for more complex settings and statistical analyses. We finally note that we used examples featuring relatively simple settings and statistical analyses in part because we do not want our examples to be used as templates. Templates are dangerous in that they are often applied in a rote and recipe-like manner, much like NHST is currently applied. Instead, we are acutely aware that the implementation of our principles and guidelines ought to vary across and within settings and statistical analyses.

However, this does raise an important issue: many statistical analyses commonly employed but heretofore unmentioned are—as practiced—tethered to NHST. For example, floodlight analysis, mediation analysis, instrumental variables analysis, regression discontinuity analysis, and placebo analysis as well as other analyses employed to argue for the absence of some undesired effect all as practiced use

*P*-values or related statistical measures in the conventional, dichotomous but inappropriate manner. Abandoning this practice when employing these statistical analyses is part and parcel of our proposal to abandon NHST and the *P*-value thresholds intrinsic to it.

In closing, we reiterate the final point made in our Introduction: we believe that no single statistical approach is suitable for all research questions, and thus we advocate a “toolkit” approach that chooses the best one for the job at hand. To be clear, we have no desire—or, for that matter, authority—to “ban” any statistical approach. Nonetheless—and the central point of this article—we believe that it is always inappropriate to use *P*-values and related statistical measures (such as the limits of interval estimates, likelihood ratios, posterior probabilities, and Bayes factors) in the conventional, dichotomous manner, that is, to declare “statistical (non)significance” and decide whether a result proves or disproves a scientific hypothesis based on where the value stands relative to some threshold. Insofar as we have emphasized *P*-values in this article, we have done so only because *P*-values are by far the most common statistical measure used to make such declarations and decisions.

## Appendix

To illustrate the degree to which (1) NHST is employed, (2) problems associated with it are fallen prey to, and (3) our guidelines for statistical analysis and reporting are adhered to in marketing, we reviewed the most recently published empirical article authored by the 33 2023 Marketing Science Institute (MSI) Scholars (excluding the first author of this article). We emphasize that the purpose of this review is to serve as an illustration rather than to be systematic and comprehensive. However, insofar as the MSI Scholars “are among the most prominent marketing scholars in the world” (MSI 2022) as claimed by John Lynch, the Executive Director of MSI, it seems not unreasonable to consider their practices as exemplary and noteworthy.

We report results in Table 1. All 33 papers employed NHST, the target hypothesis of no association or no effect, and the conventional .05 threshold. All erred in reasoning typically by wrongly interpreting results that attained “statistical significance” as demonstrating an effect and those that failed to do so as demonstrating no effect.

While all of the papers reported point estimates, adherence to our other guidelines for statistical analysis and reporting ranged from nil to partial. About one-sixth of the papers reported interval estimates. Of the nine papers that discussed the practical importance of results, they did so only for point estimates: none discussed the practical importance of either, let alone both, the lower and upper limits of interval estimates. About half of the papers reported *P*-values continuously and about half reported them as binary inequalities. All of the papers reported that results were “statistically (non)significant,” with about half reporting that they were “marginally statistically significant” or some other such phrase; in doing so, they typically

**Table 1.** Literature Review Results.

Reference	Practice	Papers
Introduction	Employs null hypothesis significance testing	33
Problem 1	Employs the target hypothesis of no association or no effect	33
Problem 2	Employs the conventional .05 threshold	33
Problem 3	Misinterprets the <i>P</i> -value	N.A.
Problem 4	Errs in reasoning	33
Problem 5	Biases the literature or employs harmful research practices	N.A.
Guideline 1	Reports point estimates	33
	Reports interval estimates	5
	When discussing practical importance, does so not only for point estimates but also for both the lower and upper limits of interval estimates	0
Guideline 2	Reports interval estimates at multiple levels	0
Guideline 3	Reports <i>P</i> -values continuously	18
	Reports <i>P</i> -values for relevant nonnull values	0
	Avoids reporting <i>P</i> -values as binary inequalities	16
	Avoids reporting that results are “statistically significant” or “statistically nonsignificant”	0
	Avoids reporting that results are “marginally statistically significant” or “approaching statistical significance” or some other such phrase	16
	Avoids asterisks or other adornments that signify thresholds	11
Guideline 4	Reports the rationale for the sample size	17
Guideline 5	Eschews binary decisions	16

Notes: None of the 33 papers offered an interpretation of the *P*-value, and therefore none were eligible to misinterpret it. We did not attempt to adjudicate whether a paper biases the literature or employs harmful research practices. N.A. = not applicable.

failed to include the word “statistically,” thereby creating ambiguity as to whether the claimed “significance” was one of “statistical significance” or practical importance. Relatedly, two-thirds of the papers employed asterisks or other adornments that signify thresholds. Finally, about half of the papers reported a rationale for the sample size and about half made decisions typically by declaring what consumers or firms “should” do.

We note that many of the coding decisions required to obtain the results reported in Table 1 were relatively unambiguous, for example, whether a paper employed NHST, the target hypothesis of no association or no effect, and the conventional .05 threshold. However, some required judgment. For example, if a paper almost always reported point estimates but reported only observed test statistics or observed *P*-values for a small number of secondary statistical analyses, it was coded as reporting point estimates. Or, if a paper almost always failed to report interval estimates but reported them only for mediation analyses and used them only to test the target hypothesis of no association or no effect, it was coded as failing to report interval estimates.

## Authors Note

This article originated in Autumn 2019 when David W. Stewart, in his role as Vice President of Publications of the American Marketing Association (AMA) and chair of the AMA Publications Policy Board, asked Robert J. Meyer to convene a committee to discuss problems associated with null hypothesis significance testing and to make a proposal for statistical analysis and reporting so that marketing could keep pace with similar discussions and proposals that have in recent years appeared throughout the biomedical and social sciences. The committee report was delivered in Summer 2020 and evolved into this article.

## Associate Editor

Rajdeep Grewal


## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## ORCID iD

John G. Lynch  <https://orcid.org/0000-0002-4094-3738>

## References

- Altman, Douglas G and J. Martin Bland (1995), “Statistics Notes: Absence of Evidence Is Not Evidence of Absence,” *BMJ*, 311 (7003), 485.
- Amrhein, Valentin and Sander Greenland (2022), “Discuss Practical Importance of Results Based on Interval Estimates and *p*-Value Functions, Not Only on Point Estimates and Null *p*-Values,” *Journal of Information Technology*, 37 (3), 316–20.
- Amrhein, Valentin, Sander Greenland, and Blake McShane (2019a), “Scientists Rise up Against Statistical Significance,” *Nature*, 567 (7748), 305–07.
- Amrhein, Valentin, Sander Greenland, and Blakeley B. McShane (2019b), “Statistical Significance Gives Bias a Free Pass,” *European Journal of Clinical Investigation*, 49 (12), e13176.
- Amrhein, Valentin, David Trafimow, and Sander Greenland (2019), “Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis If We Don’t Expect Replication,” *The American Statistician*, 73 (Supp. 1), 262–70.
- Anderson, David R., Kenneth P. Burnham, and William L. Thompson (2000), “Null Hypothesis Testing: Problems, Prevalence, and an Alternative,” *Journal of Wildlife Management*, 64, 912–23.
- Arbuthnott, John (1710), “II. An Argument for Divine Providence, Taken from the Constant Regularity Observ’d in the Births of Both Sexes. By Dr. John Arbuthnott, Physitian in Ordinary to Her Majesty, and Fellow of the College of Physitians and the Royal Society,” *Philosophical Transactions of the Royal Society*, 27 (328), 186–90.
- Bakan, David (1966), “The Test of Significance in Psychological Research,” *Psychological Bulletin*, 66 (6), 423–37.

- Bayarri, M.J. and James O. Berger (2000), “*P* Values for Composite Null Models,” *Journal of the American Statistical Association*, 95 (452), 1127–42.
- Bayarri, M. Jesús and James O. Berger (2004), “The Interplay of Bayesian and Frequentist Analysis,” *Statistical Science*, 19, 58–80.
- Berkson, Joseph (1938), “Some Difficulties of Interpretation Encountered in the Application of the Chi-Square Test,” *Journal of the American Statistical Association*, 33 (203), 526–36.
- Bernard, Christophe (2019), “Changing the Way We Report, Interpret, and Discuss Our Results to Rebuild Trust in Our Research,” *Eneuro*, 6 (4), <https://doi.org/10.1523/ENEURO.0259-19.2019>.
- Bijak, Jakub (2019), “P-Values, Theory, Replicability, and Rigour,” *Demographic Research*, 41, 949–52.
- Birnbaum, Allan (1961), “A Unified Theory of Estimation, I,” *The Annals of Mathematical Statistics*, Pages, 32 (1), 112–35.
- Bonovas, Stefanos and Daniele Piovani (2023), “On *p*-Values and Statistical Significance,” *Journal of Clinical Medicine*, 12 (3), 900.
- Boring, Edwin G. (1919), “Mathematical vs. Scientific Significance,” *Psychological Bulletin*, 16 (10), 335–38.
- Box, George E.P. (1980), “Sampling and Bayes’ Inference in Scientific Modelling and Robustness,” *Journal of the Royal Statistical Society: Series A (General)*, 143 (4), 383–404.
- Breese, Lauren (2019), “Do We Give Too Much Significance to Statistical Significance?” *Canadian Journal of Hospital Pharmacy*, 72 (5), 339–40.
- Briggs, William M. (2016), *Uncertainty: The Soul of Modeling, Probability and Statistics*. Springer.
- Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg (2016), “Star Wars: The Empirics Strike Back,” *American Economic Journal: Applied Economics*, 8 (1), 1–32.
- Brown, Sacha D., David Furrow, Daniel F. Hill, Jonathon C. Gable, Liam P. Power, and W. Jake Jacobs (2014), “A Duty to Describe: Better the Devil You Know Than the Devil You Don’t,” *Perspectives on Psychological Science*, 9 (6), 626–40.
- Butler, Ruth C. (2022), “Popularity Leads to Bad Habits: Alternatives to ‘the Statistics’ Routine of Significance, ‘Alphabet Soup’ and Dynamite Plots,” *Annals of Applied Biology*, 180 (2), 182–95.
- Carlsson, Sigrid V. and Mithat Gönen (2020), “Towards Wiser Use and Interpretation of *P* Values,” *Journal of Sexual Medicine*, 17 (1), 1–3.
- Charlesworth, M. and J. J. Pandit (2020), “Negative Outcomes in Critical Care Trials: Applying the Wrong Statistics—or Asking the Wrong Questions?” *Anaesthesia*, 75 (10), 1284–88.
- Chernev, Alexander, Ulf Böckenholt, and Joseph Goodman (2015), “Choice Overload: A Conceptual Review and Meta-Analysis,” *Journal of Consumer Psychology*, 25 (2), 333–58.
- Cochran, William G. (1976), “Early Development of Techniques in Comparative Experimentation,” in *On the History of Statistics and Probability*, D.B. Owen, ed. Dekker, 1–25.
- Cohen, Jacob (1994), “The Earth Is Round ( $p < .05$ ),” *American Psychologist*, 49 (12), 997–1003.
- Cowles, Michael and Caroline Davis (1982), “On the Origins of the .05 Level of Significance,” *American Psychologist*, 44 (5), 1276–84.
- Cox, David R. (1958), “Some Problems Connected with Statistical Inference,” *Annals of Mathematical Statistics*, 29 (2), 357–72.
- Cox, David R. (1977), “The Role of Significance Tests,” *Scandinavian Journal of Statistics*, 4 (2), 49–70.
- Cox, David R. and Christl A. Donnelly (2011), *Principles of Applied Statistics*. Cambridge University Press.
- Cox, David R. and David V. Hinkley (1974), *Theoretical Statistics*. Chapman and Hall.
- Curran-Everett, Douglas (2019), “Statistical Considerations for Occupational and Environmental Physiology,” *Temperature*, 6 (3), 179–80.
- Curran-Everett, Douglas (2020), “Evolution in Statistics: P values, Statistical Significance, Kayaks, and Walking Trees,” *Advances in Physiology Education*, 44 (2), 221–24.
- Davidson, Andrew (2019), “Embracing Uncertainty: The Days of Statistical Significance Are Numbered,” *Paediatric Anaesthesia*, 29 (10), 978–80.
- De Koning, Jos J. and Dionne A. Noordhof (2019), “Embrace Uncertainty,” *International Journal of Sports Physiology and Performance*, 14 (6), 697.
- Dirnagl, Ulrich (2019), “The *p* Value Wars (Again),” *European Journal of Nuclear Medicine and Molecular Imaging*, 46, 2421–23.
- Edwards, Ward, Harold Lindman, and Leonard J. Savage (1963), “Bayesian Statistical Inference for Psychological Research,” *Psychological Review*, 70 (3), 193–242.
- Efron, Bradley and Trevor Hastie (2016), *Computer Age Statistical Inference*. Cambridge University Press.
- Elkins, Mark R., Rafael Zambelli Pinto, Arianne Verhagen, Monika Grygorowicz, Anne Söderlund, Matthieu Guemann, Antonia Gómez-Conesa, Sarah Blanton, Jean-Michel Brismée, Shabnam Agarwal, Alan Jette, Sven Karstens, Michele Harms, Geert Verheyden, and Umer Sheikh (2022), “Statistical Inference Through Estimation: Recommendations from the International Society of Physiotherapy Journal Editors,” *Journal of Physiotherapy*, 68 (1), 1–4.
- Fabrigar, Leandre R. and Duane T. Wegener (2016), “Conceptualizing and Evaluating the Replication of Research Results,” *Journal of Experimental Social Psychology*, 66, 68–80.
- Filippini, Tommaso and Silvio Roberti Vinceti (2022), “The Role of Statistical Significance Testing in Public Law and Health Risk Assessment,” *Journal of Preventative Medicine and Hygiene*, 61 (1), E161–5.
- Fingerhut, Abe (2023), “Probability, Values, and Statistical Significance: Instructions for Use by Surgeons,” *British Journal of Surgery*, 110 (4), 399–400.
- Fisher, Ronald Aylmer (1926), “The Arrangement of Field Experiments,” *Journal of the Ministry of Agriculture*, 33, 503–13.
- Fisher, Ronald Aylmer (1934), *Statistical Methods for Research Workers*. Oliver and Boyd.
- Fisher, Ronald Aylmer (1956), *Statistical Methods and Scientific Inference*. Hafner Publishing Co.
- Fisher, Ronald Aylmer (1935a), *The Design of Experiments*. Oliver & Boyd.
- Fisher, Ronald Aylmer (1935b), “Letter to the Editor: Statistical Tests,” *Nature*, 136, 474.
- Folks, J. Leroy (1981), *Ideas of Statistics*. John Wiley & Sons.
- Freeman, P.R. (1993), “The Role of *p*-Values in Analysing Trial Results,” *Statistics in Medicine*, 12 (15–16), 1443–52.

- Gardner, Martin J. and Douglas G. Altman (1986), "Confidence Intervals Rather Than P Values: Estimation Rather Than Hypothesis Testing," *British Medical Journal (Clinical Research Edition)*, 292 (6522), 746–50.
- Gelman, Andrew (2016), "The Problems with p-Values Are Not Just with p-Values," *The American Statistician, Online Discussion*, [https://stat.columbia.edu/~gelman/research/published/asa\\_pvalues.pdf](https://stat.columbia.edu/~gelman/research/published/asa_pvalues.pdf).
- Gelman, Andrew (2017), "Ethics and Statistics: Honesty and Transparency Are Not Enough," *Chance*, 30 (1), 37–39.
- Gelman, Andrew (2020), "Alexey Guzey's Sleep Deprivation Self-Experiment," blog entry (May 26), *Statistical Modeling, Causal Inference, and Social Science*, <https://statmodeling.stat.columbia.edu/2020/05/26/alexey-guzeys-sleep-deprivation-self-experiment/>.
- Gelman, Andrew and John Carlin (2014), "Beyond Power Calculations Assessing Type s (Sign) and Type m (Magnitude) Errors," *Perspectives on Psychological Science*, 9 (6), 641–51.
- Gelman, Andrew (2019a), "Comment on 'Post-Hoc Power Using Observed Estimate of Effect Size Is Too Noisy to Be Useful,'" *Annals of Surgery*, 270 (2), e64.
- Gelman, Andrew (2019b), "Don't Calculate Post-Hoc Power Using Observed Estimate of Effect Size," *Annals of Surgery*, 269 (1), e9–e10.
- Gelman, Andrew and Eric Loken (2013), "The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No 'Fishing Expedition' or 'p-Hacking' and the Research Hypothesis Was Posited Ahead of Time," (November 14), <http://www.stat.columbia.edu/~gelman/research/unpublished/forking.pdf>.
- Gelman, Andrew and Eric Loken (2014), "The Statistical Crisis in Science," *American Scientist*, 102 (6), 460–65.
- Gelman, Andrew and Hal Stern (2006), "The Difference Between 'Significant' and 'Not Significant' Is Not Itself Statistically Significant," *The American Statistician*, 60 (4), 328–31.
- Gigerenzer, Gerd (2004), "Mindless Statistics," *Journal of Socio-Economics*, 33 (5), 587–606.
- Gigerenzer, Gerd (2018), "Statistical Rituals: The Replication Delusion and How We Got There," *Advances in Methods and Practices in Psychological Science*, 1 (2), 198–218.
- Gigerenzer, Gerd, S. Krauss, and O. Vitouch (2004), "The Null Ritual: What You Always Wanted to Know About Null Hypothesis Testing but Were Afraid to Ask," in *The SAGE Handbook of Quantitative Methodology for the Social Sciences*, David Kaplan, ed. Sage Publications, 389–406.
- Gill, Jeff (1999), "The Insignificance of Null Hypothesis Significance Testing," *Political Research Quarterly*, 52 (3), 647–74.
- Goodman, Steven N. (1992), "A Comment on Replication, P-Values, and Evidence," *Statistics in Medicine*, 11 (7), 875–79.
- Goodman, Steven N. (2008), "A Dirty Dozen: Twelve p-Value Misconceptions," *Seminars in Hematology*, 45 (3), 135–40.
- Greenland, Sander (1987), "Quantitative Methods in the Review of Epidemiologic Literature," *Epidemiologic Reviews*, 9 (1), 1–30.
- Greenland, Sander (2012), "Nonsignificance Plus High Power Does Not Imply Support for the Null over the Alternative," *Annals of Epidemiology*, 22 (5), 364–68.
- Greenland, Sander (2016), "The ASA Guidelines and Null Bias in Current Teaching and Practice," *The American Statistician*, 70 (Supp. 10), [https://www.tandfonline.com/action/downloadSupplement?doi=10.1080%2F00031305.2016.1154108&file=utas\\_a\\_1154108\\_sm5079.pdf](https://www.tandfonline.com/action/downloadSupplement?doi=10.1080%2F00031305.2016.1154108&file=utas_a_1154108_sm5079.pdf).
- Greenland, Sander (2017), "Invited Commentary: The Need for Cognitive Science in Methodology," *American Journal of Epidemiology*, 186 (6), 639–45.
- Greenland, Sander (2019), "Valid P-Values Behave Exactly as They Should: Some Misleading Criticisms of P-Values and Their Resolution with S-Values," *The American Statistician*, 73 (Supp. 1), 106–14.
- Greenland, Sander (2023a), "Connecting Simple and Precise P-Values to Complex and Ambiguous Realities (Includes Rejoinder to Comments on 'Divergence vs. Decision P-Values')," *Scandinavian Journal of Statistics*, 50 (3), 899–914.
- Greenland, Sander (2023b), "Divergence Versus Decision P-Values: A Distinction Worth Making in Theory and Keeping in Practice: Or, How Divergence P-Values Measure Evidence Even When Decision P-Values Do Not," *Scandinavian Journal of Statistics*, 50 (1), 54–88.
- Greenland, Sander, Mohammad Ali Mansournia, and Michael Joffe (2022), "To Curb Research Misreporting, Replace Significance and Confidence by Compatibility: A Preventive Medicine Golden Jubilee Article," *Preventive Medicine*, 164, 107127.
- Greenland, Sander and Keith O'Rourke (2008), "Meta-Analysis," in *Modern Epidemiology*, 3rd ed., Kenneth J. Rothman, Sander Greenland, and Timothy L. Lash, eds. Lippincott, Williams, and Wilkins, 652–82.
- Greenland, Sander, James J. Schlesselman, and Michael H. Criqui (1986), "The Fallacy of Employing Standardized Regression Coefficients and Correlations as Measures of Effect," *American Journal of Epidemiology*, 123 (2), 203–08.
- Greenland, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman (2016), "Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations," *The American Statistician*, 70 (2), 1–12. (Online Supplement).
- Harrington, David, Ralph B. D'Agostino Sr., Constantine Gatsonis, Joseph W. Hogan, David J. Hunter, Sharon-Lise T. Normand, Jeffrey M. Drazen, and Mary Beth Hamel (2019), "New Guidelines for Statistical Reporting in the Journal," *New England Journal of Medicine*, 381, 285–86.
- Harvey, Lisa A. and Martin W.G. Brinkhof (2019), "Imagine a Research World Without the Words 'Statistically Significant'. Is it Really Possible?" *Spinal Cord*, 57, 437–38.
- Hassler, Uwe (2023), "From Fact to Fake: The Importance of Being Significant," *Research in Statistics*, 1 (1), 2236995.
- Hayat, Matthew J., Vincent Staggs, Todd A. Schwartz, Melinda Higgins, Andres Azuero, Chakra Budhathoki, Rameela Chandrasekhar, Paul Cook, Emily Cramer, Mary S. Dietrich, Mauricio Garnier-Villarreal, Alexandra Hanlon, Jianghua He, Jinxiang Hu, MyoungJin Kim, Martina Mueller, Joseph R. Nolan, Yelena Perkhounkova, Janet Rothers, Glenna Schluck, et al. (2019), "Moving Nursing Beyond  $p < 0.05$ ," *Research in Nursing and Health*, 42 (4), 244–45.



- Head, Megan L., Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions (2015), "The Extent and Consequences of p-Hacking in Science," *PLoS Biology*, 13 (3), e1002106.
- Heckelei, Thomas, Silke Hüttel, Martin Odening, and Jens Rommel (2021), "The Replicability Crisis and the p-value Debate—What Are the Consequences for the Agricultural and Food Economics Community?" technical report, <https://doi.org/10.20944/preprints202201.0311.v1>.
- Hoening, John M. and Dennis M. Heisey (2001), "The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis," *The American Statistician*, 55 (1), 19–24.
- Holman, C. D'Arcy J., Diane E. Arnold-Reed, Nicholas de Klerk, Christine McComb, and Dallas R. English (2001), "A Psychometric Experiment in Causal Inference to Estimate Evidential Weights Used by Epidemiologists," *Epidemiology*, 12 (2), 246–55.
- Hubbard, Raymond (2004), "Alphabet Soup: Blurring the Distinctions Between p's and a's in Psychological Research," *Theory and Psychology*, 14 (3), 295–327.
- Hunter, John E. (1997), "Needed: A Ban on the Significance Test," *Psychological Science*, 8 (1), 3–7.
- Imbens, Guido W. (2021), "Statistical Significance, p-Values, and the Reporting of Uncertainty," *Journal of Economic Perspectives*, 35 (3), 157–74.
- John, Leslie K., George Loewenstein, and Drazen Prelec (2012), "Measuring the Prevalence of Questionable Research Practices with Incentives for Truth-Telling," *Psychological Science*, 23 (5), 524–32.
- Johnson, Samantha L., Whitley J. Stone, Jennifer A. Bunn, T. Scott Lyons, and James W. Navalta (2020), "New Author Guidelines in Statistical Reporting: Embracing an Era Beyond  $p < .05$ ," *International Journal of Exercise Science*, 13 (1), 1–5.
- Kelley, Ken, Scott E. Maxwell, and Joseph R. Rausch (2003), "Obtaining Power or Obtaining Precision: Delineating Methods of Sample-Size Planning," *Evaluation & the Health Professions*, 26 (3), 258–87.
- Kempthorne, Oscar (1976), "Of What Use Are Tests of Significance and Tests of Hypothesis," *Communications in Statistics – Theory and Methods*, 5 (8), 763–77.
- Kempthorne, Oscar and Leroy Folks (1971), *Probability, Statistics, and Data Analysis*. Iowa State University Press.
- Knottnerus, J. André and Peter Tugwell (2020), "Thresholds and Innovation: Discussion on Statistical Significance," *Journal of Clinical Epidemiology*, 118, A5–7.
- Lane, David M. and William P. Dunlap (1978), "Estimating Effect Size: Bias Resulting from the Significance Criterion in Editorial Decisions," *British Journal of Mathematical and Statistical Psychology*, 31 (2), 107–12.
- Lehmann, Erich L. (1986), *Testing Statistical Hypotheses*, 3rd ed. Springer Science & Business Media.
- Lehmann, Erich L. and Joseph P. Romano (2005), *Testing Statistical Hypotheses*, 3rd ed. Springer Science & Business Media.
- Light, Richard J. and David B. Pillemer (1984), *Summing up: The Science of Reviewing Research*. Harvard University Press.
- Lowe, Nancy K. (2019), "The Push to Move Health Care Science Beyond  $p < .05$ ," *Journal of Obstetric, Gynecologic, and Neonatal Nursing*, 48 (5), 493–94.
- Lykken, David T. (1968), "Statistical Significance in Psychological Research," *Psychological Bulletin*, 70 (3, Part 1), 151–59.
- Macarthur, John W. (1926), "Linkage Studies with the Tomato," *Genetics*, 11 (4), 387–405.
- Marshall, Andrea P. (2019), "Living with Uncertainty in Clinical Research," *Australian Critical Care*, 32 (3), 183–84.
- Marshall, Andrea P. and Ian Hughes (2020), "Statistics: The Grammar of Science," *Australian Critical Care*, 33 (2), 113–15.
- Masicampo, E.J. and Daniel R. Lalande (2012), "A Peculiar Prevalence of p Values Just Below .05," *Quarterly Journal of Experimental Psychology*, 65 (11), 2271–79.
- Maula, Markku and Wouter Stam (2020), "Enhancing Rigor in Quantitative Entrepreneurship Research," *Entrepreneurship Theory and Practice*, 44 (6), 1059–90.
- McCloskey, Deidre N. and Stephen T. Ziliak (1996), "The Standard Error of Regressions," *Journal of Economic Literature*, 34 (1), 97–114.
- McShane, Blakeley B. and Ulf Böckenholt (2014), "You Cannot Step into the Same River Twice: When Power Analyses Are Optimistic," *Perspectives on Psychological Science*, 9 (6), 612–25.
- McShane, Blakeley B. and Ulf Böckenholt (2016), "Planning Sample Sizes When Effect Sizes Are Uncertain: The Power-Calibrated Effect Size Approach," *Psychological Methods*, 21 (1), 47–60.
- McShane, Blakeley B. and Ulf Böckenholt (2017), "Single Paper Meta-Analysis: Benefits for Study Summary, Theory-Testing, and Replicability," *Journal of Consumer Research*, 43 (6), 1048–63.
- McShane, Blakeley B. and Ulf Böckenholt (2018), "Multilevel Multivariate Meta-Analysis with Application to Choice Overload," *Psychometrika*, 83 (1), 255–71.
- McShane, Blakeley B. and Ulf Böckenholt (2022), "Meta-Analysis of Studies with Multiple Contrasts and Differences in Measurement Scales," *Journal of Consumer Psychology*, 32 (1), 23–40.
- McShane, Blakeley B., Ulf Böckenholt, and Karsten T. Hansen (2016), "Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes," *Perspectives on Psychological Science*, 11 (5), 730–49.
- McShane, Blakeley B., Ulf Böckenholt, and Karsten T. Hansen (2020), "Average Power: A Cautionary Note," *Advances in Methods and Practices in Psychological Science*, 3 (2), 185–99.
- McShane, Blakeley B., Ulf Böckenholt, and Karsten T. Hansen (2022), "Variation and Covariation in Large-Scale Replication Projects: An Evaluation of Replicability," *Journal of the American Statistical Association*, 117 (540), 1605–21.
- McShane, Blakeley B. and David Gal (2016), "Blinding Us to the Obvious? The Effect of Statistical Training on the Evaluation of Evidence," *Management Science*, 62 (6), 1707–18.
- McShane, Blakeley B. and David Gal (2017), "Statistical Significance and the Dichotomization of Evidence," *Journal of the American Statistical Association*, 112 (519), 885–95.
- McShane, Blakeley B., David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett (2019), "Abandon Statistical Significance," *The American Statistician*, 73 (Supp. 1), 235–45.
- McShane, Blakeley B. and Andrew Gelman (2022), "Selecting on Statistical Significance and Practical Importance Is Wrong," *Journal of Information Technology*, 37 (3), 312–5.

- Meehl, Paul E. (1967), "Theory-Testing in Psychology and Physics: A Methodological Paradox," *Philosophy of Science*, 34 (2), 103–15.
- Meehl, Paul 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, 806–34.
- Meehl, Paul E. (1990), "Why Summaries of Research on Psychological Theories Are Often Uninterpretable," *Psychological Reports*, 66 (1), 195–244.
- Michel, Martin C., T.J. Murphy, and Harvey J. Motulsky (2020), "New Author Guidelines for Displaying Data and Reporting Data Analysis and Statistical Methods in Experimental Biology," *Journal of Pharmacology and Experimental Therapeutics*, 372 (1), 136–47.
- Montero, Olimpio, Mikael Hedeland, and David Balgoma (2023), "Trials and Tribulations of Statistical Significance in Biochemistry and Omics," *Trends in Biochemical Sciences*, 48 (6), 503–12.
- Morken, Nils-Halvdan (2019), "Victims and Addicts of Biostatistics," *Acta Obstetricia et Gynecologica Scandinavica*, 98 (9), 1085.
- Morrison, D.E. and R.E. Henkel (1970), *The Significance Test Controversy*. Aldine.
- MSI (2022), "2022 MSI Scholars Announced," (November 4), <https://www.msi.org/article/2023-msi-scholars-announced/>.
- Nguyen, Tuan V., Fernando Rivadeneira, and Roberto Civitelli (2019), "New Guidelines for Data Reporting and Statistical Analysis: Helping Authors with Transparency and Rigor in Research," *Journal of Bone and Mineral Research*, 34 (11), 1981–84.
- Oakes, Michael (1986), *Statistical Inference: A Commentary for the Social and Behavioral Sciences*. John Wiley & Sons.
- O'Connor, Christopher M. (2019), "A Call for Change: Level of Statistical Significance," *Journal of the American College of Cardiology: Heart Failure*, 7 (11), 993–94.
- Parsons, Nick, Richard Carey-Smith, Melina Dritsaki, Xavier Griffin, David Metcalfe, Daniel Perry, Dirk Stengel, and Matthew Costa (2019), "Statistical Significance and p-Values: Guidelines for Use and Reporting," *Bone and Joint Journal*, 101-B (10), 1179–83.
- Pearson, Karl (1900), "On the Criterion That a Given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen from Random Sampling," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50 (302), 157–75.
- Pearson, Karl (1904), "Report on Certain Enteric Fever Inoculation Statistics," *British Medical Journal*, 2 (2288), 1243–46.
- Pearson, Karl (1906), "Note on the Significant or Non-Significant Character of a Sub-Sample Drawn from a Sample," *Biometrika*, 5 (1–2), 181–83.
- Pearson, Karl (1935), "Letter to the Editor: Statistical Tests," *Nature*, 136, 550.
- Pickler, Rita H. (2019), "The Problem with p and Statistical Significance," *Nursing Research*, 68 (6), 421–22.
- Poole, Charles (1987), "Beyond the Confidence Interval," *American Journal of Public Health*, 77 (2), 195–99.
- Price, Robert, Rob Bethune, and Lisa Massey (2020), "Problem with p Values: Why p Values Do Not Tell You If Your Treatment Is Likely to Work," *Postgraduate Medical Journal*, 96 (1131), 1–3.
- Putnam, Persis (1927), "Sex Differences in Pulmonary Tuberculosis Deaths," *American Journal of Hygiene*, 7 (6), 663–705.
- Putt, Mary E. (2021), "Assessing Risk Factors with Information Beyond P Value Thresholds: Statistical Significance Does Not Equal Clinical Importance," *Cancer*, 127 (8), 1180–85.
- Rafi, Zad and Sander Greenland (2020), "Semantic and Cognitive Tools to Aid Statistical Science: Replace Confidence and Significance by Compatibility and Surprise," *BMC Medical Research Methodology*, 20 (1), 1–13.
- Rafi, Zad and Andrew D. Vigotsky (2020), *concurve: Computes and Plots Compatibility (Confidence) Intervals, P-Values, S-Values, & Likelihood Intervals to Form Consonance, Surprisal, & Likelihood Functions*, <https://CRAN.R-project.org/package=concurve>, r package version 2.4.2.
- Robins, James M., Aad van der Vaart, and Valérie Ventura (2000), "Asymptotic Distribution of p Values in Composite Null Models," *Journal of the American Statistical Association*, 95 (452), 1143–56.
- Robinson, R. and J.S. Haviland (2021), "Understanding Statistical Significance and Avoiding Common Pitfalls," *Clinical Oncology*, 33 (12), 804–06.
- Rosenthal, Robert (1990), "Replication in Behavioral Research," *Journal of Social Behavior and Personality*, 5 (4), 1–30.
- Rosnow, Ralph L. and Robert Rosenthal (1989), "Statistical Procedures and the Justification of Knowledge in Psychological Science," *American Psychologist*, 44 (10), 1276–84.
- Rothman, Kenneth J. (1978), "A Show of Confidence," *New England Journal of Medicine*, 299 (24), 1362–63.
- Rothman, Kenneth J. (1986), "Significance Questing," *Annals of Internal Medicine*, 105 (3), 445–47.
- Rozeboom, William W. (1960), "The Fallacy of the Null Hypothesis Significance Test," *Psychological Bulletin*, 57, 416–28.
- Salsburg, David S. (1985), "The Religion of Statistics as Practiced in Medical Journals," *The American Statistician*, 39 (3), 220–23.
- Santibáñez, Miguel, Juan Luis García-Rivero, and Esther Barreiro (2020), "P of Significance: Is it Better to Avoid It If It Is Poorly Understood?" *Archivos de Bronconeumología*, 56 (10), 613–14.
- Sawyer, Alan G. and J. Paul Peter (1983), "The Significance of Statistical Significance Tests in Marketing Research," *Journal of Marketing Research*, 20 (2), 122–33.
- Schmidt, Frank L. (1996), "Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for the Training of Researchers," *Psychological Methods*, 1 (2), 115–29.
- Senn, Stephen (2002), "Letter to the Editor: A Comment on Replication, p-Values, and Evidence," *Statistics in Medicine*, 21, 2437–44.
- Serlin, Ronald C. and Daniel K. Lapsley (1993), "Rational Appraisal Psychological Research and the Good Enough Principle," in *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, Gideon Keren and Charles Lewis, eds. Lawrence Erlbaum Associates, 199–228.
- Shafer, Glenn (2020), "On the Nineteenth-Century Origins of Significance Testing and p-Hacking," working paper 55, Probability and Finance (July 18), <http://www.probabilityandfinance.com/articles/55.pdf>.
- Shannon, Claude Elwood (1948), "A Mathematical Theory of Communication," *Bell System Technical Journal*, 27 (3), 379–423.
- Staggs, Vincent S. (2019), "Why Statisticians Are Abandoning Statistical Significance," *Research in Nursing and Health*, 42 (3), 159–60.
- Sullivan, Kevin M. and David A. Foster (1990), "Use of the Confidence Interval Function," *Epidemiology*, 1 (1), 39–42.

- Sun, C.P. (1928), "On the Examination of Final Digits by Experiments in Artificial Sampling," *Biometrika*, 20A (1–2), 64–68
- Tijssen, Jan G.P. (2021), "More Confidence Intervals and Fewer p Values: A Positive Trend?" *Journal of the American College of Cardiology*, 77 (12), 1562–63.
- Tukey, John W. (1969), "Analyzing Data: Sanctification or Detective Work?" *American Psychologist*, 24 (2), 83–91.
- Tukey, John W. (1991), "The Philosophy of Multiple Comparisons," *Statistical Science*, 6, 100–116.
- Van Witteloostuijn, Arjen (2020), "New-Day Statistical Thinking: A Bold Proposal for a Radical Change in Practices," *Journal of International Business Studies*, 51 (2), 274–78.
- Verykoui, Eleni and Christos T. Nakas (2023), "Adaptations on the Use of  $p$ -Values for Statistical Inference: An Interpretation of Messages from Recent Public Discussions," *Stats*, 6 (2), 539–51.
- Wasserstein, Ronald L. and Nicole A. Lazar (2016), "The ASA's Statement on  $p$ -Values: Context, Process, and Purpose," *The American Statistician*, 70 (2), 129–33.
- Wasserstein, Ronald L., Allen L. Schirm, and Nicole A. Lazar (2019), "Moving to a World Beyond ' $p < 0.05$ ,'" *The American Statistician*, 73 (Supp. 1), 1–19.
- Wilkinson, Leland (1999), "Statistical Methods in Psychology Journals: Guidelines and Explanations," *American Psychologist*, 54 (8), 594–604.
- Yuan, Ke-Hai and Scott E. Maxwell (2005), "On the Post Hoc Power in Testing Mean Differences," *Journal of Educational and Behavioral Statistics*, 30 (2), 141–67.