JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2017, VOL. 112, NO. 519, 885–908, Applications and Case Studies
https://doi.org/10.1080/01621459.2017.1289846

Taylor & Francis
Taylor & Francis Group



# Statistical Significance and the Dichotomization of Evidence

Blakeley B. McShane[a] and David Gal[b]

[a]Kellogg School of Management, Northwestern University, Evanston, IL; [b]College of Business Administration, University of Illinois at Chicago, Chicago, IL

**ABSTRACT**

In light of recent concerns about reproducibility and replicability, the ASA issued a *Statement on Statistical Significance and p-values* aimed at those who are not primarily statisticians. While the ASA Statement notes that statistical significance and *p*-values are "commonly misused and misinterpreted," it does not discuss and document broader implications of these errors for the interpretation of evidence. In this article, we review research on how applied researchers who are not primarily statisticians misuse and misinterpret *p*-values in practice and how this can lead to errors in the interpretation of evidence. We also present new data showing, perhaps surprisingly, that researchers who *are* primarily statisticians are also prone to misuse and misinterpret *p*-values thus resulting in similar errors. In particular, we show that statisticians tend to interpret evidence dichotomously based on whether or not a *p*-value crosses the conventional 0.05 threshold for statistical significance. We discuss implications and offer recommendations.

**ARTICLE HISTORY**
Received May 2016
Revised December 2016

**KEYWORDS**
Null hypothesis significance testing; *p*-value; Statistical significance; Sociology of science



## 1. Introduction

In light of a number of recent high-profile academic and popular press articles critical of the use of the null hypothesis significance testing (NHST) paradigm in applied research as well as concerns about reproducibility and replicability more broadly, the Board of Directors of the American Statistical Association (ASA) issued a *Statement on Statistical Significance and p-values* (Wasserstein and Lazar 2016). The ASA Statement, aimed at "researchers, practitioners, and science writers who are not primarily statisticians," consists of six principles:

P1. *p*-values can indicate how incompatible the data are with a specified statistical model.
P2. *p*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
P3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.
P4. Proper inference requires full reporting and transparency.
P5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
P6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

The ASA Statement notes "Nothing in the ASA statement is new. Statisticians and others have been sounding the alarm about these matters for decades, to little avail" (Wasserstein and Lazar 2016). Indeed, P1, P2, and P5 follow from the definition of the *p*-value; P3 and P5 are repeatedly emphasized in introductory textbooks; P4 is a general principle of epistemology; and P6 has long been a subject of research (Edwards, Lindman, and Savage 1963; Berger and Sellke 1987; Cohen 1994; Hubbard and Lindsay 2008; Johnson 2013).

Among these six principles, considerable attention has been given to P3, which covers issues surrounding the dichotomization of evidence based solely on whether or not a *p*-value crosses a specific threshold such as the hallowed 0.05 threshold. For example, in the press release of March 7, 2016 announcing the publication of the ASA Statement, Ron Wasserstein, Executive Director of the ASA, was quoted as saying:

> The *p*-value was never intended to be a substitute for scientific reasoning. Well-reasoned statistical arguments contain much more than the value of a single number and whether that number exceeds an arbitrary threshold. The ASA statement is intended to steer research into a "post $p < 0.05$ era."

Additionally, the ASA Statement concludes with the sentence "No single index should substitute for scientific reasoning."

While the ASA Statement notes that statistical significance and *p*-values are "commonly misused and misinterpreted" (Wasserstein and Lazar 2016) in applied research, in line with its focus on general principles it does not discuss and document broader implications of these errors for the interpretation of evidence. Thus, in this article, we review research on how applied researchers who are not primarily statisticians misuse and misinterpret *p*-values in practice and how this can lead to errors in the interpretation of evidence. We also present new data showing, perhaps surprisingly, that researchers who *are* primarily statisticians are also prone to misuse and misinterpret *p*-values thus resulting in similar errors. In particular, we show that—like applied researchers who are not primarily statisticians—statisticians also tend to fail to heed P3, interpreting evidence dichotomously based on whether or not a *p*-value crosses the

**CONTACT** Blakeley B. McShane b-mcshane@kellogg.northwestern.edu Associate Professor, Marketing Department, Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60611.

conventional 0.05 threshold for statistical significance. In sum, the assignment of evidence to the different categories "statistically significant" and "not statistically significant" appears to be simply too strong an inducement to the conclusion that the items thusly assigned are categorically different—even to those who are most aware of and thus should be most resistant to this line of thinking. We discuss implications and offer recommendations.

## 2. Misuse and Misinterpretation of *p*-Values in Applied Research

There is a long line of work documenting how applied researchers misuse and misinterpret $p$-values in practice. In this section, we briefly review some of this work that relates to P2, P3, and P5 with a focus on P3.

While formally defined as the probability of observing data as extreme or more extreme than that actually observed assuming the null hypothesis is true, the $p$-value is often misinterpreted by applied researchers not only as "the probability that the studied hypothesis is true or the probability that the data were produced by random chance alone" (P2) but also as the probability that the null hypothesis is true and one minus the probability of replication. For example, Gigerenzer (2004) reported an example of research conducted on psychology professors, lecturers, teaching assistants, and students (see also Haller and Krauss (2002), Oakes (1986), and Gigerenzer, Krauss, and Vitouch (2004)). Subjects were given the result of a simple $t$-test of two independent means ($t = 2.7$, $df = 18$, $p = 0.01$) and were asked six true or false questions based on the result and designed to test common misinterpretations of the $p$-value. All six of the statements were false and, despite the fact that the study materials noted "several or none of the statements may be correct," (i) none of the 44 students, (ii) only four of the 39 professors and lectures who did not teach statistics, and (iii) only six of the 30 professors and lectures who did teach statistics marked all as false (members of each group marked an average of 3.5, 4.0, and 4.1 statements respectively as false).

The results reported by Gigerenzer (2004) are, unfortunately, robust. For example, Cohen (1994) reported that Oakes (1986), using the same study materials discussed above, found 68 out of 70 academic psychologists misinterpreted the $p$-value as the probability that the null hypothesis is true while 42 believed a $p$-value of 0.01 implied a 99% chance that a replication would yield a statistically significant result. Falk and Greenbaum (1995) also found similar results—despite adding the explicit option "none of these statements is correct" and requiring their subjects to read an article (Bakan 1966) warning of these misinterpretations before answering the questions. For more details and examples of these mistakes in textbooks and applied research, see Sawyer and Peter (1983), Gigerenzer (2004), and Kramer and Gigerenzer (2005).

More broadly, statisticians have long been critical of the various forms of dichotomization intrinsic to the NHST paradigm such as the dichotomy of the null hypothesis versus the alternative hypothesis and the dichotomization of results into the different categories statistically significant and not statistically significant. For example, Gelman et al. (2003) stated that the dichotomy of $\theta = \theta_0$ versus $\theta \neq \theta_0$ required by sharp point null hypothesis significance tests is an "artificial dichotomoty" and that "difficulties related to this dichotomy are widely acknowledged from all perspectives on statistical inference." More specifically, the sharp point null hypothesis of $\theta = 0$ used in the overwhelming majority of applications has long been criticized as always false—if not in theory at least in practice (Berkson 1938; Edwards, Lindman, and Savage 1963; Bakan 1966; Tukey 1991; Cohen 1994; Briggs 2016); in particular, even were an effect truly zero, experimental realities dictate that the effect would generally not be exactly zero in any study designed to test it. In addition, statisticians have noted the 0.05 threshold (or for that matter any other threshold) used to dichotomize results into statistically significant and not statistically significant is arbitrary (Fisher 1926; Yule and Kendall 1950; Cramer 1955; Cochran 1976; Cowles and Davis 1982) and thus this dichotomization has "no ontological basis" (Rosnow and Rosenthal 1989).

One consequence of this dichotomization is that applied researchers often confuse statistical significance with practical importance (P5). Freeman (1993) discussed this confusion in the analysis of clinical trials via an example of four hypothetical trials in which subjects express a preference for treatment A or treatment B. The four trials feature sequentially smaller effect sizes (preferences for treatment A of 75.0%, 57.0%, 52.3%, and 50.07% respectively) but larger sample sizes (20, 200, 2,000, and 2,000,000 respectively) such that all yield the same statistically significant $p$-value of about 0.04; the effect size in the largest study shows that the two treatments are nearly identical and thus researchers err greatly by confusing statistical significance with practical importance. Similarly, in a discussion of trials comparing subcutaneous heparin with intravenous heparin for the treatment of deep vein thrombosis, Messori, Scrocarro, and Martini (1993) stated their findings are "exactly the opposite" of those of Hommes et al. (1992) based solely on considerations relating to statistical significance that entirely ignore the similarity of the estimates of two sets of researchers (Messori, Scrocarro, and Martini (1993) estimated the odds ratio at 0.61 (95% confidence interval: 0.298–1.251), whereas Hommes et al. (1992) estimated the odds ratio at 0.62 (95% confidence interval: 0.39–0.98); for additional discussion of this example and others, see Healy (2006)).

An additional consequence of this dichotomization is that applied researchers often make scientific conclusions largely if not entirely based on whether or not a $p$-value crosses the 0.05 threshold instead of taking a more holistic view of the evidence (P3) that includes "the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis" (Wasserstein and Lazar 2016). For example, Holman et al. (2001) showed that epidemiologists incorrectly believe a result with a $p$-value below 0.05 is evidence that a relationship is causal; further, they give little to no weight to other factors such as the study design and the plausibility of the hypothesized biological mechanism.

The tendency to focus on whether or not a $p$-value crosses the 0.05 threshold rather than taking a more holistic view of the evidence has frequently led researchers astray and caused them to make rather incredible claims. For example, consider the now notorious claim that posing in open, expansive postures—so-called "power poses"—for two minutes causes changes in neuroendocrine levels, in particular increases in testosterone and

decreases in cortisol (Carney, Cuddy, and Yap 2010). The primary evidence adduced for this claim were two $p$-values that crossed the 0.05 threshold. Scant attention was given to other factors such as the design of the study (here two conditions, between-subjects), the quality of the measurements (here from saliva samples), the sample size (here 42), or potential biological pathways or mechanisms that could explain the result. Consequently, it should be unsurprising that this finding has failed to replicate (Ranehill et al. 2015; we note the first author of Carney, Cuddy, and Yap (2010) no longer believes in, studies (and discourages others from studying), teaches, or speaks to the media about these power pose effects (Carney 2016)).

As another example, consider the claim—which has been well-investigated by statisticians over the decades (Diaconis 1978; Diaconis and Graham 1981; Diaconis and Mosteller 1989; Briggs 2006) and which has surfaced again recently (Bem 2011)—that there is strong evidence for the existence of psychic powers such as extrasensory perception. Again, the primary evidence adduced for this claim were several $p$-values that crossed the 0.05 threshold and scant attention was given to other important factors. However, as Diaconis (1978) said decades ago, "The only widely respected evidence for paranormal phenomena is statistical...[but] in complex, badly controlled experiments simple chance models cannot be seriously considered as tenable explanations; hence, rejection of such models is not of particular interest."

Such incredible claims are by no means unusual in applied research—even that published in top-tier journals as were the two examples given above. However, given that the primary evidence adduced for such claims is typically one or more $p$-values that crossed the 0.05 threshold with relatively little or no attention given to other factors such as the study design, the data quality, and the plausibility of the mechanism, it should be unsurprising that support for these claims is often found to be lacking when others have attempted to replicate them or have put them to more rigorous tests (see, e.g., Open Science Collaboration 2015 and Johnson et al. 2016).

A closely related consequence of the various forms of dichotomization intrinsic to the NHST paradigm is that applied researchers tend to think of evidence in dichotomous terms (P3). For example, they interpret evidence that reaches the conventionally defined threshold for statistical significance as a demonstration of a difference and in contrast they interpret evidence that fails to reach this threshold as a demonstration of no difference. In other words, the assignment evidence to different categories induces applied researchers to conclude that the items thusly assigned are categorically different.

An example of dichotomous thinking is provided by Gelman and Stern (2006), who show applied researchers often fail to appreciate that "the difference between 'significant' and 'not significant' is not itself statistically significant." Instead, applied researchers commonly (i) report an effect for one treatment based on a $p$-value below 0.05, (ii) report no effect for another treatment based on a $p$-value above 0.05, and (iii) conclude that the two treatments are different—even when the difference between the two treatments is not itself statistically significant. In addition to the examples of this error in applied research provided by Gelman and Stern (2006), Gelman continues to document and discuss contemporary examples of this error on his blog (e.g., Blackwell, Trzesniewski, and Dweck (2007), Hu et al. (2015), Haimovitz and Dweck (2016), Pfattheicher and Schindler (2016) as well as Thorstenson, Pazda and Elliot (2015), which was retracted for this error after being discussed on the blog), while Nieuwenhuis, Forstmann, and Wagenmakers (2011) documented that it is rife in neuroscience, appearing in half of neuroscience papers in top journals such as *Nature* and *Science* in which the authors might have the opportunity to make the error.

This error has dire implications for perceptions of replication among applied researchers because the common definition of replication employed in practice is that a subsequent study successfully replicates a prior study if either both fail to attain statistical significance or both attain statistical significance and are directionally consistent. Consequently, applied researchers will often claim replication failure if a prior study attains statistical significance and a subsequent study fails to attain statistical significance—even when the two studies are themselves not statistically significantly different. This suggests that perceptions of replication failure may be overblown.

Additional examples of dichotomous thinking are provided in a series of studies conducted by McShane and Gal (2016) involving applied researchers across a wide variety of fields including medicine, epidemiology, cognitive science, psychology, business, and economics. In these studies, researchers were presented with a summary of a hypothetical experiment comparing two treatments in which the $p$-value for the comparison was manipulated to be statistically significant or not statistically significant; they were then asked questions, for example to interpret descriptions of the data presented in the summary or to make likelihood judgments (i.e., predictions) and decisions (i.e., choices) based on the data presented in the summary. The results show that applied researchers interpret $p$-values dichotomously rather than continuously, focusing solely on whether or not the $p$-value is below 0.05 rather than the magnitude of the $p$-value. Further, they fixate on $p$-values even when they are irrelevant, for example when asked about descriptive statistics. In addition, they ignore other evidence, for example the magnitude of treatment differences.

In sum, there is ample evidence that applied researchers misuse and misinterpret $p$-values in practice and that these errors directly relate to several principles articulated in the ASA Statement.

## 3. Misuse and Misinterpretation of $p$-Values by Statisticians

### 3.1. Overview

It is natural to presume that statisticians, given their advanced training and expertise, would be extremely familiar with the principles articulated in the ASA Statement. Indeed, this is reflected by the fact that the ASA Statement notes that nothing in it is new and that it is aimed at those who are not primarily statisticians. Consequently, this suggests that statisticians, in contrast to applied researchers, would be relatively unlikely to misuse and misinterpret $p$-values particularly in ways that relate to the principles articulated in the ASA Statement.

For example, perhaps dichotomous thinking and similar errors that relate to P3 are not intrinsic consequences of statistical significance and $p$-values per se but rather arise from the rote and recipe-like manner in which statistics is taught in the biomedical and social sciences and applied in academic research (Preece 1984; Cohen 1994; Gigerenzer 2004). Supporting this view, McShane and Gal (2016) found that when applied researchers were presented with not only a $p$-value but also with a posterior probability based on a noninformative prior, they were less likely to make dichotomization errors. This is interesting because objectively the posterior probability is a redundant piece of information: under a noninformative prior it is one minus half the two-sided $p$-value. While applied researchers might not consider the posterior probability unless prompted to do so or may not recognize that it is redundant with the $p$-value, statisticians can be expected to more comprehensively evaluate the informational content of a $p$-value. Thus, if rote and recipe-like training in and application of statistical methods is to blame, those deeply trained in statistics should not make these dichotomization errors.

However, by replicating the studies by McShane and Gal (2016) but using authors of articles published in this very journal as subjects, we find that expert statisticians—while less likely to make dichotomization errors than applied researchers—are nonetheless highly likely to make them. In our first study, we show that statisticians fail to identify a difference between groups when the $p$-value is above 0.05. In our second study, we show that statisticians' judgment of a difference between two treatments is disproportionately affected by whether or not the $p$-value is below 0.05 rather than the magnitude of the $p$-value; encouragingly, however, their decision-making may not be so dichotomous.

### 3.2. Study 1

*Objective:* The goal of Study 1 was to examine whether the various forms of dichotomization intrinsic to the NHST paradigm would lead even expert statisticians to engage in dichotomous thinking and thus misinterpret data. To systematically examine this question, we presented statisticians with a summary of a hypothetical study comparing two treatments in which the $p$-value for the comparison was manipulated to be statistically significant or not statistically significant and then asked them to interpret descriptions of the data presented in the summary.

*Subjects:* Subjects were the authors of articles published in the 2010–2011 volumes of the *Journal of the American Statistical Association* (*JASA*; issues 105(489)–106(496)). A link to our survey was sent via email to the 531 authors who were not personal acquaintances or colleagues of the authors; about 50 email addresses were incorrect. 117 authors responded to the survey, yielding a response rate of 24%.

*Procedure:* Subjects were asked to respond sequentially to two versions of a principal question followed by several follow-up questions. The principal question asked subjects to choose the most accurate description of the results from a study summary that showed a difference in an outcome variable associated with an intervention. Whether this difference attained ($p = 0.01$) or failed to attain ($p = 0.27$) statistical significance was manipulated within subjects.

Subjects were randomly assigned to one of four conditions following a two by two design. The first level of design varied whether subjects were presented with the $p = 0.01$ version of the question first and the $p = 0.27$ version second or whether they were presented with the $p = 0.27$ version of the question first and the $p = 0.01$ version second. The second level of the design varied the wording of the response options to test for robustness. The $p = 0.01$ version of the principal question using response wording one was as follows:

> Below is a summary of a study from an academic paper.
> The study aimed to test how different interventions might affect terminal cancer patients' survival. Subjects were randomly assigned to one of two groups. Group A was instructed to write daily about positive things they were blessed with while Group B was instructed to write daily about misfortunes that others had to endure. Subjects were then tracked until all had died. Subjects in Group A lived, on average, 8.2 months post-diagnosis whereas subjects in Group B lived, on average, 7.5 months post-diagnosis ($p = 0.01$).
> Which statement is the most accurate summary of the results?
>
> A. Speaking only of the subjects who took part in this particular study, the average number of post-diagnosis months lived by the subjects who were in Group A was *greater* than that lived by the subjects who were in Group B.
> B. Speaking only of the subjects who took part in this particular study, the average number of post-diagnosis months lived by the subjects who were in Group A was *less* than that lived by the subjects who were in Group B.
> C. Speaking only of the subjects who took part in this particular study, the average number of post-diagnosis months lived by the subjects who were in Group A was *no different* than that lived by the subjects who were in Group B.
> D. Speaking only of the subjects who took part in this particular study, it *cannot be determined* whether the average number of post-diagnosis months lived by the subjects who were in Group A was greater/no different/less than that lived by the subjects who were in Group B.

After seeing this question, each subject was asked the same question again but $p = 0.01$ was switched to $p = 0.27$ (or vice versa for the subjects in the condition that presented the $p = 0.27$ version of the question first). Response wording two was identical to response wording one above except it omitted the phrase "Speaking only of the subjects who took part in this particular study" from each of the four response options.

Subjects were then asked a series of optional follow-up questions. First, to gain insight into subjects' reasoning, subjects were asked to explain why they chose the option they chose for each of the two principle questions and were provided with a text box to do so. Next, subjects were asked a multiple choice question about their statistical model for the data which read as follows:

> Responses in the treatment and control group are often modeled as a parametric model, for example, as independent normal with two different means or independent binomial with two different proportions.
> An alternative model under the randomization assumption is a finite population model under which the permutation distribution of the conventional test statistic more or less coincides with the distribution given by the parametric model.
> Which of the following best describes your modeling assumption as you were considering the prior questions?
>
> A. I was using the parametric model.
> B. I was using the permutation model.

C. I was using some other model.
D. I was not using one specific model.

Finally, they were then asked a multiple choice question about their primary area of expertise (modeling: statistics, biostatistics, computer science, econometrics, psychometrics, etc.; substantive area: basic science, earth science, medicine, genetics, political science, etc.; or other in which case a text box was provided); a multiple choice question about their statistical approach (frequentist, Bayesian, neither, or both); a multiple choice question about how often they read Andrew Gelman's blog, which frequently discusses issues related to the dichotomization of evidence (daily; not daily but at least once a week; not weekly but at least once a month; less often than once a month; I do not read Andrew Gelman's blog but I know who he is; or I do not know who Andrew Gelman is); and a free response question asking at what $p$-value statistical significance is conventionally defined. After this, the survey terminated.

*Results:* The pattern of results was not substantially affected by the order in which the $p$-value was presented. Consequently, we collapse across both order conditions and present our results in Figure 1(a). For the principal question shown above, the correct answer is option *A* regardless of the $p$-value and the response wording: all four response options are descriptive statements and indeed the average number of post-diagnosis months lived by the subjects who were in Group A was greater than that lived by the subjects who were in Group B (i.e., $8.2 > 7.5$). However, subjects were much more likely to answer the question correctly when the $p$-value in the question was set to 0.01 than to 0.27 (84% versus 49%). Further, the response wording did not substantially affect the pattern of results.

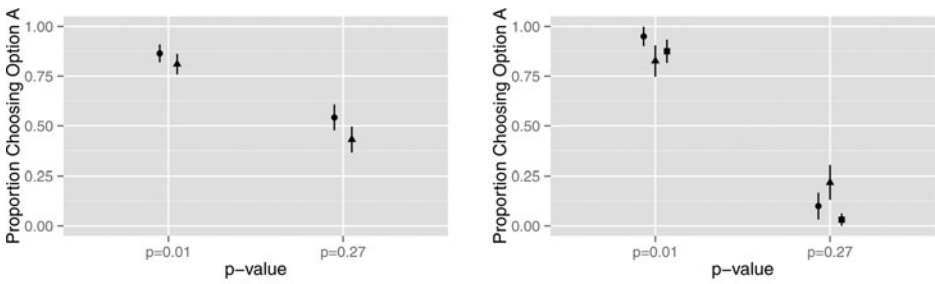
These results are striking and suggest that the dichotomization of evidence intrinsic to the NHST paradigm leads even expert statisticians to think dichotomously. In particular, about half the subjects failed to identify differences that were not statistically significant as different.

Nonetheless, as illustrated in Figure 1(b), the statisticians who were the subjects in this study performed better in this respect than the applied researchers who were the subjects in McShane and Gal (2016). Encouragingly, this suggests that a deep as opposed to cursory training in statistics that includes exposure to forms of statistical reasoning outside the NHST paradigm does help subjects focus on the descriptive nature of the question. Nonetheless, such training does not appear sufficient to entirely eliminate dichotomous thinking.

*Text Responses:* To gain additional insight into subjects' reasoning, we examined their explanations for their answers. The responses of the fifty-seven subjects who chose option *A* for the $p = 0.27$ version of the question tended to correctly identify that the question was about descriptive statistics; representative responses include: "The statement simply asked whether the average in group A was larger than group B. It was. It never asked us to conclude whether a general patient given treatment A can be expected to live longer than one given treatment B;" "The question was not about $p$-values, or inference to a larger population, it was just about the average of a set of numbers;" and "The $p$-value was irrelevant to the question and answer."

The responses of the 26 subjects who chose option *C* for the $p = 0.27$ version of the question tended to focus on statistical significance and the 0.05 threshold; representative responses include: "I based my conclusion on the observed $p$-value using the customary rule of $p < 0.05$ for a significant difference;" "The first was statistically significant and the second was not;" and "In the first question, the $p$-value is above the usual threshold. So, the difference is considered to be insignificant. In the second question, what we can say here is that the difference is statistically significant at 1% level." This was also the case of the responses of the 20 subjects who chose option *D* for the $p = 0.27$ version of the question but who did not choose option *D* for the $p = 0.01$ version of the question; representative responses include: "The result in the first question was statistically significant...for the second question, the result is not statistically significant;" "For (1) the null of equal survival can be rejected, for (2) this is not the case;" "I first looked at the different number of months for the two outcomes, then used the $p$-value to assess whether the difference was significant;" and "The $p$-value is less than 0.05 in first study."

Finally, the responses of the 14 subjects who answered option *D* to both the $p = 0.01$ and $p = 0.27$ versions of the question tended to either focus on statistical significance or emphasize additional considerations; responses representative of the former were similar to the above while responses representative of the latter include: "A $p$-value is not enough to see if the difference actually exist. Many other factors may also be important but are not available from the short story provided;" "No sample size




(a) *JASA* (b) *NEJM*

**Figure 1.** Data from Study 1 (left) and McShane and Gal (2016) Study 1 (right). Points denote $\hat{p}_A$, the proportion of subjects choosing option *A*, and lines denote $\hat{p}_A \pm \sqrt{\hat{p}_A(1-\hat{p}_A)/n}$. Response wording one is indicated by a circle, response wording two by a triangle, and response wording three (used only in McShane and Gal (2016) Study 1) by a square. Regardless of response wording, the vast majority of subjects in Study 1 correctly answered option *A* when $p = 0.01$ but only about half did when $p = 0.27$. Nonetheless, the statisticians (i.e., *JASA* authors) who were the subjects in Study 1 performed better than the applied researchers (i.e., *New England Journal of Medicine* (*NEJM*) authors) who were the subjects in McShane and Gal (2016) Study 1.

for comparison is given to see if the $p$-value is representative for first question. And no information such as demographics, medical history, and concomitant medication to see if patients' treatments are confounding with the other factors which may affect the survival.;" "I would like to make sure that the characteristics of the patients from two groups are similar (post hoc check; random assignment does not always guarantee that). Moreover, the $p$-value is not a good measure of the evidence, even if the sample sizes were known. We also need to know what the life expectancy was for each patient (without intervention...if these cancers have known history, this could be computed) and then see how different the actual life span was. We can, then, use each patient as a control for himself/herself. The information is insufficient to make a conclusion."

In sum, the text responses of the subjects who did not choose option *A* emphasized that they were thinking dichotomously in a manner consistent with the dichotomization of evidence intrinsic to the NHST paradigm.

*Additional Considerations:* One potential criticism of our findings is that we asked a trick question: our subjects clearly know that 8.2 is greater than 7.5 but perceive that asking whether 8.2 is greater than 7.5 is too trivial thereby leading them to instead answer whether or not the difference attains or fails to attain statistical significance. However, asking whether a $p$-value of 0.27 attains or fails to attain statistical significance is also trivial. Consequently, this criticism does not resolve why subjects focus on the statistical significance of the difference rather than on the difference itself. Further, we note the text responses presented above do not suggest subjects necessarily found the question too trivial.

A related potential criticism is that by including a $p$-value, we naturally led our subjects to focus on statistical significance. This is not really a criticism but rather is essentially our point: our subjects are so trained to focus on statistical significance that the mere presence of a $p$-value leads them to automatically view everything through the lens of the NHST paradigm—even when it is unwarranted.

In further response to such criticisms, we note that our response options stopped just short of explicitly telling subjects that we were asking for a description of the observed data rather than asking them to make a statistical inference. For example, in the context of the study summary "the average number of post-diagnosis months lived by the subjects who were in Group A" pretty clearly refers to the number 8.2 rather than to some hypothetical population parameter.

We also note two further points. First, even had we asked subjects to conduct a hypothesis test, option *C* is never correct: a failure to reject the null hypothesis does not imply or prove that the two treatments do not differ. Second, and again assuming we asked subjects to conduct a hypothesis test, there is a sense in which option *D* is always correct since at no particular $p$-value is the null definitively overturned. Nonetheless, 27 subjects chose option *C* for one or both versions of the question while only 14 chose option *D* for both versions.

We also analyzed data from the follow-up questions for exploratory purposes. Only four subjects reported using the permutation model justified by the randomization assumption; 30 reported using a parametric model and 67 no specific model. Eighty-five subjects reported their expertise in modeling while 48 reported taking a frequentist approach to statistics and forty both a frequentist and Bayesian approach. Unfortunately, few subjects reported being frequent readers of Andrew Gelman's blog with only one daily and two weekly readers; 47 reported not reading it at all while a further 22 reported not knowing who Gelman is. This is unfortunate as the blog often covers topics related to the dichotomization of evidence (particularly with regard to the 0.05 threshold) and we would have thus expected frequent readers to perform better on the $p = 0.27$ version of the question.

Using a parametric model seems associated with worse performance on the $p = 0.27$ version of the question: only six of the 30 subjects who reported using the parametric model chose option *A*. Further, this seems to be the only follow-up variable associated with choosing option *A* for this version of the question (none seems to be associated with choosing option *A* for the $p = 0.01$ version of the question).

### 3.3. Study 2

*Objective:* The goal of Study 2 was to examine whether the pattern of results observed in Study 1 extends from the interpretation of data to likelihood judgments (i.e., predictions) and decisions (i.e., choices) made based on data. A further goal was to examine how varying the degree to which the $p$-value is above the threshold for statistical significance affects likelihood judgments and decisions. To systematically examine these questions, we presented statisticians with a summary of a hypothetical study comparing two treatments in which the $p$-value for the comparison was manipulated to one of four values and then asked them to make likelihood judgments and decisions based on the data presented in the summary.

*Subjects:* Subjects were the authors of articles published in the 2012–2013 volumes of *JASA* (issues 107(497)–108(503)). A link to our survey was sent via email to the 565 authors who were not personal acquaintances or colleagues of the authors and who were not sent a link to Study 1; about 50 email addresses were incorrect. 140 authors responded to the survey, yielding a response rate of 27%.

*Procedure:* Subjects completed a likelihood judgment question followed by a choice question. Subjects were randomly assigned to one of four conditions that varied whether the $p$-value was set to 0.025, 0.075, 0.125, or 0.175. Subjects saw the same $p$-value in the choice question as they saw in the preceding likelihood judgment question.

The judgment question was as follows:

> Below is a summary of a study from an academic paper.
> The study aimed to test how two different drugs impact whether a patient recovers from a certain disease. Subjects were randomly drawn from a fixed population and then randomly assigned to Drug A or Drug B. Fifty-two percent (52%) of subjects who took Drug A recovered from the disease while forty-four percent (44%) of subjects who took Drug B recovered from the disease.
> A test of the null hypothesis that there is no difference between Drug A and Drug B in terms of probability of recovery from the disease yields a $p$-value of 0.025.
> Assuming no prior studies have been conducted with these drugs, which of the following statements is most accurate?

A. A person drawn randomly from the same population as the subjects in the study is *more likely* to recover from the disease if given Drug A than if given Drug B.
B. A person drawn randomly from the same population as the subjects in the study is *less likely* to recover from the disease if given Drug A than if given Drug B.
C. A person drawn randomly from the same population as the subjects in the study is *equally likely* to recover from the disease if given Drug A than if given Drug B.
D. It *cannot be determined* whether a person drawn randomly from the same population as the subjects in the study is more/less/equally likely to recover from the disease if given Drug A or if given Drug B.

After answering this judgment question, subjects were presented with the same study summary with the same $p$-value but were instead asked to make a hypothetical choice. The choice question was as follows:

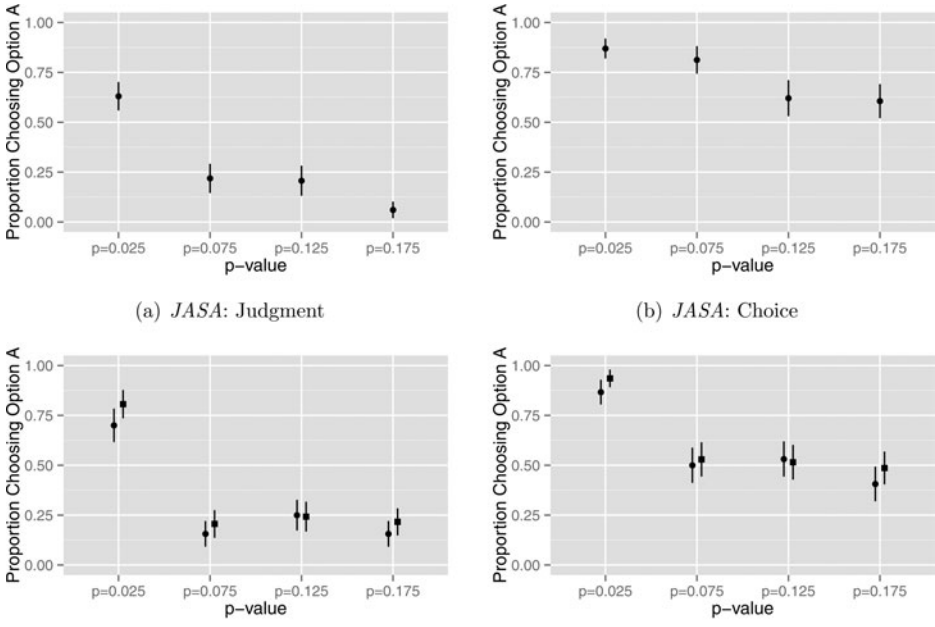
> Assuming no prior studies have been conducted with these drugs, if you were a patient from the same population as the subjects in the study, what drug would you prefer to take to maximize your chance of recovery?
> A. I prefer Drug A.
> B. I prefer Drug B.
> C. I am indifferent between Drug A and Drug B.

Subjects were then asked the same series of optional follow-up questions that were asked of subjects in Study 1.

*Results:* We present our results in Figures 2(a) and 2(b). We note that the issue at variance in both the likelihood judgment question and choice question is fundamentally a predictive one: they both ask about the relative likelihood of a new patient drawn from the subject population—whether a hypothetical one as in the likelihood judgment question or the self as in the choice question—recovering if given Drug A rather than Drug B. This in turn clearly depends on whether or not Drug A is more effective than Drug B. The $p$-value is of course one measure of the strength of the evidence regarding the likelihood that it is. However, the level of the $p$-value does not alter the correct response option for either question: the correct answer is option *A* as Drug A is more likely to be more effective than Drug B in each of the four respective $p$-value settings. Indeed, under the noninformative prior encouraged by the question wording, the probability that Drug A is more effective than Drug B is a decreasing linear function of the $p$-value (i.e., it is one minus half the two-sided $p$-value or 0.9875, 0.9625, 0.9375, and 0.9125 when the $p$-value is set respectively to 0.025, 0.075, 0.125, and 0.175).

The proportion of subjects who chose option *A* for the judgment question dropped sharply once the $p$-value rose above 0.05 but it was relatively stable thereafter (63% versus 22%, 21%, and 6%, respectively). This provides further evidence that the dichotomization of evidence intrinsic to the NHST paradigm leads even expert statisticians to think dichotomously.

In contrast, the proportion of subjects who chose Drug A for the choice question was 87%, 81%, 62%, and 61% for each of the four respective $p$-value settings. This appears best described by either a decreasing linear function of the $p$-value or a step function with a single step at a $p$-value of 0.10 or thereabouts and suggests that when it comes to making decisions—particularly personally consequential ones—expert statisticians may not dichotomize evidence (or at least may not do so around a $p$-value of 0.05).




(a) *JASA*: Judgment (b) *JASA*: Choice

(c) *AJE*: Judgment (d) *AJE*: Choice

**Figure 2.** Data from Study 2 (top) and McShane and Gal (2016) Study 2 (bottom). Points denote $\hat{p}_A$, the proportion of subjects choosing option *A*, and lines denote $\hat{p}_A \pm \sqrt{\hat{p}_A(1-\hat{p}_A)/n}$. A treatment difference of 52% versus 44% is indicated by a circle and a treatment difference of 57% versus 39% (used only in McShane and Gal (2016) Study 2) by a square. For the likelihood judgment question, the proportion of subjects in Study 2 who chose option *A* dropped sharply once the $p$-value rose above 0.05, but it was relatively stable thereafter; for the choice question, the proportion appears best described by either a decreasing linear function of the $p$-value or a step function with a single step at a $p$-value of 0.10 or thereabouts. The statisticians (i.e., *JASA* authors) who were the subjects in Study 2 performed similarly to the applied researchers (i.e., *American Journal of Epidemiology* (*AJE*) authors) who were the subjects in McShane and Gal (2016) Study 2 on the likelihood judgment question but better on the choice question.

In sum, the results of the likelihood judgment question are consistent with the results of Study 1 and the notion that the dichotomization of evidence intrinsic to the NHST paradigm leads even expert statisticians to think dichotomously. Encouragingly, they do not seem to do this for the choice question which may most realistically demonstrate how statisticians are likely to behave when making recommendations based on evidence.

As illustrated in Figures 2(c) and 2(d), the statisticians who were the subjects in this study performed similarly in this respect to the applied researchers who were the subjects in McShane and Gal (2016) on the likelihood judgment question but better on the choice question thus providing further support for the notion that a deep as opposed to cursory training in statistics that includes exposure to forms of statistical reasoning outside the NHST paradigm helps attenuate dichotomous thinking even if it cannot entirely eliminate it.

That said, given the posterior probability that Drug A was more effective than Drug B was larger than 90% in each of the four $p$-value settings, it is perhaps discouraging that nearly all statisticians did not select option *A* for both the likelihood judgment and choice questions.

*Text Responses:* To gain additional insight into subjects' reasoning, we examined their explanations for their answers. We begin by discussing the responses of subjects assigned to the $p = 0.025$ condition. Twenty-nine of these chose option *A* for the likelihood judgment question, all of whom also chose option *A* for the choice question. Responses tended to focus either on the observed differences, statistical significance, or both; representative responses include: "I chose the one with the higher probability;" "The statistical tests suggests that Drug A is significantly more efficient than Drug B;" and "The point estimate of the efficacy of Drug A (compared to Drug B) along with the corresponding $p$-value are the only information available and from that A is appears to be better. It is therefore the better bet." Among the 17 who did not choose option *A* for the likelihood judgment question, there seemed to be no systematic pattern to the responses except perhaps for a tendency to emphasize that when forced to make a choice they would choose the drug that performed better empirically.

More interesting are the responses of subjects assigned to the three conditions where the $p$-value was set above 0.05. Eleven of these chose option *A* for the likelihood judgment question, of whom only nine eleven chose option *A* for the choice question. Responses tended to focus on the observed differences; representative responses include: "You asked if it was 'more likely'; it is more likely. It's not significantly more likely, but you didn't ask this; you only asked about directionality. In Q2, you now asked my preference about the drugs. Again, even though the finding isn't statistically significant, if I were choosing the drug, I'd go with the one that had performed better;" "Because a higher percentage of the sample that took Drug A recovered than Drug B;" "As a Bayesian the higher success rate for Drug A is some evidence, even though it is not significant;" of the two subjects who curiously switched to option *C* for the choice question, only one left a text response and the response indicated confusion.

Twelve subjects chose option *C* for the likelihood judgment question, and, of these, seven switched to option *A* for the choice question while the remaining five stuck with option *C*. Responses tended to tended to focus on statistical significance and the 0.05 threshold although those who switched indicated they would lay aside concerns about statistical significance when making a choice; representative responses of two switchers versus nonswitchers respectively include: "In question one, the $p$-value is relatively large, we fail to reject $H_0$ but do not say $H_0$ is true. If we collect more samples, we may have a significant result that A is better than B. In the current situation, I choose A in the second question to maximize my chance or minimize my loss." and "The second question is conditional on me having to take one of the two." versus "The probability of recovery for the two drugs is not significantly different at level $\alpha = 0.05$." and "For the first question the $p$-value does not suggest any difference between the drugs. For the second, since no significant difference was found, I do not prefer any drug."

Sixty-seven chose option *D* for the likelihood judgment question, and, of these, 44 chose option *A* for the choice question, while the remaining 23 chose option *C*. As with those who chose option *C* for the likelihood judgment question, responses tended to tended to focus on statistical significance and the 0.05 threshold; responses of those who chose option *A* for the choice question also indicated they would lay aside concerns about statistical significance and mentioned that the posterior probability that Drug A was more effective than Drug B was above a half. Thus, representative responses were similar to those presented in the prior paragraph.

In sum, the text responses of the subjects who did not choose option *A* for the likelihood judgment question emphasized that they were thinking dichotomously in a manner consistent with the dichotomization of evidence intrinsic to the NHST paradigm but that the choice question prompted other considerations such as the observed difference and posterior probabilities.

*Additional Considerations:* One potential criticism of our findings is that there is a sense in which option *D* is the correct option for the likelihood judgment question (i.e., because at no particular $p$-value is the null hypothesis definitively overturned). More specifically, which drug is "more likely" to result in recovery depends upon the parameters governing the probability of recovery for each drug, and these parameters are unknown and unknowable under a classical frequentist interpretation of the question. However, subjects generally chose option *A* for the likelihood judgment question when the $p$-value was set below 0.05 but option *D* when it was set above 0.05 rather than option *D* regardless. Thus, this criticism does not stand.

We again analyzed data from the follow-up questions for exploratory purposes. Only seven subjects reported using the permutation model justified by the randomization assumption; 41 reported using a parametric model and 51 no specific model. Eighty-four subjects reported their expertise in modeling, while 48 reported taking a frequentist approach to statistics, 24 a Bayesian approach, and 33 both a frequentist and Bayesian approach. Unfortunately, again few subjects reported being frequent readers of Andrew Gelman's blog with only one daily and six weekly readers; 37 reported not reading it at all while a further 31 reported not knowing who Gelman is.

Curiously, those who reported taking a Bayesian approach to statistics seemed to have performed worse on the choice question when the $p$-value was set above 0.05. Further, this seems to

be the only follow up variable associated with choosing option *A* for the choice question (none seems to be associated with choosing option *A* for the likelihood judgment question).

## 4. Discussion

We have shown that even expert statisticians are sometimes prone to misuse and misinterpret *p*-values. Thus, the ASA Statement is relevant not only for those who are not primarily statisticians but also for statisticians. In particular, the principle that "Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold" (P3)—or, more poetically, as Rosnow and Rosenthal (1989) famously put it, "Surely, God loves the 0.06 nearly as much as the 0.05. Can there be any doubt that God views the strength of evidence for or against the null as a fairly continuous function of the magnitude of *p*?"—bears repetition and emphasis even among statisticians and even though there is nothing new about it.

Our most discouraging findings were (i) that about half the subjects in Study 1 failed to identify differences that were not statistically significant as different and (ii) that the vast majority of the subjects in Study 2 failed to select option *A* for both the likelihood judgment and choice question (i.e., because the posterior probability that Drug A was more effective than Drug B was larger than 90% in each of the four *p*-value settings). On the other hand, it was quite encouraging that statisticians did not seem to dichotomize evidence around the 0.05 threshold for the choice question in Study 2 as this question may most realistically demonstrate how they are likely to behave when making recommendations based on evidence. It was also encouraging—if not entirely surprising—that statisticians performed better in these studies than applied researchers as it suggests a deep as opposed to cursory training in statistics that includes exposure to forms of statistical reasoning outside the NHST paradigm can help attenuate dichotomous thinking even if it cannot entirely eliminate it.

While some may argue that the presence of a *p*-value in our questions naturally led our subjects to focus on statistical significance, we reiterate that this is not really a criticism but rather is essentially our point: our subjects are so trained to focus on statistical significance that the mere presence of a *p*-value leads them to automatically view everything through the lens of the NHST paradigm—even in cases where it is unwarranted. We further note that the text responses of our subjects emphasized that they were thinking dichotomously in a manner consistent with the dichotomization of evidence intrinsic to the NHST paradigm and that response to our principal questions did not associate particularly strongly with responses to our follow up questions.

We also note that the studies reported by McShane and Gal (2016)—while not conducted on statisticians but on applied researchers across a wide variety of fields including medicine, epidemiology, cognitive science, psychology, business, and economics—lend further support to our conclusion. For example, they show that undergraduates who have not taken a statistics course—and thus are unlikely or even unable to focus on statistical significance—perform similarly on the versions of the questions where the *p*-value is versus is not statistically significant. They also show, as discussed, that applied researchers presented with not only a *p*-value but also with a posterior probability based on a noninformative prior were less likely to make dichotomization errors. Further, they show, as illustrated in Figures 2(c) and 2(d), that applied researchers tend to ignore the magnitude of treatment differences. Finally, they also show that when subjects are asked to make a choice on behalf of a psychologically close other (i.e., a loved one) as compared to a psychologically distant other (i.e., physicians treating patients), they are more likely to choose Drug A when the *p*-value is not statistically significant; this, in combination with subjects' superior performance on the choice question as compared to the likelihood judgment question, suggests that the presence of a *p*-value may lead to dichotomous thinking by default but that other considerations (e.g., the degree to which something is personally consequential) can shift the focus away from whether a result attains or fails to attain statistical significance and toward a more holistic view of the evidence.

In addition, in a yet to be published study, when responses to the likelihood judgment question were solicited on a continuous scale rather than via a multiple choice question, applied researchers continued to interpret evidence dichotomously. In particular, when subjects were asked to rate on a one hundred point scale how confident they were that "A person drawn randomly from the same patient population as the patients in the study is more likely to recover from the disease if given Drug A than if given Drug B," the average confidence dropped precipitously as the *p*-value rose above the 0.05 threshold but did not decrease further as the *p*-value increased beyond 0.05.

Given that these findings appear quite robust, they (in particular the finding that statisticians performed better in these studies than applied researchers) naturally raise the question of what can be done in graduate training to help eliminate dichotomous thinking. Our suggestions are similar to many of those directed at applied researchers in the ASA Statement, and, like it, are not particularly new or original.

We should further expand on our efforts to emphasize that evidence lies on a continuum. For example, rather than treating effects as "real" or "not real" and statistical analysis, particularly via NHST, as the method for determining this, we should further emphasize and embrace the variation in effects and the uncertainty in our results. We may also want to consider emphasizing not only variation but also individual-level and group-level moderators of this variation that govern the generalizability of effects in other subjects and subject populations, at other times, and in different contexts. Further, as noted in the ASA Statement, we should emphasize not only statistical considerations but also take a more holistic and integrative view of evidence that includes prior and related evidence, the type of problem being evaluated, the quality of the data, the model specification, the effect size, and real world costs and benefits, and other considerations.

Perhaps most importantly we should move away from any forms of dichotomous or categorical reasoning whether in the form of NHST or otherwise (e.g., confidence intervals evaluated only on the basis of whether or not they contain zero or some other number, posterior probabilities evaluated only on the basis of whether or not they are above some particular threshold, Bayes Factors evaluated only in terms of discrete categories). While NHST clearly has its place, it also seems to be the case

that estimation (including variation and uncertainty estimation) and full decision analyses (particularly ones that account for real world costs and benefits as well as variation and uncertainty in them) are often more appropriate and fruitful in applied settings.

Moving away from graduate training of statisticians to training in statistics more broadly, Wasserstein and Lazar (2016) echo George Cobb's concern about circularity in curriculum and practice: we teach NHST because that's what the scientific community and journal editors use but they use NHST because that's what we teach them. Indeed, statistics at the undergraduate level as well as at the graduate level in applied fields is often taught in a rote and recipe-like manner that typically focuses nearly exclusively on the NHST paradigm. To be fair, statisticians are only partially at fault for this: statisticians are often not responsible for teaching statistics courses in applied fields (this is probably especially the case at the graduate level as compared to the undergraduate level) and, even when they are, institutional realities often constrain the curriculum.

The recent trend toward so-called "data science" curricula may prove helpful in facilitating a reevaluation and relaxation of these institutional constraints. In particular, it may provide statisticians with the institutional leverage necessary to move curricula away from the rote and recipe-like application of NHST in training and toward such topics as estimation, variability, and uncertainty as well as exploratory and graphical data analysis, model checking and improvement, and prediction. Further, these curricula may help facilitate a move away from point-and-click statistical software and toward scripting languages. This in and of itself is likely to encourage a more holistic view of the evidence; for example, data cleaning in a scripting language naturally prompts questions about the quality of the data and measurement while coding a model oneself increases understanding and likely promotes deeper reflection on model specification and model fit. Thus, recent developments in curricula may well help mitigate dichotomous thinking errors.

In closing, we do not believe the fault for dichotomous thinking errors shown by our subjects lies with them *per se*. Indeed, evaluating evidence under uncertainty is well-known to be quite difficult (Tversky and Kahneman 1974). Instead, we believe the various forms of dichotomization intrinsic to the NHST paradigm such as the dichotomy of the null hypothesis versus the alternative hypothesis and the dichotomization of results into the different categories statistically significant and not statistically significant almost necessarily results in some forms of dichotomous thinking: the assignment of evidence to different categories is simply just too strong an inducement to the conclusion that the items thusly assigned are categorically different—even to those who are most aware of and thus should be most resistant to this line of thinking! Thus, although statisticians and researchers more broadly are generally aware that statistical significance at the 0.05 level is a mere convention, our findings highlight that this convention strongly affects the interpretation of evidence. We thus hope that our findings will raise awareness of this phenomenon and thereby lead researchers to adopt the ASA Statement's suggestions that they take a more holistic and integrative view of evidence (and thus correspondingly reduce their reliance on statistical significance) in their interpretation of evidence and that $p$-values be supplemented, if not altogether replaced, by other approaches.

## References


Bakan, D. (1966), "The Test of Significance in Psychological Research," *Psychological Bulletin*, 66, 423–437. [886]

Bem, D. J. (2011), "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect," *Journal of Personality and Social Psychology*, 100, 407–425. [887]

Berger, J. O., and Sellke, T. (1987), "Testing a Point Null Hypothesis: The Irreconciliability of *P* Values and Evidence," *Journal of the American Statistical Association*, 82, 112–122. [885]

Berkson, J. (1938), "Some Difficulties of Interpretation Encountered in the Application of the Chi-Square Test," *Journal of the American Statistical Association*, 33, 526–536. [886]

Blackwell, L. S., Trzesniewski, K. H., and Dweck, C. S. (2007), "Implicit Theories of Intelligence Predict Achievement Across an Adolescent Transition: A Longitudinal Study and an Intervention," *Child Development*, 78, 246–263. [887]

Briggs, W. M. (2006), *So, You Think You're Psychic?* New York: Lulu. [887]

Briggs, W. M. (2016), *Uncertainty: The Soul of Modeling, Probability and Statistics*, New York: Springer. [886]

Carney, D. R. (2016), "My Position on 'Power Poses'," Technical report, Haas School of Business, University of California at Berkeley. [887]

Carney, D. R., Cuddy, A. J., and Yap, A. J. (2010), "Power Posing Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance," *Psychological Science*, 21, 1363–1368. [887]

Cochran, W. G. (1976), "Early Development of Techniques in Comparative Experimentation," in *On the History of Statistics and Probability*, ed. D. B. Owens, New York: Marcel Dekker Inc. [886]

Cohen, J. (1994), "The Earth is Round ($p < .05$)," *American Psychologist*, 49, 997–1003. [885,886,888]

Cowles, M., and Davis, C. (1982), "On the Origins of the .05 Level of Significance," *American Psychologist*, 44, 1276–1284. [886]

Cramer, H. (1955), *The Elements of Probability Theory*, New York: Wiley. [886]

Diaconis, P. (1978), "Statistical Problems in ESP Research," *Science*, 201, 131–136. [887]

Diaconis, P., and Graham, R. (1981), "The Analysis of Sequential Experiments with Feedback to Subjects," *The Annals of Statistics*, 9, 3–23. [887]

Diaconis, P., and Mosteller, F. (1989), "Methods for Studying Coincidences," *Journal of the American Statistical Association*, 84, 853–861. [887]

Edwards, W., Lindman, H., and Savage, L. J. (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70, 193. [885,886]

Falk, R., and Greenbaum, C. W. (1995), "Significance Tests Die Hard The Amazing Persistence of a Probabilistic Misconception," *Theory & Psychology*, 5, 75–98. [886]

Fisher, R. A. (1926), "The Arrangement of Field Experiments," *Journal of the Ministry of Agriculture*, 33, 503–513. [886]

Freeman, P. R. (1993), "The Role of $p$-values in Analysing Trial Results," *Statistics in Medicine*, 12, 1443–1452. [886]

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003), *Bayesian Data Analysis* (2nd ed.), Boca Raton, FL: Chapman and Hall. [886]

Gelman, A., and Stern, H. (2006), "The Difference Between 'Significant' and 'Not Significant' is not Itself Statistically Significant," *The American Statistician*, 60, 328–331. [887]

Gigerenzer, G. (2004), "Mindless Statistics," *Journal of Socio-Economics*, 33, 587–606. [886,888]

Gigerenzer, G., Krauss, S., and Vitouch, O. (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*, ed. D. Kaplan, Thousand Oaks, CA: SAGE, pp. 391–408. [886]

Haimovitz, K., and Dweck, C. S. (2016), "What Predicts Children's Fixed and Growth Intelligence Mind-Sets? Not Their Parents' Views of Intelligence but Their Parents' Views of Failure," *Psychological Science*, p. 0956797616639727. [887]

Haller, H., and Krauss, S. (2002), "Misinterpretations of Significance: A Problem Students Share with their Teachers?" *Methods of Psychological Research*, 7, available at *https://www.metheval.uni-jena.de/lehre/0405-ws/evaluationuebung/haller.pdf*. [886]
Healy, D. (2006), "The Antidepressant Tale: Figures Signifying Nothing?," *Advances in Psychiatric Treatment*, 12, 320–328. [886]
Holman, C. J., Arnold-Reed, D. E., de Klerk, N., McComb, C., and English, D. R. (2001), "A Psychometric Experiment in Causal Inference to Estimate Evidential Weights used by Epidemiologists," *Epidemiology*, 12, 246–255. [886]
Hommes, D. W., Bura, A., H. Buller, L. M., and ten Cate, J. W. (1992), "Subcutaneous Heparin Compared with Continuous Intravenous Heparin Administration in the Initial Treatment of Deep Vein Thrombosis," *Annals of Internal Medicine*, 116, 279–284. [886]
Hu, X., Antony, J. W., Creery, J. D., Vargas, I. M., Bodenhausen, G. V., and Paller, K. A. (2015), "Unlearning Implicit Social Biases During Sleep," *Science*, 348, 1013–1015. [887]
Hubbard, R., and Lindsay, R. M. (2008), "Why P Values Are Not a Useful Measure of Evidence in Statistical Significance Testing," *Theory and Psychology*, 18, 69–88. [885]
Johnson, V. E. (2013), "Uniformly Most Powerful Bayesian Tests," *Annals of Statistics*, 41, 1716–1741. [885]
Johnson, V. E., Payne, R. D., Wang, T., Asher, A., and Mandal, S. (2016), "On the Reproducibility of Psychological Science," *Journal of the American Statistical Association*, 112, 1–10. [887]
Kramer, W., and Gigerenzer, G. (2005), "How to Confuse with Statistics or: The Use and Misuse of Conditional Probabilities," *Statistical Science*, 20, 223–230. [886]
McShane, B. B., and Gal, D. (2016), "Blinding Us to the Obvious? The Effect of Statistical Training on the Evaluation of Evidence," *Management Science*, 62, 1707–1718. [887,888,889,892,893]
Messori, A., Scrocarro, G., and Martini, N. (1993), "Calculation Errors in Meta-Analysis," *Annals of Internal Medicine*, 118, 77–78. [886]
Nieuwenhuis, S., Forstmann, B. U., and Wagenmakers, E.-J. (2011), "Erroneous Analyses of Interactions in Neuroscience: A Problem of Significance," *Nature neuroscience*, 14, 1105–1107. [887]
Oakes, M. (1986), *Statistical Inference: A Commentary for the Social and Behavioral Sciences*, New York: Wiley. [886]
Open Science Collaboration (2015), "Estimating the Reproducibility of Psychological Science," *Science*, 349, aac4716. [887]
Pfattheicher, S., and Schindler, S. (2016), "Misperceiving Bullshit as Profound Is Associated with Favorable Views of Cruz, Rubio, Trump and Conservatism," *PloS one*, 11, e0153419. [887]
Preece, D. (1984), "Biometry in the Third World: Science Not Ritual," *Biometrics*, 40, 519–523. [888]
Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., and Weber, R. A. (2015), "Assessing the Robustness of Power Posing: No Effect on Hormones and Risk Tolerance in a Large Sample of Men and Women," *Psychological Science*, 26, 653–656. [887]
Rosnow, R. L., and Rosenthal, R. (1989), "Statistical Procedures and the Justification of Knowledge in Psychological Science," *American Psychologist*, 44, 1276–1284. [886,893]
Sawyer, A. G., and Peter, J. P. (1983), "The Significance of Statistical Significance Tests in Marketing Research," *Journal of Marketing Research*, 20, 122–133. [886]
Thorstenson, C. A., Pazda, A. D., and Elliot, A. J. (2015), "Sadness Impairs Color Perception," *Psychological Science*, 26, 1822–1822. [887]
Tukey, J. W. (1991), "The Philosophy of Multiple Comparisons," *Statistical Science*, 6, 100–116. [886]
Tversky, A., and Kahneman, D. (1974), "Judgment under Uncertainty: Heuristics and Biases," *Science*, 185, 1124–1131. [894]
Wasserstein, R. L., and Lazar, N. A. (2016), "The ASA's Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133. [885,886,894]
Yule, G. U., and Kendall, M. G. (1950), *An Introduction to the Theory of Statistics* (14 ed.), London: Griffin. [886]

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2017, VOL. 112, NO. 519, Applications and Case Studies
https://doi.org/10.1080/01621459.2017.1316279



# A *p*-Value to Die For

Donald Berry

Department of Biostatistics, University of Texas M. D. Anderson Cancer Center, Houston, TX



McShane and Gal expose statisticians as not understanding what is the very substance of our expertise. Only some of the "experts" failed the authors' tests. Still, such failure impugns our profession. We deserve criticism, whether the tests measure the right thing or not. We are too smug in thinking that we understand the elementary stuff. But we do not, in part because it is not elementary. And our failures are detrimental to society at large, and of course to our profession.

My commentary has two parts. One is a critique of the McShane and Gal article. The other addresses an issue regarding *p*-values that is more serious and problematic for statisticians and other scientists than the ones addressed by these authors.

McShane and Gal rail against treating evidence as binary. None of what they say is new, as they indicate. But it bears repeating. We fall prey to this yes-no silliness because many decisions are binary. But believing or advertising something as true and acting as though it is true are very different kettles of fish.

Evaluating evidence in the context of uncertainty is difficult. Communicating such evidence is more difficult yet. And there are subtleties in communicating to us about how poorly we communicate with others.

A case in point is Study 1 of McShane and Gal. They ask questions of statisticians who had published articles in *JASA*. When someone asks a question, part of the information conveyed is the fact that they asked the question. Why did they ask? To teach the respondents something? To demonstrate that they know more than the respondents? To get wrong answers so they can write an article arguing that some respondents are clueless?

**CONTACT** Donald Berry *dberry@mdanderson.org* Department of Biostatistics, University of Texas M. D. Anderson Cancer Center, 1515 Holcombe, Houston, TX 77030.

All the possible responses McShane and Gal give in the Study 1 question begin with "Speaking only of the subjects who took part in this particular study." They say

> One potential criticism of our findings is that we asked a trick question: our subjects clearly know that 8.2 is greater than 7.5 but perceive that asking whether 8.2 is greater than 7.5 is too trivial thereby leading them to instead answer whether or not the difference attains or fails to attain statistical significance.

Indeed, I wondered if the "Speaking only" question was about what statisticians usually call "the sample" and "the sample average." Or did the authors really mean to say, or imply, "Speaking only of the *population of* subjects who took part in this particular study"? If so then "average" would mean population average. Evidence for the latter interpretation is that they said "$p = 0.01$," followed by "Which statement is the most accurate summary of the results?" The "results" include the $p$-value 0.01. Why say these things if they did not mean "population"? It is confusing.

Taking this point a bit further, they write as though this simple phrase is unambiguous. It is not. A reader could easily take this as assurance from the authors that they are describing the population from which the sample was taken. For example, in interpreting the results of Study 1 they may want to clearly restrict to authors who publish in *JASA*, or those who published in *JASA* in 2010 or 2011.

Some of the respondents to Study 1 must have been confused about this point. It might partly explain the low response rate of 24%. The point is sufficiently important that I completely discount the conclusions of Study 1.

The remainder of my commentary focuses on what I regard to be the most serious and most insidious threat to the credibility and reputation of statisticians and of traditional statistical arguments.

We have saddled ourselves with perversions of logic—$p$-values—and so we deserve our collective fate. I forgive nonstatisticians who cannot provide a correct interpretation of $p < 0.05$. $p$-Values are fundamentally un-understandable. I cannot forgive statisticians who give understandable—and therefore wrong—definitions of $p$-values to their nonstatistician colleagues. But I have some sympathy for their tack. If they provide a correct definition then they will end up having to disagree with an unending sequence of "in other words." And the colleague will come away confused and thinking that statistics is nutty.

Much has been written about the misunderstandings and misinterpretations of $p$-values. The cumulative impact of such criticisms in statistical practice and on empirical research has been nil. I am witness to the collective ignorance regarding $p$-values in medicine. And I also see the herd mentality that $p < 0.05$ means true and $p > 0.05$ means not true. This mentality leads to inappropriate clinical attitudes and guidelines, and consequently to poor treatment of patients. $p$-Values are life and death quantities, and hence the title of this piece. We must have better teachers in elementary statistics courses and we must communicate better the role or nonrole of $p$-values in decision making.

What is a $p$-value? It is a statistic. Calculate the sample mean, subtract some hypothetical mean, and divide by the sample mean's standard error. Then look up the corresponding tail probability in some table, usually the standard normal. That is fine. The number of standard errors from 0 is a convenient representation of extremity for the numbers observed.

The problem arises when one attributes an inference to a $p$-value. We encourage researchers to claim statistical significance or not, implying some sort of reality, or truth. There are millions of articles in substantive fields that conclude or imply that $p < 0.05$ means the null hypothesis is wrong. So the risk factor is important, the therapy has an effect, or the biomarker is predictive of treatment response. I just Googled "$p < 0.05$ implies statistical significance" and found this garbage on the first site listed: "Most authors refer to statistically significant as $P < 0.05$ and statistically highly significant as $P < 0.001$ (less than one in a thousand chance of being wrong)." Many of these articles have statisticians as co-authors. They are our students, our colleagues, us!

Maybe we should deputize a statistical posse to ferret out drivel and label the person responsible for the drivel as *statistica non grata*.

What is the harm? Well, not much if the context is a protocol with the focus being a primary end point from the protocol and a prospective analysis of that end point. In a Bayesian perspective, this prospective analysis implies some level of prior probability associated with the alternative hypothesis. The problem occurs when there is a lot of numerical information about many variables, and the report concerns just one of those variables. That is, multiplicities (Berry 2012), traditionally called multiple comparisons. Having many dimensions and many comparisons is standard fare in biostatistics and epidemiology today. $p$-Values proliferate. And most of them are inferentially meaningless. Very few conclusions of statistical significance can be reproduced.

Principle 4 in the ASA statement (Wasserstein and Lazar 2016) is that "Proper inference requires full reporting and transparency and multiplicities." This essentially never happens.

We created a monster. And we keep feeding it, hoping that it will stop doing bad things. It is a forlorn hope. No cage can confine this monster. The only reasonable route forward is to kill it.

Inferences from an experiment depend on the data from that experiment. A $p$-value calculated from some numbers is a descriptive summary of those numbers. As such it has no inferential content. The critical issue is the interpretation of "the data" in the $p$-value definition. Inferences require a broader interpretation of "data" than one based on numbers alone. My dictionary says data are "things known or assumed as facts, making the basis of reasoning or calculation." $p$-Values calculated from numbers ignore many aspects of the evidence in the experiment at hand, including information that is quite evident. One important piece of "data" is the fact that somebody gave the statistician a spreadsheet of numbers and requested a $p$-value. What was the reason for the request? Was there something unusual about the outcomes? Sometimes the most important statistical analysis occurs before the statistician sees the numbers.

The specifics of data collection and curation and even the investigator's intentions and motivation are critical for making inferences. What has the investigator not told the statistician? Did the investigator delete some data points or experimental

units, possibly because they seemed unusual? Are some entries actually the average of two or more measurements made on the same experimental unit? If so, why were there more measurements on some units than on others? Has the investigator conducted other experiments addressing the same or related questions and decided that this was the most relevant experiment to present to the statistician? And on and on. The answers to these questions may be more important for making inferences than the numbers themselves. They set the context for properly interpreting the numerical aspects of the "data." Viewed alone, *p*-values calculated from a set of numbers and assuming a statistical model are of limited value and are frequently meaningless.

How can one incorporate the answers to questions such as those above into a statistical analysis? Standard Bayesian data-analytic measures have the same fundamental limitation as *p*-values. Subjective Bayesian approaches have some hope, but exhibiting a full likelihood function for nonquantifiable data such as previously described may be difficult or impossible. As a practical matter, when I worry that I do not know enough about the extra-numerical aspects of the "data" or about the possibility of incorporating this information into a quantitative measure of evidence then I resort to including a "black-box warning" in the publication:

> Our study is exploratory and we make no claims for generalizability. Statistical calculations such as *p*-values and confidence intervals are descriptive only and have no inferential content.

## References

Berry, D. A. (2012), "Multiplicities in Cancer Research: Ubiquitous and Necessary Evils," *Journal of the National Cancer Institute*, 104, 1125–1133. [896]

Wasserstein, R. L., and Lazar, N. A. (2016), "The ASA's Statement on *p*-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133. [896]

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2017, VOL. 112, NO. 519, Applications and Case Studies
https://doi.org/10.1080/01621459.2017.1311264



# The Substitute for *p*-Values

William M. Briggs

New York, NY

**ABSTRACT**

If it was not obvious before, after reading McShane and Gal, the conclusion is that *p*-values should be proscribed. There are no good uses for them; indeed, every use either violates frequentist theory, is fallacious, or is based on a misunderstanding. A replacement for *p*-values is suggested, based on predictive models.

**ARTICLE HISTORY**
Received March 2017
Revised March 2017

**KEYWORDS**
Cause; Model selection; *p*-Values; Philosophy of probability; Variable selection



There are no good reasons nor good ways to use *p*-values. They should be retired forthwith. The reasons for this are many, and most are well explicated in McShane and Gal's fine article. It is well to complain about something broken; it is better to provide a solution. This I will do, while highlighting philosophical deficiencies and fallacies of *p*-value-driven NHST.

A person is interested in a probability model. But guided by the philosophy of *p*-values, he asks no questions about this model, and instead asks what is the probability, given the data and some other model, which is *not* the model of interest, of seeing an ad hoc statistic larger than some value. (*Any* change in a model produces a different model.) Since there are an infinite number of models that are not the model of interest, and since there are an infinite number of statistics, the creation of *p*-values can go on forever. Yet none have *anything* to say about the model of interest.

Why? Fisher (1970) said: "Belief in null hypothesis as an accurate representation of the population sampled is confronted by a logical disjunction: *Either* the null is false, *or* the *p*-value has attained by chance an exceptionally low value."

Fisher's "logical disjunction" is evidently not one, since the either-or describes different propositions. A real disjunction can however be found: *Either* the null is false and we see a small *p*-value, *or* the null is true and we see a small *p*-value. Or just: *Either* the null is true or it is false and we see a small *p*-value. Since "*Either* the null is true or it is false" is a tautology, and is therefore necessarily true, we are left with, "We see a small *p*-value." The *p*-value casts no light on the truth or falsity of the null.

Frequentist theory claims, assuming the truth of the null, we can equally likely see any *p*-value whatsoever. And since we always do (see any value), all *p*-values are logically evidence *for*

**CONTACT** William M. Briggs matt@wmbriggs.com 340 E. 64st Apt. 9A, New York, NY 10065.

the null and not against it. Yet practice insists small $p$-value is evidence the null is (likely) false. That is because people argue: For most small $p$-values I have seen in the past, the null has been false; I now see a new small $p$-value, therefore the null hypothesis in this new problem is likely false. That argument works, but it has no place in frequentist theory (which anyway has innumerable other difficulties).

Any use of $p$-values in deciding model truth thus involves a fallacy or misunderstanding. This is formally proven by Briggs (2016, chap. 9), a work which I draw from to suggest a replacement for $p$-values, which is this. Clients ask, "What's the probability that if I know $X$, $Y$ will be true?" Instead of telling them that, we give them $p$-values. This is like a customer asking for a Cadillac and being given a broken-down rickshaw without wheels. Why not give customers what they want?

Without elaborating the mathematical details, which will anyway be obvious to readers, here is the scheme in simplified form. We have interest in proposition $Y$, which might be "This patient gets better." We want the probability $Y$ is true given we know $X_0 =$ "The patient will be treated by the usual protocol" or $X_1 =$ "The patient will be treated by the *New & Improved!* protocol." We have a collection of observations $D$ detailing where patients improved or not and which protocol they received. We want

$$\Pr(Y|X_i D). \tag{1}$$

This could be deduced in many cases using finite, discrete probability, but that is hard work; instead, a probability model relating $D$ and $X$ to $Y$ is proposed. This model will be parameterized with continuous-valued parameters. Since all observations are finite and discrete, this model will be an approximation, though it can be an excellent one. The parameters are, of course, of no interest whatsoever to man or beast; they serve only to make the model function. They are a nuisance and no help in answering the question of interest, so they are "integrated out." The end result is this:

$$\Pr(Y|X_i DM), \tag{2}$$

where $M$ is a complicated proposition that gives details about the model proposed by the statistician. This is recognized as the predictive posterior distribution given $M$ (e.g., Bernardo and Smith 2000). $M$ thus also contains assumptions made about the approximate parameters, that is, whether to use "flat" priors and so on.

This form has enormous benefits. It is in plain language; specialized training is not need to grasp model statements, though advanced work (or better software) is needed to implement (2). Everything is put in terms of observables. The model is also made prominent, in the sense that it is plain there is a specific probability model with definite assumptions in use, and thus it is clear that answers will be different if a different model or different assumptions about that model are used ("maxent" priors versus "flat," say).

*Anybody* can check (2)'s predictions, even if they do not know $D$ or $M$'s details. Given $M$ and $D$, authors might claim there is a 55% chance Y is true under the new protocol. Any reader can verify whether this prediction is useful for him or not, whether the predictions are calibrated, etc. We do not have to take authors at their word about what they discovered. Note: because finding wee $p$-values is trivial, many "novel" theories vanish under (2).

A prime reason $p$-values were embraced was that they made automatic, *universal* decisions about whether to "drop" variables or to keep them (in a given model schema). But probability is not decision; $p$-values conflated the concepts. $p$-Values cannot discover cause.

There are an infinite number of "variables" (observations, assumptions, information, premises, etc.) that can be added to the right-hand-side of (2). In our example, these can be anything—they can always be anything!—from a measure of hospital cleanliness to physician sock color to the number of three-teated cows in Cleveland. The list really is endless. Each time one is put into or removed from (2), the probability changes. Which list of variables is correct? They all are. This is true because all probability is conditional: there is no such thing as unconditional probability (this is also proven by Briggs 2016).

The goal of all modeling is to find a list of true premises (which might include data, etc.), which allow us to determine or know the cause of $Y$. This list (call it $C$) will give extreme probabilities in (2), that is,

$$\Pr(Y|X_i DC) = 0 \text{ or } 1. \tag{3}$$

Note that *to determine* and *to cause* are not the same; the former means *to ascertain*, while the latter is more complex. Readers generally think of efficient causes, and that is enough for here, though these comprise only one aspect of cause. (Because of underdetermination, $C$ is also not unique.) Discovering cause is rare because of the complexity of $C$ (think of the myriad causes of patient improvement). It is still true that the probabilities in (2) are correct when $M \not\equiv C$, for they are calculated based on different assumptions.

What goes into $M$? Suppose (observation, assumption, etc.) $W$ is considered. The (conditional) probability of $Y$ with and without $W$ in (2) is found; if these differ such that the model user would make different decisions, $W$ is kept; else not. Since decision is relative there is thus *no* universal solution to variable selection. A model or variable important to one person can be irrelevant to another. Since model creators always had infinite choice, this was always obvious.

## References

Bernardo, J. M., and Smith, A. F. M. (2000), *Bayesian Theory*, New York: Wiley. [898]

Briggs, W. M. (2016), *Uncertainty: The Soul of Probability, Modeling & Statistics*, New York: Springer. [898]

Fisher, R. (1970), *Statistical Methods for Research Workers* (14th ed.), Edinburgh, UK: Oliver and Boyd. [897]

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2017, VOL. 112, NO. 519, Applications and Case Studies
https://doi.org/10.1080/01621459.2017.1311263



# Some Natural Solutions to the *p*-Value Communication Problem—and Why They Won't Work

Andrew Gelman[a] and John Carlin[b]

[a]Department of Statistics and Department of Political Science, Columbia University, New York; [b]Clinical Epidemiology and Biostatistics, Murdoch Children's Research Institute, and Centre for Epidemiology & Biostatistics, University of Melbourne, Parkville, Victoria, Australia



## 1. Significance Testing in Crisis

It is well known that even experienced scientists routinely misinterpret *p*-values in all sorts of ways, including confusion of statistical and practical significance, treating nonrejection as acceptance of the null hypothesis, and interpreting the *p*-value as some sort of replication probability or as the posterior probability that the null hypothesis is true.

A common conceptual error is that researchers take the rejection of a straw-man null as evidence in favor of their preferred alternative (Gelman 2014). A standard mode of operation goes like this: $p < 0.05$ is taken as strong evidence against the null hypothesis, $p > 0.15$ is taken as evidence in favor of the null, and $p$ near 0.10 is taken either as weak evidence for an effect or as evidence of a weak effect.

Unfortunately, none of those inferences is generally appropriate: a low *p*-value is not necessarily strong evidence against the null (see, e.g., Morris 1987; Gelman and Carlin 2014), a high *p*-value does not necessarily favor the null (the strength and even the direction of the evidence depends on the alternative hypotheses), and *p*-values are in general not measures of the size of any underlying effect. But these errors persist, reflecting (a) inherent difficulties in the mathematics and logic of *p*-values, and (b) the desire of researchers to draw strong conclusions from their data.

Continued evidence of these and other misconceptions and their dire consequences for science (the "replication crisis" in psychology, biology, and other applied fields), especially in light of new understanding of how common it is that abundant "researcher degrees of freedom" (Simmons, Nelson, and Simonsohn 2011) and "gardens of forking paths" (Gelman and Loken 2014) allow researchers to routinely obtain statistically significant and publishable results from noise, motivated the American Statistical Association to release a Statement on Statistical Significance and *p*-values in an attempt to highlight the magnitude and importance of problems with current standard practice (Wasserstein and Lazar 2016).

At this point, it would be natural for statisticians to think that this is a problem of education and communication. If we could just add a few more paragraphs to the relevant sections of our textbooks, and persuade applied practitioners to consult more with statisticians, then all would be well, or so goes this logic.

In their article, McShane and Gal present survey data showing that even authors of published articles in a top statistics journal are often confused about the meaning of *p*-values, especially by treating 0.05, or the range 0.05–0.15, as the location of a threshold. The underlying problem seems to be deterministic thinking. To put it another way, applied researchers and also statisticians are in the habit of demanding more certainty than their data can legitimately supply. The problem is not just that 0.05 is an arbitrary convention; rather, even a seemingly wide range of *p*-values such as 0.01–0.10 cannot serve to classify evidence in the desired way (Gelman and Stern 2006).

It is shocking that these errors seem so hard-wired into statisticians' thinking, and this suggests that our profession really needs to look at how it teaches the interpretation of statistical inferences. The problem does not seem just to be technical misunderstandings; rather, statistical analysis is being asked to do something that it simply cannot do, to bring out a signal from any data, no matter how noisy. We suspect that, to make progress in pedagogy, statisticians will have to give up some of the claims we have implicitly been making about the effectiveness of our methods.

## 2. Some Natural Solutions That Won't, on Their Own, Work

### 2.1. Listen to the Statisticians, or Clarity in Exposition

It would be nice if the statistics profession was offering a good solution to the significance testing problem and we just needed to convey it more clearly. But, no, as McShane and Gal reveal, many statisticians misunderstand the core ideas too. It might be a good idea for other reasons to recommend that students take more statistics classes—but this would not solve the problems if textbooks point in the wrong direction and instructors do not understand what they are teaching. To put it another way, it is not that we are teaching the right thing poorly; unfortunately, we have been teaching the wrong thing all too well.

This is one of the difficulties we had with the American Statistical Association's statement on *p*-values: the statistics profession has been spending decades selling people on the idea of statistics as a tool for extracting signal from noise, and

**CONTACT** Andrew Gelman gelman@stat.columbia.edu Department of Statistics and Department of Political Science, Columbia University, 1016 Social Work Building, New York, NY 10027.

our journals and textbooks are full of triumphant examples of learning through statistical significance; so it is not clear why we as a profession should be trusted going forward, at least not until we take some responsibility for the mess we have helped to create.

### 2.2. Confidence Intervals Instead of Hypothesis Tests

A standard use of a confidence interval is to check whether it excludes zero. In this case, it is a hypothesis test under another name.

Another use is to consider the interval as a statement about uncertainty in a parameter estimate. But this can give nonsensical answers, not just in weird trick problems but for real applications. For example, Griskevus et al. (2014) used data from a small survey to estimate that single women were 20 percentage points more likely to vote for Barack Obama during certain days in their monthly cycle. This estimate was statistically significant with a reported standard error of 8 percentage points; thus the classical 95% interval for the effect size was (4%, 36%), an interval that makes no sense on either end! Even an effect of 4% is implausible given what we know about the low rate of opinion change during presidential election campaigns (e.g., Gelman et al. 2016)—and it would certainly be a mistake to use this survey to rule out zero or small negative net effects.

So, although confidence intervals contain some information beyond that in *p*-values, they do not resolve the larger problems that arise from attempting to get near-certainty out of noisy estimates.

### 2.3. Bayesian Interpretation of One-Sided p-Values

Consider a parameter estimate that is greater than zero and whose statistical significance is being assessed using a *p*-value. Under a symmetric continuous model such as the normal distribution, the one-sided *p*-value or tail-area probability is identical to the posterior probability that the parameter of interest is negative, given the data and a uniform prior distribution. This mathematical identity has led Greenland and Poole (2013) to suggest that "*p* values can be incorporated into a modern analysis framework that emphasizes measurement of fit, distance, and posterior probability in place of 'statistical significance' and accept/reject decisions." We agree with that last bit about moving away from binary decisions but, as Greenland and Poole (2013) note, the Bayesian interpretation of the *p*-value is not particularly helpful except as some sort of bound.

The problem comes with the uniform prior distribution. We tend to be most concerned with overinterpretation of statistical significance in problems where underlying effects are small and variation is high, and in these settings the use of classical inferences—or their flat-prior Bayesian equivalents—will lead to systematic overestimation of effect sizes and over-certainty regarding their signs: high type M and type S errors, in the terminology of Gelman and Tuerlinckx (2000) and Gelman and Carlin (2014). We do *not* consider it reasonable in general to interpret a *z*-statistic of 1.96 as implying a 97.5% chance that the corresponding estimate is in the right direction.

### 2.4. Focusing on "Practical Significance" Instead of "Statistical Significance"

Realistically, all statistical hypotheses are false: effects are not exactly zero, groups are not exactly identical, distributions are not really normal, measurements are not quite unbiased, and so on. Thus, with enough data it should be possible to reject any hypothesis. It is a commonplace among statisticians that a $\chi^2$ test (and, really, any *p*-value) can be viewed as a crude measure of sample size, and this can be framed as the distinction between practical and statistical significance, as can be illustrated with a hypothetical large study in which an anti-hypertension drug is found to reduce blood pressure by 0.3 mmHg with a standard error of 0.1. This estimate is clearly statistically significantly different from zero but is tiny on a substantive scale.

So, in a huge study, comparisons can be statistically significant without having any practical importance. Or, as we would prefer to put it, effects can vary: +0.3 for one group in one scenario might become −0.2 for a different group in a different situation. Tiny effects are not only possibly trivial, they can also be unstable, so that for future purposes an estimate of $0.3 \pm 0.1$ might not even be so likely to remain positive. To put it another way, the characterization of an effect as "small" or "not of practical significance" is relative to some prior understanding of underlying variation.

That said, the distinction between practical and statistical significance does *not* resolve the difficulties with *p*-values. The problem is not so much with large samples and tiny but precisely measured effects but rather with the opposite: large effect-size estimates that are hopelessly contaminated with noise. Consider an estimate of 30 with standard error 10, of an underlying effect that cannot realistically be much larger than 1. In this case, the estimate is statistically significant and also practically significant but is essentially entirely the product of noise. This problem is central to the recent replication crisis in science (see Button et al. 2013; Loken and Gelman 2017) but is not at all touched by concerns of practical significance.

### 2.5. Bayes Factors

Another direction for reform is to preserve the idea of hypothesis testing but to abandon tail-area probabilities (*p*-values) and instead summarize inference by the posterior probabilities of the null and alternative models, a method associated with Jeffreys (1961) and discussed recently by Rouder et al. (2009). The difficulty of this approach is that the marginal likelihoods of the separate models (and thus the Bayes factor and the corresponding posterior probabilities) depend crucially on aspects of the prior distribution that are typically assigned in a completely arbitrary manner by users. For example, consider a problem where a parameter has been assigned a normal prior distribution with center 0 and scale 10, and where its estimate is likely to be in the range (−1, 1). The chosen prior is then essentially flat, as would also be the case if the scale were increased to 100 or 1000. But such a change would divide the Bayes factor by 10 or 100.

Beyond this technical criticism, which is explored further by Gelman and Rubin (1995) and Gelman et al. (2013, chap. 8), the use of Bayes factors for hypothesis testing is also subject to many of the problems of *p*-values when used for that same purpose

and which are discussed by McShane and Gal: the temptation to discretize continuous evidence and to declare victory from the rejection of a point null hypothesis that in most cases cannot possibly be true.

## 3. Where Next?

Our own preferred replacement for hypothesis testing and *p*-values is model expansion and Bayesian inference, addressing concerns of multiple comparisons using hierarchical modeling (Gelman, Hill, and Yajima 2012) or through non-Bayesian regularization techniques such as lasso (Lockhart et al. 2013). The general idea is to use Bayesian or regularized inference as a replacement of hypothesis tests but in the manner of Kruschke (2013), through estimation of continuous parameters rather than by trying to assess the probability of a point null hypothesis. And, as we discuss in Sections 2.2–2.4 above, informative priors can be crucial in getting this to work. Indeed, in many contexts it is the prior information rather than the Bayesian machinery that is the most important. Non-Bayesian methods can also incorporate prior information in the form of postulated effect sizes in post-data design calculations (Gelman and Carlin 2014).

In short, we would prefer to avoid hypothesis testing entirely and just perform inference using larger, more informative models.

To stop there, though, would be to deny one of the central goals of statistical science. As Morey et al. (2014) wrote, "Scientific research is often driven by theories that unify diverse observations and make clear predictions. ...Testing a theory requires testing hypotheses that are consequences of the theory, but unfortunately, this is not as simple as looking at the data to see whether they are consistent with the theory." To put it in other words, there is a demand for hypothesis testing. We can shout till our throats are sore that rejection of the null should not imply the acceptance of the alternative, but acceptance of the alternative is what many people want to hear. There is a larger problem of statistical pedagogy associating very specific statistical "hypotheses" with scientific hypotheses and theories, which are nearly always open-ended.

As we wrote in response to the ASA's much-publicized statement from last year, we think the solution is not to reform *p*-values or to replace them with some other statistical summary or threshold, but rather to move toward a greater acceptance of uncertainty and embracing of variation (Carlin 2016; Gelman 2016).

We will end not on this grand vision but on an emphasis on some small steps that we hope can make a difference. If we do the little things right, those of us who write textbooks can then convey some of this sensibility into our writings.

To start with, we recommend saying No to binary conclusions in our collaboration and consulting projects: resist giving clean answers when that is not warranted by the data. Instead, do the work to present statistical conclusions with uncertainty rather than as dichotomies. Also, remember that most effects cannot be zero (at least in social science and public health), and that an "effect" is usually a mean in a population (or something similar such as a regression coefficient)—a fact that seems to be lost from consciousness when researchers slip into binary statements about there being "an effect" or "no effect" as if they are writing about constants of nature. Again, it will be difficult to resolve many problems with *p*-values and "statistical significance" without addressing the mistaken goal of certainty which such methods have been used to pursue.

## References


Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafo, M. R. (2013), "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience," *Nature Reviews Neuroscience*, 14, 365–376. [900]

Carlin, J. B. (2016), "Is Reform Possible Without a Paradigm Shift?" *The American Statistician*. [901]

Gelman, A. (2014), "Confirmationist and Falsificationist Paradigms of Science," *Statistical Modeling, Causal Inference, and Social Science Blog*, 5 Sept. Available at *http://andrewgelman.com/2014/09/05/confirmationist-falsificationist-paradigms-science/* [899]

——— (2016), "The Problems With p-Values are Not Just With p-Values," *The American Statistician*, Available at *http://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108* [901]

Gelman, A., and Carlin, J. B. (2014), "Beyond Power Calculations: Assessing Type S (sign) and Type M (Magnitude) Errors," *Perspectives on Psychological Science*, 9, 641–651. [899,900,901]

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian Data Analysis* (3rd ed.), London: Chapman and Hall. [900]

Gelman, A., Goel, S., Rivers, D., and Rothschild, D. (2016), "The Mythical Swing Voter," *Quarterly Journal of Political Science*, 11, 103–130. [900]

Gelman, A., Hill, J., and Yajima, M. (2012), "Why We (Usually) Don't Have to Worry About Multiple Comparisons," *Journal of Research on Educational Effectiveness*, 5, 189–211. [901]

Gelman, A., and Loken, E. (2014), "The Statistical Crisis in Science," *American Scientist*, 102, 460–465. [899]

Gelman, A., and Rubin, D. B. (1995), "Avoiding Model Selection in Bayesian Social Research," *Sociological Methodology*, 25, 165–173. [900]

Greenland, S., and Poole, C. (2013), "Living With *P*-Values: Resurrecting a Bayesian Perspective on Frequentist Statistics," *Epidemiology*, 24, 62–68. [900]

Jeffreys, H. (1961), *Theory of Probability* (3rd ed.), Oxford, UK: Oxford University Press. [900]

Kruschke, J. K. (2013), "Bayesian Estimation Supersedes the *t* Test," *Journal of Experimental Psychology: General*, 142, 573–603. [901]

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2013), "A Significance Test for the Lasso," Technical Report, Department of Statistics, Stanford University. [901]

Loken, E., and Gelman, A. (2017), "Measurement Error and the Replication Crisis," *Science*, 355, 584–585. [900]

Morey, R. D., Rouder, J. N., Verhagen, J., and Wagenmakers, E. J. (2014), "Why Hypothesis Tests are Essential for Psychological Science," *Psychological Science*, 25, 1289–1290. [901]

Morris, C. N. (1987), "Comment," *Journal of the American Statistical Association*, 82, 131–133. [899]

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009), "Bayesian *t* Tests for Accepting and Rejecting the Null Hypothesis," *Psychonomic Bulletin & Review*, 16, 225–237. [900]

Simmons, J., Nelson, L., and Simonsohn, U. (2011), "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allow Presenting Anything as Significant," *Psychological Science*, 22, 1359–1366. [899]

Wasserstein, R. L., and Lazar, N. A. (2016), "The ASA's Statement on *p*-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133. [899]

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2017, VOL. 112, NO. 519, Applications and Case Studies
https://doi.org/10.1080/01621459.2017.1311265



# Statistical Significance and the Dichotomization of Evidence: The Relevance of the ASA Statement on Statistical Significance and p-Values for Statisticians

Eric B. Laber[a] and Kerby Shedden[b]

[a]Department of Statistics, NC State University, Raleigh, NC; [b]Department of Statistics, University of Michigan, Ann Arbor, MI



## 1. Introduction

Empirical efforts to document the practices and thought processes of data analysis are a promising way to understanding the real-world impact of statistical methodology. The empirical findings of McShane and Gal provide new insights into long-standing debates about dichotomization and NHST. We appreciate having the opportunity to comment on this important work. Our discussion is divided into two sections. First, we comment narrowly on the new empirical findings. Second, we discuss in broader terms the interpretations drawn from these studies, and present some of our own points of view about $p$-values, dichotomization, and NHST in applied research.

## 2. The McShane and Gal Empirical Studies

The principle argument of studies 1 and 2 is that statistical researchers endorse incorrect statistical statements about an applied study's findings, and more critically, that these incorrect assertions disproportionately arise when the evidence level crosses the $p = 0.05$ threshold. We especially appreciate the uniqueness and novelty of the latter finding. We do however have some reservations about the context in which these findings were elicited, and would be interested to see if they persist elsewhere.

A notable aspect of study 1 is that the correct line of reasoning in all cases is to simply state the numerical characteristics of a sample, avoiding any consideration of whether the results generalize to a population. While a literal reading of the study questions supports this approach, commenting only on the sample runs against our usual practice, especially when assessing a randomized trial. This presents each participant with a dilemma—can the question truly mean what it seems to say, given that there is almost no conceivable setting in which the stated fact (that the numerical means of the survival times differ between the groups) would provide a justified basis for action?

Challenges in communication are particularly prominent when we aim to maintain a sharp distinction between the sample and the population. Many applied researchers choose to speak primarily about their data, but usually recognize that there is a difference between their data and "the truth." Statisticians have extensive experience with the rhetorical means of maintaining a distinction between the sample and population. Effective collaboration may require us to take a somewhat figurative view of the language employed by applied researchers. Even expressions such as "speaking only of the subjects in this study," which reads as an unmistakable cue once we are aware of the study's intentions, can have a different impact if read slightly less literally. Acknowledging that McShane and Gal asked the participants for their own interpretation of the hypothetical research findings, the reality is that we are constantly engaged in interactions with applied researchers, and these interactions impact our communication. Thinking of our many discussions with applied researchers, "speaking only of the subjects in this study" could be taken to indicate that the *inferences* are to be made based only on the data observed in this study, not, for example, using other studies of the same treatment.

Study 2 raises different, more subtle issues. As written, question 1 appears to ask about the values of unknown population parameters, whereas the intended focus may have been the direction of evidence in the observed data. Under the former interpretation, the correct answer is arguably 'd' (cannot be determined). The authors reason that because respondents were more likely to select option 'd' if the $p$-value was above 0.05, respondents were interpreting the question as intended, and their answers are evidence of dichotomous thinking. An alternative explanation is that respondents first noted that the question, as stated, could not be answered using the observed data and therefore opted to "read between the lines" and answer the question they felt the investigators meant to ask. Statisticians often have the role of properly qualifying study findings that are imprecisely reported, for example, by deeming a result statistically significant or not. It seems conceivable that some respondents may have opted to answer a question about the statistical significance of the findings rather than the direction of evidence.

While we appreciate the importance of using terminology appropriately, and share to some extent the authors' dismay at the number of research statisticians who "failed" these tests, it is not clear to us that statistics as a discipline will have its greatest possible positive impact if we default to taking the most literal interpretation of statements made by researchers about their data, especially when such statements are at odds with the best practices of our field. Fortunately, in the real world we are rarely placed in a position where our response is limited to a list of pre-defined choices. By engaging researchers in a discussion about their data and the scientific context of their research, misstatements relating to uncertainty and evidence can usually be avoided.

**CONTACT** Eric B. Laber eblaber@ncsu.edu North Carolina State University, Department of Statistics, 2311 Stinson Drive, Campus Box 8203, Raleigh, NC 27695-8203.

## 3. Varying Points of View About NHST and Dichotomization of Evidence

### 3.1. Holistic Interpretation of Evidence

A consistent theme in McShane and Gal's article is that when interpreting statistical evidence, one should take "a more holistic and integrative view," suggesting that additional factors to consider should include "prior and related evidence," the "type of problem being evaluated," the "quality of the data," the model specification, the effect size, and real world costs and benefits. However even when taking an integrated view there will be pressure to make a binary decision to accept or reject a study's claims, and inevitably there will be near hits and near misses. Thus, while holistic interpretation is a laudable practice, we view this as a largely distinct issue from the problem of dichotomization.

Practices vary among scientific disciplines, here we speak of our own experiences. In years of working with life science researchers, especially around investigations of the molecular mechanisms of human diseases, we have noted that most scientists are intensely concerned about many of the evidentiary factors cited above. The relationship between novel and prior work, formulation of the hypotheses, and especially issues of data quality are all of great concern to many scientists. While statisticians are usually the first voice in the room when discussing statistical uncertainty, unfortunately we are too often excluded (or self-exclude) from discussions of other important aspects of data analysis, perhaps because these discussions tend to involve more specialized aspects of the subject area.

We have the following concern: if we raise doubts about the value of our narrow contributions relating to formalized statistical inference, and at the same time fail to engage seriously in other aspects of research, statisticians will lose a great deal of hard-won standing. It is appropriate to flag misuse of NHST and sometimes to counsel against inappropriate dichotomous thinking. At the same time, we need to intensively seek out other contributions we can make to the practice of data-driven research, and to train the next generation of statistical researchers to think beyond stylized hypothesis testing.

### 3.2. Continuous Evidence and Actionability

There has been a great deal of discussion over the years about deficiencies of various statistical frameworks, but we do not believe any existing framework performs so flawlessly that it automates the process of reasoning with uncertainty. The specific emphasis of McShane's and Gal's argument is the notion of "dichotomous thinking," specifically using $p$-values and arbitrary thresholds to make binary decisions. McShane and Gal encouraged us to think continuously about evidence. We are strongly in agreement to the extent that evidence does, nearly always, arrive to us in a continuous form. Many features of data or of a model are not of decisive importance, and can be presented simply as an estimate with standard error or other uncertainty measure. Nevertheless, decisions do arise that cannot be made continuously. When making a binary decision, there will inevitably be near-hits and near misses. This arbitrariness is inevitable in any setting where discrete actions must be taken.

One often-proposed way to resolve this difficulty is to work from a cost-based perspective, in which the decision, while (often) binary, is based on both the weight of evidence, and the costs of the two types of errors that can be committed. Ultimately this is still a binary decision, albeit using a threshold that is adapted to the context of the problem. Again there will be near hits and near misses, even when costs are taken into account. Furthermore, the "crisis of reproducibility" in science has most often been discussed in the context of basic research. In that setting, what is the cost of presenting a result that is not true, or that is only true in limited and difficult-to-replicate settings?

### 3.3. Adaptability of p-Values and NHST

It is notable how novel and sophisticated developments and extensions of the NHST framework continue to regularly arise. For example, the rapidly growing toolbox of false discovery rate (FDR) methods has in our view been very successful at addressing concerns about multiple hypothesis testing and inference for exploratory analysis, in spite of being built on the binary notion of discoveries being either "false" or "true." Along these lines, the work by Efron (2004) and others on empirical null distributions, and the recent "knockoff" approach by Candès, and Barber (2015) are elegant and powerful approaches to inference that rest on the NHST framework.

In our view, most failures of statistical inference result from poor understanding of the sources of variation in the systems being studied, not from generic failures of inferential tools. Insights from domain-specific research including "cryptic relatedness" (Pritchard and Voight 2005) and other forms of population structure in genetics, subtle placebo effects in clinical research (Howick et al. 2013), batch effects in genomic research (Leek et al. 2010), and false positives deriving from complex spatial noise in brain imaging (Knutsson, Eklund, and Nichols 2016) provided us with a mechanistic basis for understanding previous inference failures. These insights do not mainly posit a failure of methods or of practitioners, but rather advance novel and fundamental mechanisms of variation that clarify the basis for past failures of NHST. Arguably, each of these mechanistic factors would affect "dichotomous" or "continuous" reasoning in statistical inference to similar degrees.

The most salient critique of NHST, in our view, is that rejecting a "straw man" null hypothesis resting on a simplistic model does not provide much evidence in favor of any particular alternative. But when the default "null model" is a rich and complex model, fit using efficient methods to large and carefully modeled data, a NHST targeted to the effects of interest can become quite compelling. As a case in point, in linguistics, there has been much discussion lately about specification of mixed effects models, with one community suggesting to take the "maximal random effects structure" (Barr 2017), meaning that every plausible random effect should be included. This nearly saturates the correlation model, with the view being that any parameter contrasts that appear strong against this correlational backdrop stand a good chance of being real.

There is good reason to be optimistic about the future of statistical inference as a relevant tool for discovery in science,

including frequentist and NHST-based inference. As has been intensively discussed elsewhere, we are likely to be increasingly working with extensive volumes of fine-scale data on the systems we study. It has also been noted that "big data needs big models" (Gelman 2014). These big models, including models derived from machine learning methods, as well as flexible procedures deriving from classical statistics such as semiparametric, empirical likelihood, dimension reduction, and localized methods, can be powerful tools for improving the properties of NHST. Recent work on high dimensional inference is providing new tools to build such models while not saturating the models to the point where parameter estimates become meaningless. However, most applied researchers and many statisticians are not using these new tools to their full potential. The findings of McShane and Gal make clear that in terms of communication, training, and methods development, there is still a lot of room to grow.

## Funding

The authors gratefully acknowledge funding from the National Science Foundation (DMS-1555141, DMS-1557733, DMS-1513579, DMS-1317631) and the National Institutes of Health (P01 CA142538).

## References

Barr, D. J., Levy, R., Scheepers, C., and Tily, Harry, J. (2013), "Random Effects Structure for Confirmatory Hypothesis Testing: Keep It Maximal," *Journal of Memory and Language*, 68. [903]

Candès, E., and Barber, R. (2015), "Controlling the False Discovery Rate via Knockoffs," *Annals of Statistics*, 43, 2055–2085. [903]

Efron, B. (2004), "Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis," *Journal of the American Statistical Association*, 99, 96–104. [903]

Gelman, A. (2014), "Big Data Needs Big Model," available at *http://andrewgelman.com/2014/05/22/big-data-needs-big-model*. [904]

Howick, J., Friedemann, C., Tsakok, M., Watson, R., Tsakok, T., Thomas, J., Perera, R., Fleming, S., and Heneghan, C. (2013), "Are Treatments More Effective than Placebos? A Systematic Review and Meta-Analysis," *PLoS One*, 11, e0147354. [903]

Knutsson, H., Eklund, A., and Nichols, T. E. (2016), "Cluster Failure: Why fmri Inferences for Spatial Extent Have Inflated False-Positive Rates," *Proceedings of the National Academy of Sciences of the United States of America*, 113, 7900–7905. [903]

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A., (2010), "Tackling the Widespread and Critical Impact of Batch Effects in High-Throughput Data," *Nature Reviews Genetics*, 11, 733–739. [903]

Pritchard, J. K., and Voight, B. F. (2005), "Confounding From Cryptic Relatedness in Case–Control Association Studies," *PLoS Genetics*, 1, 302–311. [903]

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2017, VOL. 112, NO. 519, Applications and Case Studies
https://doi.org/10.1080/01621459.2017.1323642



# Rejoinder: Statistical Significance and the Dichotomization of Evidence

Blakeley B. McShane[a] and David Gal[b]

[a]Kellogg School of Management, Northwestern University, Evanston, IL; [b]College of Business Administration, University of Illinois at Chicago, Chicago, IL



We heartily thank editor Montserrat Fuentes for selecting our article (McShane and Gal 2017) for discussion. We are grateful for the opportunity to receive feedback on our work from four sets of distinguished discussants who possess a tremendous breadth of knowledge and expertise, and we deeply thank them for the time and effort they put into contemplating and responding to our article. We were delighted that our principal point—namely, that even expert statisticians are sometimes prone to misuse and misinterpret *p*-values and that these errors disproportionally arise from interpreting evidence dichotomously based on whether or not a *p*-value crosses the conventional 0.05 threshold for statistical significance—was both clear to and appreciated by our four sets of discussants.

In this rejoinder, we aim to do three things. First, we clarify and expound on certain aspects of our study designs and results to respond to some potential alternative accounts and criticisms raised in the discussion. Second, we tie together several broad themes that emerged in the discussion. Finally, we explore issues related to statistical significance and the dichotomization of evidence in the domain in which we most often work, namely, social psychology and consumer behavior.

In the remainder of this rejoinder, we abbreviate the discussions as DAB (Berry 2017), WMB (Briggs 2017), GC (Gelman and Carlin 2017), and LS (Laber and Shedden 2017).

## 1. Study Designs and Results

### 1.1. Study 1

DAB and LS both raise a concern regarding a potential misinterpretation by our subjects of the principal question asked in Study 1, in particular a confusion over whether the question we asked was one about the sample (i.e., about descriptive statistics) or about the population (i.e., about statistical inference). Both key in on the phrase "Speaking only of the subjects who took part in this particular study" used in the response options as

**CONTACT** Blakeley B. McShane *b-mcshane@kellogg.northwestern.edu* Associate Professor, Marketing Department, Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60611.

potentially responsible for our results, with DAB regarding the phrase as ambiguous and LS regarding it as an "unmistakable cue" (at least *ex post*). We note that, due to the design of our study, this phrase—and its ambiguity or clarity—cannot be responsible for our results. The reason for this is (i) subjects were randomized to one of two wordings of the response options where response wording one included the phrase and response wording two omitted it (this was the sole difference between the two response wordings) and (ii) the results were not substantially affected by the response wording (see Figure 1a of our article; see also Figure 1b, which shows the same was true in Study 1 of McShane and Gal (2016) where a third response wording was used and subjects were authors of articles published in the *New England Journal of Medicine*).

However, it is possible that a different confusion between sample and population may have arisen. In particular, while responses in treatment and control groups are often modeled using infinite population parametric models (e.g., independent normal with different means or independent binomial with different proportions), randomization secures only a finite population permutation model: under randomization, the population in question does not consist of additional subjects who were not included in the study but rather consists of both potential outcomes (i.e., under treatment and under control of which of course only one is observed) of each subject included in the study (generalization to additional subjects is a distinct matter). Under the permutation model, it could be argued that statements such as "the average for the treatment" can be ambiguous in terms of whether they refer to the average for those subjects who actually received the treatment in the study (i.e., the sample average) versus the average for all subjects under the hypothetical that they all received the treatment (i.e., the population average); under the latter interpretation, one might perhaps be justified in giving a different response for the $p = 0.01$ and $p = 0.27$ versions of the question. However, as only four subjects reported using the permutation model, this explanation cannot hold in practice. Further, our response wording generally precluded the latter interpretation (i.e., by asking about the average of "participants who were in Group A" it is unreasonable to assume we were asking about a hypothetical under which all participants were assigned to Group A).

We also wish to reiterate that the claim that the mere presence of a $p$-value in the question naturally led our subjects to focus on statistical inference rather than description is not really a criticism but rather is essentially our point: our subjects are so trained to focus on statistical significance that the mere presence of a $p$-value leads them to automatically view everything through the lens of the null hypothesis significance testing (NHST) paradigm—even in cases where it is unwarranted.

Further, as acknowledged by LS, subjects were asked to explain in their own words why they chose the options they chose. As shown in our article, their text responses emphasized that they were thinking dichotomously in a manner consistent with the dichotomization of evidence intrinsic to the NHST paradigm. Moreover, their responses to the two versions of our principal question did not associate particularly strongly with their responses to our various follow-up questions.

A final concern raised by DAB was that our study had a "low response rate" due to potential confusion over whether the question we asked was one about the sample or about the population and generated by the "Speaking only" phrase. In response, we note that our response rate of 27% is actually rather high for this kind of survey and subject population. Further, due to the design of our study, this phrase cannot be responsible for our response rate as (i) subjects were randomized to one of two wordings of the response options where response wording one included the phrase and response wording two omitted it and (ii) those randomized to response wording one would have seen the phrase only after they had already responded to the survey. Instead, if the phrase were to have had an impact, it would have been on the completion rate of our survey rather than the response rate to it. However, our completion rate did not substantially differ by the response wording and, at 94%, is extremely high.

### 1.2. Study 2

DAB accepts the results of Study 2 for Bayesians but not for frequentists. We do not necessarily disagree with his underlying logic but wish to expound upon this. First, subjects' responses to our follow-up question regarding statistical approach (frequentist, Bayesian, neither, or both) did not particularly strongly associate with their responses to either of the principal questions. Second, the text responses of our subjects provide little support for any concern about Bayesian versus frequentist reasoning. Third, any concern about Bayesian versus frequentist reasoning seems most germane to the likelihood judgment question rather than the choice question. However, as noted in our article and by LS, there is a sense in which option *D* is the correct frequentist option for the likelihood judgment question (i.e., because at no particular $p$-value is the null hypothesis definitively overturned). More specifically, which drug is "more likely" to result in recovery depends upon the parameters governing the probability of recovery for each drug, and these parameters are unknown and unknowable under a classical frequentist interpretation of the question. However, subjects generally chose option *A* for the likelihood judgment question when the $p$-value was set below 0.05 but option *D* when it was set above 0.05 rather than option *D* regardless. Thus, it seems improbable that subjects approached the question in this manner.

LS suggest that perhaps some of our subjects were engaging in response substitution (Gal and Rucker 2011), in particular, that subjects who were presented with a $p$-value greater than 0.05 "'read' between the 'lines' and answer[ed] the question they felt the investigators meant to ask," namely, one of statistical significance. Were subjects engaging in response substitution, we might have expected their text responses to reflect it. In particular, we might have expected them to say something along the lines of, "Drug A is more likely to lead to recovery from the disease than Drug B, but it is not statistically significantly more likely to lead to recovery." However, we did not see text responses of this sort. We further note that, while the likelihood judgment question may allow for this interpretation, the choice question allows little room for it; nonetheless, a meaningful share of subjects did not choose Drug A.

We also note that, while we agree with DAB that "For a different prior distribution it is quite possible for the [posterior] probability that Drug A is more effective than Drug B to be 0.99, say, and yet Drug B have the greater [posterior] mean and so be the correct choice" (although for it to be "correct" requires some additional assumptions about the loss function), we believe this is not relevant to our study as the question wording explicitly encouraged a noninformative prior (i.e., "Assuming no prior studies have been conducted with these drugs, which of the following statements is most accurate?").

## 2. Themes

There were several broad themes that emerged in the discussion to which we would like to draw attention. There was agreement that the very definition or logic of the *p*-value is problematic in and of itself. This was put perhaps with greatest flourish by DAB who stated that *p*-values are "perversions of logic" that are "fundamentally un-understandable" and led some to the conclusion that "the only reasonable path forward is to kill [*p*-values]" (DAB) and that "there are no good reasons nor good ways to use *p*-values. They should be retired forthwith" (WMB). GC go further and argue that many oft-suggested replacements for *p*-values such as confidence intervals and Bayes factors share some of the same problems in terms of inducing dichotomous (or more broadly categorical) thinking.

Related to dichotomous thinking is what GC term deterministic thinking, namely, "demanding more certainty than [the] data can legitimately supply" (GC) and the related "mentality that $p < 0.05$ means true and $p > 0.05$ means not true" (DAB) (or, as we put it, the assignment of evidence to the different categories "statistically significant" and "not statistically significant" naturally leads to the conclusion that the treatments thusly assigned are categorically different). This becomes particularly problematic and pronounced when, as GC note, most effects measured in applied research represent a mean in some population (or something similar such as a regression coefficient)—a fact which they note "seems to be lost from consciousness when researchers slip into binary statements about there being 'an effect' or 'no effect' as if they are writing about constants of nature;" this issue is strongly compounded in the biomedical and social sciences where an effect (i.e., mean, regression coefficient) of zero is generally implausible.

Hypothesized zero mean effects tie nicely to the issue of the "straw man" null hypotheses decried by LS (but used in the overwhelming majority of applications) as well as the fact that, as per GC, there is generally no clean mapping between a scientific hypothesis (or theory) on one hand and a statistical hypothesis on the other hand with the latter often being one of many possible particular and concrete operationalizations of the former. Nonetheless, GC are correct that "there is a demand for hypothesis testing": applied researchers want to accept and reject hypotheses (and theories) and are not content with admonitions that they may only "retain the null" or that "rejection of the null should not imply acceptance of the alternative." A closer mapping between the scientific hypothesis and its operationalization as a statistical hypothesis as well as using a "default 'null model' [that] is a rich and complex model" (LS) may help in this regard where it is possible.

An additional theme concerned the notion that "probability is not a decision" (WMB)—beliefs are not actions—and the fact that "we have fallen prey to this accept-reject silliness (i.e., dichotomous thinking) because many decisions are binary" (DAB). However, as rightly pointed out by LS, when faced with a decision, the proper course of action is not to make it based on statistical hypotheses or probabilities alone but rather to conduct a full decision analysis that accounts for the costs and benefits of the various alternatives and to choose the one with, for example, the greatest expected value (although see Diaconis (2003) for a humorous cautionary note on conducting decision analyses); while, as per LS, there will still be "near hits and near misses" when using a decision analysis, a decision analysis nonetheless constitutes a major improvement over using the outcome of a statistical hypothesis test alone as the decision. In this regard, the point made by WMB that decisions are relative to person and situation and that a probability model that is useful for a given person in a given situation can be irrelevant to another person in another situation is important to bear in mind.

We further note that, while we agree with DAB that "many decisions are binary" (or at least categorical) in nature and consequently with LS that "decisions do arise that cannot be made continuously," we urge caution in this matter as many decisions that appear on the surface as binary or categorical are actually—or can be reframed to be—continuous. For example, a decision about whether or not to invest in some project can be viewed as a decision about how much to invest in the project. We believe such a continuous view of the underlying decision will naturally lead to a more continuous view of the evidence and make the issue of near hits and near misses less relevant.

Finally, while, as DAB notes, *p*-values may not cause "much harm if the focus is the primary endpoint from a protocol and the *p*-value is calculated based on a prospective analysis of that endpoint," all discussants brought up that fact that multiple comparisons—including multiple potential comparisons or the "garden of forking paths" (Gelman and Loken 2014)—are the norm in applied research and the consequence that—strictly speaking—this in practice invalidates all *p*-values except those from studies with preregistered protocols and data analysis procedures; as DAB put it, "a *p*-value has no inferential role outside the rigidness of a protocol" (and, we note, it may not inside if the underlying model that generated the *p*-value is misspecified in an important manner; we further note that while we view preregistration as often laudable, it has several limitations including being typically confirmatory and possible only in certain applied domains). This led to a discussion of alternative methods including posterior predictive probabilities of observables (WMB), hierarchical modeling and penalized (or regularized) inference techniques (GC), and false discovery rate methods (LS). We agree that all of these methods constitute a large improvement on the rote and recipe-like application of NHSTs but share the concern expressed by LS that "most applied researchers and many statisticians are not using these new tools to their full potential."

## 3. Social Psychology and Consumer Behavior Research

While we share the discussants' enthusiasm for recent methodological developments, we do question their applicability to the

domain in which we most often work, namely, social psychology and consumer behavior. In this domain, the fundamental unit of analysis is the individual study, and the prototypical study follows a two-by-two between-subjects design where interest centers on demonstrating multiple effects—both null and nonnull—by using the linear model to conduct NHSTs on contrasts of the means of the individual-level observations in each condition. In the best of cases (even if this is not all that common), the study measures a single dependent measure and both the contrasts of interest and the data analysis procedures (e.g., outlier exclusion rules, covariates to be included in the analysis) are specified in advance.

As can be seen, dichotomization is rife in this paradigm. Not only are there the aforementioned dichotomy of the null hypothesis versus the alternative hypothesis; the dichotomization of results into the different categories statistically significant and not statistically significant; and the dichotomous thinking about there being an effect or no effect when such effects are contrasts of means, but also there is dichotomization built into the very experimental design: each experimental factor is manipulated in a dichotomous manner as if it were a light being switched on and off.

Beyond dichotomous thinking, the NHST paradigm causes additional problems in this domain. For example, because individual-level measurements are typically quite errorful, sample sizes are not especially large, and effects are small and variable, study estimates are themselves often rather noisy; noisy estimates in combination with the fact that the publication process typically screens for statistical significance results in published estimates that are biased upward (potentially to a large degree) and often of the wrong sign (Gelman and Carlin 2014). Further, the screening of estimates for statistical significance by the publication process to some degree almost encourages researchers to conduct studies with errorful measurements and small sample sizes because such studies will often yield one or more statistically significant results. Of course, all of these issues are further compounded when researchers engage in multiple comparisons—whether actual or potential.

Nonetheless, as GC noted, "there is a demand for hypothesis testing" in this domain to demonstrate effects ("to establish stylized facts" in the language of Gelman (2017)). Unfortunately, these effects are typically demonstrated by rejecting the straw man null hypothesis of zero effect decried by LS; however, it is unclear whether the rich and complex null models LS favor are possible or realistic for this data. Further, it is also unclear whether recent methodological developments can play much of a role because, for example, researchers seldom have observables for which they seek posterior probabilities, studies have no hierarchical structure, and adjustment for multiplicities via penalized inference techniques or false discovery rate methods makes little sense when zero effects are generally implausible (in this domain, there are not a small number of large effects coupled with a large number of zero effect but rather a large number of small and variable effects).

Consequently, we have been developing and encouraging the use of methods that concord with GC's call for "a greater acceptance of uncertainty and embracing of variation" while simultaneously satisfying researchers' demand to demonstrate effects. One particular area of focus has been attempting to divert attention away from individual studies, which as noted above can often be noisy, by developing meta-analytic methods (i.e., hierarchical models) that are specially tailored to the single paper meta-analysis of the multiple studies of a common phenomenon that appear in a typical research paper (McShane and Böckenholt 2017) as well as the more traditional meta-analysis of multiple studies from multiple papers that vary considerably in terms of their dependent measures and moderators (i.e., experimental factors) (McShane and Böckenholt 2017). As per GC, these methods assess and account for—indeed embrace—the variation (or heterogeneity) across multiple studies and papers (including differing degrees of variation across various dependent measures) as well as the covariation induced by the fact that observations are nested within, for example, papers, studies, groups of subjects, and study conditions; *inter alia*, this can help encourage the careful consideration of potential moderators of this variation thereby resulting in deeper and richer theories.

Further, these methods are, as noted, capable of satisfying researchers' demand to demonstrate effects, in particular via meta-analytic NHSTs. However, they do so in a perhaps subversive manner: because zero effects are generally implausible in this domain and because meta-analyses generally have much greater power than single studies, meta-analytic NHSTs are highly likely to be rejected. If the rejection of these meta-analytic NHSTs can satisfy researchers' demand to demonstrate effects, this should help divert attention away from noisy single-study NHSTs (and perhaps NHSTs in general) and free it up to focus on, for example, the estimation of effect sizes and their convergence and divergence (i.e., variation) across studies and papers as well as various dependent measures. It may also lessen considerably the degree to which the publication process screens for statistical significance (at least at the level of the individual study).

Given that the demand to demonstrate effects and the dominance of the prototypical study design are both at present firmly entrenched in this domain, we believe these methods provide researchers a means of accepting uncertainty and embracing variation that is also respectful of and responsive to their goals and data. We also believe these methods—along with other measures such as more precise individual-level measurements, larger sample sizes, a greater use of within-subjects (or longitudinal) study designs, and deeper connection between theory, measurement, and data (Gelman 2017)—should also help with current difficulties in replication.

## References


Berry, D. A. (2017), "A *p*-Value to Die For," *Journal of the American Statistical Association*, 112, this issue. [904]

Briggs, W. M. (2017), "The Substitute for *p*-Values," *Journal of the American Statistical Association*, 112, this issue. [904]

Diaconis, P. (2003), "Problem of Thinking Too Much," *Bulletin of the American Academy of Arts and Sciences*, Spring 2003, 26–38. [906]

Gal, D., and Rucker, D. D. (2011), "Answering the Unasked Question: Response Substitution in Consumer Surveys," *Journal of Marketing Research*, 48, 185–195. [905]

Gelman, A. (2017), "The Failure of Null Hypothesis Significance Testing when Studying Incremental Changes, and What To Do About It," *Personality and Social Psychology Bulletin*, forthcoming. [907]

Gelman, A., and Carlin, J. (2014), "Beyond Power Calculations Assessing Type s (sign) and Type m (Magnitude) Errors," *Perspectives on Psychological Science*, 9, 641–651. [907]

——— (2017), "Some Natural Solutions to the *p*-Value Communication Problem—and Why They Won't Work," *Journal of the American Statistical Association*, 112, this issue. [904]

Gelman, A., and Loken, E. (2014), "The Statistical Crisis in Science," *American Scientist*, 102, 460–465. [906]

Laber, E. B., and Shedden, K. (2017), "Comment: Statistical Significance and the Dichotomization of Evidence: The Relevance of the ASA Statement on Statistical Significance and p-values for Statisticians," *Journal of the American Statistical Association*, 112, this issue. [904]

McShane, B. B., and Böckenholt, U. (2017), "Single Paper Meta-Analysis: Benefits for Study Summary, Theory-Testing, and Replicability," *Journal of Consumer Research*, 43, 1048–1063. [907]

——— (2017), "Multilevel Multivariate Meta-Analysis With Application to Choice Overload," *Psychometrika*. [907]

McShane, B. B., and Gal, D. (2016), "Us to the Obvious? The Effect of Statistical Training on the Evaluation of Evidence," *Management Science*, 62, 1707–1718. [905]

——— (2017), "Statistical Significance and the Dichotomization of Evidence," *Journal of the American Statistical Association*, 112, this issue. [904]