Journal of the American Statistical Association
September 2017 Volume 112 Number 519

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Statistical Significance and the Dichotomization of Evidence

Blakeley B. McShane & David Gal

To cite this article: Blakeley B. McShane & David Gal (2017) Statistical Significance and the Dichotomization of Evidence, Journal of the American Statistical Association, 112:519, 885-895, DOI: 10.1080/01621459.2017.1289846

To link to this article: http://dx.doi.org/10.1080/01621459.2017.1289846

Published online: 30 Oct 2017.

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.

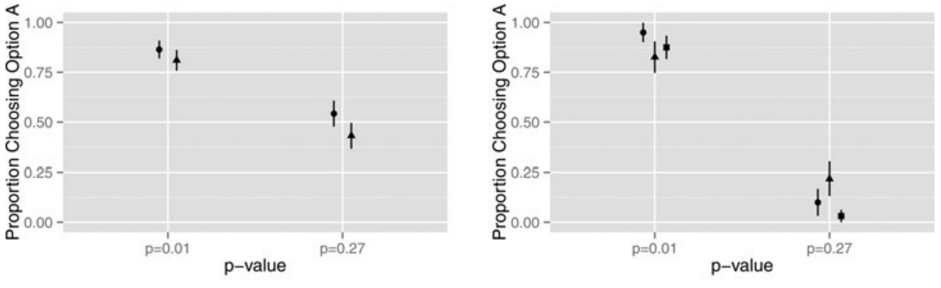
These results are striking and suggest that the dichotomization of evidence intrinsic to the NHST paradigm leads even expert statisticians to think dichotomously. In particular, about half the subjects failed to identify differences that were not statistically significant as different.

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:

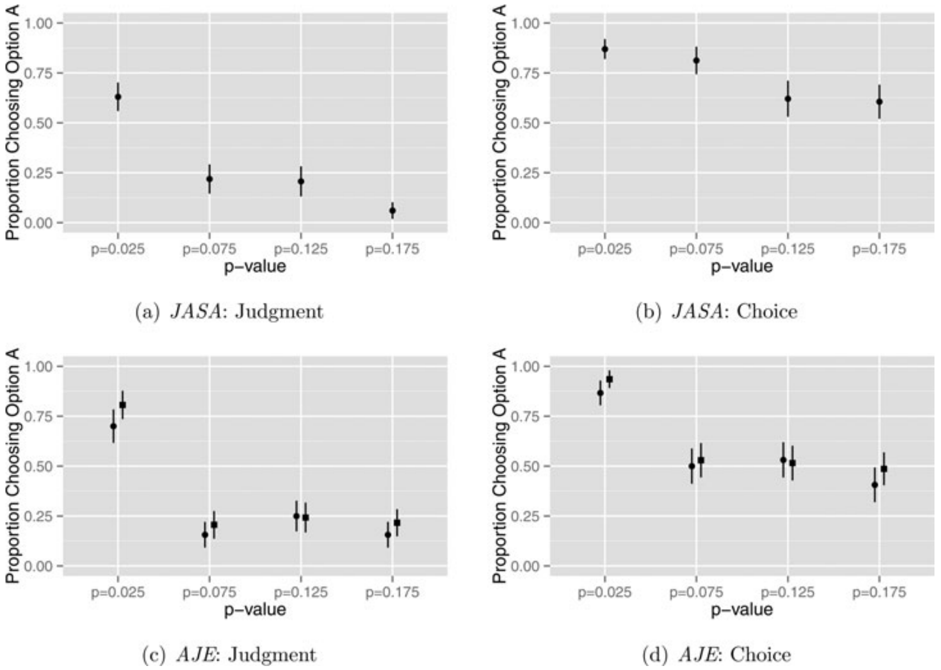
> Assuming no prior studies have been conducted with these drugs, if you were a patient from the same population as the subjects in the study, what drug would you prefer to take to maximize your chance of recovery?
> A. I prefer Drug A.
> B. I prefer Drug B.
> C. I am indifferent between Drug A and Drug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