Submitted to *Management Science*
manuscript MS-13-00804

Blinding Us to the Obvious? The Effect of Statistical Training on the Evaluation of Evidence

Blakeley B. McShane
Kellogg School of Management, Northwestern University, Evanston, IL 60208, b-mcshane@kellogg.northwestern.edu

David Gal
College of Business Administration, University of Illinois at Chicago, Chicago, IL 60607, dgaluic@gmail.com

Statistical training helps individual analyze and interpret data. However, the emphasis placed on null hypothesis significance testing in academic training and reporting may lead researchers to interpret evidence dichotomously rather than continuously. Consequently, researchers may either disregard or undervalue evidence that fails to attain statistical significance relative to evidence that attains statistical significance. Surveys of researchers across a wide variety of fields (including medicine, epidemiology, cognitive science, psychology, business, and economics) show that a substantial majority indeed does so. This phenomenon is manifest in both how researchers interpret descriptions of evidence and in their likelihood judgments. Dichotomization of evidence is reduced though still present when researchers are asked to make decisions based on the evidence, particularly when the decision outcome is personally consequential. Recommendations are offered.

Key words: sociology of science, evaluation of evidence, strength of evidence, null hypothesis, significance testing, p-values, description, inference, judgment, choice

1. Introduction

The null hypothesis significance testing (NHST) paradigm is the dominant statistical paradigm in academic training and reporting in the biomedical and social sciences (Morrison and Henkel 1970, Gigerenzer 1987, Sawyer and Peter 1983, McCloskey and Ziliak 1996, Gill 1999, Anderson et al. 2000, Gigerenzer 2004, Hubbard 2004). A prominent feature of the NHST paradigm is the enshrinement of the eponymous null hypothesis, which typically posits that there is no difference between two or more groups with respect to some underlying population parameter of interest (e.g., a mean or proportion). Pitted against the null hypothesis is the alternative hypothesis, which, in typical applications, posits that there is a difference between the groups. Standard practice involves

collecting data, computing a p-value which is a function of the data and the null hypothesis, and then retaining or rejecting the null hypothesis depending on whether the p-value is respectively above or below the size α of the hypothesis test where α is conventionally set to 0.05.

Despite the overwhelming dominance of the NHST paradigm in practice, it has received no small degree of criticism over the decades. Consider, for instance, the following passage from Gill (1999):

> It [NHST] has been described as a "strangle-hold" (Rozenboom 1960), "deeply flawed or else ill-used by researchers" (Serlin and Lapsley 1993), "a terrible mistake, basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology" (Meehl 1978), "an instance of the kind of essential mindlessness in the conduct of research" (Bakan 1966), "badly misused for a long time" (Cohen 1994), and that it has "systematically retarded the growth of cumulative knowledge" (Schmidt 1996). Or even more bluntly: "The significance test as it is currently used in the social sciences just does not work." (Hunter 1997)

Clearly NHST is not without its critics.

A major line of criticism against the NHST paradigm pertains to the misinterpretation of the p-value. 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 has often been misinterpreted as *inter alia* (i) the probability that the null hypothesis is true, (ii) one minus the probability that the alternative hypothesis is true, and (iii) one minus the probability of replication (Bakan 1966, Sawyer and Peter 1983, Nickerson 2000, Cohen 1994, Schmidt 1996, Krantz 1999, Gigerenzer 2004, Kramer and Gigerenzer 2005). For example, Gigerenzer (2004) reports that, in a survey of psychology professors, lecturers, teaching assistants, and students, 103 out of 113 endorsed one or more of six false statements about p-values (see also Haller and Krauss (2002) and Gigerenzer et al. (2004)). Similarly, Cohen (1994) reports that sixty-eight out of seventy academic psychologists misinterpreted the p-value as the probability that the null hypothesis is true while forty-two believed a p-value of 0.01 implied a 99% chance that a replication would yield a statistically significant result (see also Oakes (1986)).

These common misinterpretations can be wildly off the mark in practice and, consequently, another line of criticism against NHST is that the p-value is a poor measure of the evidence for or against a statistical hypothesis. For example, Cohen (1994) provides a setting where the p-value is less than 0.05 but the probability that the null hypothesis is true is about 0.60; consequently, he affirms that one can be very wrong by erroneously considering the p-value "as bearing on the truth of the null hypothesis." In other words, there can be a large divergence between the probability of

the null hypothesis given the data and the p-value–the probability of the data (or more extreme data) given the null hypothesis–and in a variety of settings, this can cause the p-value to exaggerate the evidence against the null (see also Berger and Delampady (1987), Berger and Sellke (1987), Casella and Berger (1987), Berger and Berry (1988), and Hubbard and Lindsay (2008)).

Another series of criticisms levied against the NHST paradigm pertain to the various forms of dichotomization associated with it, for instance the dichotomy of the null hypothesis versus the alternative hypothesis and the dichotomization of results into "statistically significant" and "not statistically significant". For instance, the influential Bayesian textbook Gelman et al. (2003) criticizes "the artificial dichotomy" required by sharp point null hypothesis significance tests of $\theta = \theta_0$ versus $\theta \neq \theta_0$ (where θ is some statistical parameter of interest and θ_0 is some value for that parameter) and notes that "difficulties related to this dichotomy are widely acknowledged from all perspectives on statistical inference"; the authors instead suggest that an estimate of the posterior distribution of θ or an interval estimate of θ provide more interesting and relevant information as compared to asking whether θ equals θ_0.

A more specific variant of this dichotomization criticism relates to the particular form that the sharp point null hypothesis takes in the overwhelming majority of empirical applications, namely sharp point null hypothesis significance tests of the form $\theta = 0$ versus $\theta \neq 0$ (e.g., tests of no difference between two or more groups). It is argued that the null hypothesis of zero effect is never and could never be precisely true in practice–particularly in the social sciences–due to factors such as measurement error and varying treatment effects (Tukey 1991, Cohen 1994, Gelman et al. 2003, Gelman 2015). Indeed, Cohen (1994) derides this form of the null hypothesis as the "nil hypothesis" and lampoons it as "always false." Similarly, Gelman et al. (2003) note that, for continuous parameters (e.g., the difference between two means or proportions), the null hypothesis of "exactly zero is rarely reasonable" while Tukey (1991) notes that two treatments are "always different."

Yet another series of criticisms of the NHST paradigm relating to dichotomization pertains to the labeling of results as statistically significant or not statistically significant depending upon whether or not the p-value is respectively below or above the size α of the test where α is conventionally set to 0.05. One well-known criticism is that the 0.05 threshold is entirely arbitrary (Fisher (1926), Yule and Kendall (1950), Cramer (1955), Cochran (1976), Cowles and Davis (1982)) and that the threshold selected should depend on the application at hand. While this argues for replacing 0.05 with a different perhaps application-specific threshold, another line of criticism suggests that the problem is with having a threshold in the first place: the dichotomization into statistically significant and not statistically significant itself has "no ontological basis" (Rosnow and Rosenthal 1989). Consequently, Rosnow and Rosenthal (1989) stress that "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?".

While these criticisms have respectively focused on the practice and metaphysics of the dichotomization of results into statistically significant and not statistically significant, still another line of criticisms has considered the potentially harmful effects of dichotomization. For instance, it is well-known that statistical significance and practical importance are often confused; indeed, this confusion is so rampant that, to preempt it, introductory statistics textbooks repeatedly affirm, with a frequency rivaled only by the frequency of declarations that correlation does not imply causation, that statistical significance is distinct from practical importance (Freedman et al. 2007). To illustrate this point, consider Freeman (1993)'s example of four hypothetical trials in which subjects express a preference for treatment A or treatment B. With sample sizes of 20, 200, 2,000, and 2,000,000 and preferences for A of 75.0%, 57.0%, 52.3%, and 50.07% respectively, all four trials produce statistically significant p-values of about 0.04; nonetheless, the effect size in the largest study shows that the two treatments are nearly identical in terms of preference. Consequently, researchers err greatly by confusing statistical significance with practical importance.

Another ill effect of the dichotomization of results into statistically significant and not statistically significant is that researchers treat results that attain statistical significance as evidence for an effect while they treat results that fail to attain statistical significance as evidence of the absence of an effect. Gelman and Stern (2006) have discussed one important implication of this practice, namely that researchers commonly infer that two treatments are significantly different when one treatment attains statistical significance while the other fails to do so. In reality, the two treatments may have a statistically similar effect, or as Gelman and Stern (2006) conclude, "the difference between 'significant' and 'not significant' is not itself statistically significant." Gelman and Stern (2006) note that this is conceptually distinct from the arbitrariness of the 0.05 threshold and provide an example where the difference between an effect that clearly attains statistical significance (i.e., $d_A = 25;\ se = 10;\ p = 0.01$) and one that clearly fails to do so (i.e., $d_B = 10;\ se = 10;\ p = 0.32$) itself clearly fails to do so (i.e., $d_A - d_B = 25 - 10;\ se = \sqrt{100^2 + 100^2};\ p = 0.29$). Further, Gelman and Stern (2006) point to examples of this mistake in the literature and trace it back to the dichotomization of evidence promoted by the NHST paradigm: assigning treatments to different categories naturally leads to the conclusion that the treatments thusly assigned are categorically different.

In this paper, we investigate one way in which the NHST paradigm may lead researchers to misinterpret evidence. In particular, given the focus on NHST and the concomitant dichotomization of results into statistically significant and not statistically significant in academic training and reporting, we hypothesized that researchers–despite general knowledge that the conventional

0.05 level of statistical significance is arbitrary–tend to think of evidence in dichotomous terms. This dichotomization may manifest itself in several ways. For example, individuals may interpret evidence that reaches the conventionally defined threshold for statistical significance as a demonstration of a difference and may in contrast interpret evidence that fails to reach this threshold as a demonstration of no difference. Similarly, individuals' confidence that a difference exists or their perceptions of the practical importance of a difference may be sharply kinked around $p = 0.05$, with a precipitous change in the confidence that a difference exists or the perception of the practical importance of a difference when the p-value crosses the $p = 0.05$ threshold.

As an example of how dichotomous thinking manifests itself, consider Messori et al. (1993)'s comparison of their findings with those of Hommes et al. (1992):

> The result of our calculation was an odds ratio of 0.61 (95% CI: $0.298 - 1.251$; $p > 0.05$); this figure differs greatly from the value reported by Hommes and associates (odds ratio: 0.62; 95% CI: $0.39 - 0.98$; $p < 0.05$)...we concluded that subcutaneous heparin is not more effective than intravenous heparin, exactly the opposite to that of Hommes and colleagues.

In other words, Messori et al. (1993) conclude that their findings are "exactly the opposite" of Hommes et al. (1992) because their odds ratio estimate failed to attain statistical significance whereas that of Hommes et al. (1992) attained statistical significance. In fact, however, the odds ratio estimates and confidence intervals of Messori et al. (1993) and Hommes et al. (1992) are highly consistent (for additional discussion of this example and others, see Healy (2006)).

The remainder of this paper is organized as follows. In Study 1, we demonstrate that researchers misinterpret mere descriptions of data depending on whether a p-value is above or below 0.05. In Study 2, we extend this result to the evaluation of evidence via likelihood judgments; we also show the effect is attenuated but not eliminated when researchers are asked to make hypothetical choices. Finally, we discuss some implications of our findings and present recommendations for statistical training and reporting.

2. Study 1: Descriptive Statements

2.1. Objective

The goal of Study 1 was to examine the hypothesis that a focus on statistical significance would lead researchers to misinterpret data. To systematically examine this question, we presented researchers with a study summary that showed a difference in an outcome variable associated with an intervention and a set of descriptions of that difference. We manipulated whether the difference in the outcome variable attained ($p = 0.01$) or failed to attain ($p = 0.27$) statistical significance. We

posited that researchers would correctly identify that the outcome variable differed when the difference attained statistical significance but would fail to identify this difference when it failed to attain statistical significance.

2.2. Participants

Participants were the authors of articles published in the 2013 volume of the *New England Journal of Medicine* (NEJM; issues 368.1-368.10). A link to our survey was sent via email to the 322 authors; about twenty email addresses were incorrect. Seventy-five authors completed the survey, yielding a completion rate of 25%.

2.3. Procedure

Participants were asked to respond sequentially to two versions of a principal question followed by several follow-up questions. As noted above, the principal question asked participants 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. We manipulated whether this difference attained ($p = 0.01$) or failed to attain ($p = 0.27$) statistical significance within subjects, with participants first asked to answer the $p = 0.27$ version of the question and then, on the next screen, the $p = 0.01$ version of the question. To test for robustness to differences in the wording of the response options, participants were randomized to one of three variations (for full details, see the online supplementary materials). For purposes of illustration, we present one below:

> 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. Participants 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. Participants were then tracked until all had died. Participants in Group A lived, on average, 8.2 months post-diagnosis whereas participants in Group B lived, on average, 7.5 months post-diagnosis ($\boldsymbol{p = 0.27}$).
>
> 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 participants who were in Group A was **greater** than that lived by the participants 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 participants who were in Group A was **less** than that lived by the participants who were in Group B.

Option	$p = 0.01$	$p = 0.27$
A	95	10
B	0	0
C	0	55
D	5	35
n	20	

(a) Wording 1

Option	$p = 0.01$	$p = 0.27$
A	83	22
B	0	0
C	0	35
D	17	43
n	23	

(b) Wording 2

Option	$p = 0.01$	$p = 0.27$
A	88	3
B	3	0
C	6	62
D	3	34
n	32	

(c) Wording 3

Table 1 Study 1 Results. Each cell gives either the percentage of participants who gave the given response option or the sample size. Participants are much more likely to answer correctly when $p = 0.01$.

C. Speaking only of the subjects who took part in this particular study, the average number of post-diagnosis months lived by the participants who were in Group A was **no different** than that lived by the participants 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 participants who were in Group A was greater/no different/less than that lived by the participants who were in Group B.

After these questions, participants were asked (i) a multiple choice question about their primary area of expertise (i.e., medicine, chemistry/biology, statistics/biostatistics, engineering, or other), (ii) a free response question asking at what p-value statistical significance is conventionally defined ($p < 0.05$; 97% of participants answered correctly), and (iii) a question about their statistical model for the data (for full details, see the online supplementary materials).

2.4. Results and Discussion

We present our results in Table 1 (for full details, see the online supplementary materials). For the principal question shown above, the correct answer is option A regardless of the p-value: all four response options are descriptive statements and indeed the average number of post-diagnosis months lived by the participants who were in Group A was greater than that lived by the participants who were in Group B (i.e., $8.2 > 7.5$). However, participants 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[1]. For instance, among participants who saw the response wording above, 95% correctly answered when the p-value was set to 0.01; on the other hand, only 10% correctly answered when the p-value was set to 0.27 with 55% choosing option C and 35% choosing option D. Responses in the other two response wording conditions were similar as can be seen in Table 1.

[1] Due to the subject matter of the paper as well as the large effect sizes observed in our data, we do not present null hypothesis significance tests and the concomitant p-values in the text. We thank the editor and the associate editor for understanding. All details of estimation can be found in the online supplementary materials.

These results are striking and suggest that, as hypothesized, a focus on statistical significance leads researchers to dichotomize evidence. In particular, participants failed to identify differences that were not statistically significant as different.

One potential criticism of our findings is that our question is essentially a trick question: researchers clearly know that 8.2 is greater than 7.5, but they might perceive that asking whether 8.2 is greater than 7.5 is too easy a question and hence they focus on whether the difference is statistically significant. However, asking whether a p-value of 0.27 is statistically significant is also trivial, so this criticism does not resolve why researchers focus on the statistical significance of the difference rather than on the difference itself. A related potential criticism regards our question as a trick question for a different reason: by including a p-value, we naturally lead researchers to focus on statistical significance. However, this is essentially our point: researchers 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 not warranted. Moreover, in further response to such criticisms, we note that we stopped just short of explicitly telling participants that we were asking for a description of the observed data rather than asking them to make a statistical inference (e.g., response options read, "**Speaking only of the subjects who took part in this particular study**, the average number of post-diagnosis months lived by the **participants who were in Group A** was greater than that lived by the **participants who were in Group B**" and similarly; emphasis added).

While not directly relevant to our hypotheses, there are two additional points worth noting. First, even if we had asked participants to make a statistical inference under the NHST paradigm rather than to simply describe the data, option C (which stated that the average number of months lived by participants in the two groups did not differ) is never correct: failure to reject the null hypothesis does not imply or prove that the two treatments do not differ. Second, and again assuming we were asking an inferential question rather than a descriptive question, there is a sense in which option D (which states that it cannot be determined whether the average number of months lived by participants in the two groups differed) is the correct answer regardless of the p-value since at no p is the null definitively overturned. However, only a relatively small proportion of participants chose option D as their response to both versions of the question (i.e., the $p = 0.01$ version and the $p = 0.27$ version), with most choosing option A for the $p = 0.01$ version and option C or option D for the $p = 0.27$ version.

2.5. Robustness

To test the robustness of the observed effect, we replicated Study 1 using different wordings for the response options, different question orders (e.g., $p = 0.01$ first versus $p = 0.27$ first), and different subject populations including researchers in psychology (members of the editorial board

of *Psychological Science*) and business (the 2013 Marketing Science Institute Young Scholars) as well as undergraduates both trained and untrained in statistics. To briefly summarize the results of these studies, neither the response wording nor the question order substantially affected the pattern of results. Further, consistent with our prediction that the focus placed on NHST in the training of professional researchers and in typical undergraduate courses would be associated with diminished performance on the $p = 0.27$ version of the question, 73% of statistically-untrained undergraduates answered $p = 0.27$ version of the question correctly compared to 17% of *Psychological Science* editorial board members, 19% of Marketing Science Institute Young Scholars, and 53% of statistically-trained undergraduates. Further, and perhaps not surprisingly, statistically-untrained undergraduates answered the $p = 0.01$ and $p = 0.27$ versions of the question correctly at the same rate. For additional details, see the online supplementary materials.

3. Study 2: Likelihood Judgments and Choices

3.1. Objective

Thus far, we have examined how differences in statistical significance affect researchers' descriptive statements about data. In Study 2, we examine whether the observed pattern of results extends from descriptive statements to the evaluation of evidence via likelihood judgments. To do so, we presented researchers with a study summary and a set of inferences that might be drawn from the data. As above, the p-value for the null hypothesis significance test of no difference between the two treatment groups was set above or below 0.05. While the p-value quantifies the strength of evidence against the null hypothesis, we hypothesized that participants would incorrectly judge (i.e., dismiss or undervalue in a relative sense) evidence that failed to attain statistical significance.

To examine whether participants' likelihood judgments would extend to decisions based on the data, we also asked participants to report how they would choose to act in light of the data. We hypothesized that when it comes to making a choice, researchers would, to some degree, shift their focus from whether a result is or is not statistically significant to which choice option represents the superior alternative[2]. As a result, we predicted that researchers would be more likely to select the superior alternative in the context of making a choice relative to the context of making a likelihood judgment. Moreover, we predicted that this effect would be more pronounced the more personally consequential the choice for the participant.

A further goal of Study 2 was to gain additional insight into researchers' reasoning when making likelihood judgments and choices by examining how varying (i) the degree to which the p-value is above the threshold for statistical significance and (ii) the magnitude of the treatment difference

[2] More precisely, here and hereafter, by the "superior alternative" we mean the alternative that is more likely to be superior which, in our setting, means the alternative that is more likely to be more effective in terms of the probability of recovery from a disease.

affects researchers' likelihood judgments and choices. We hypothesized that researchers would be substantially more likely to provide incorrect judgments when the p-value was set above 0.05 than when set below 0.05, but that (i) the degree to which the p-value exceeded 0.05 and (ii) the magnitude of the treatment difference would have little impact on the results as researchers' would focus almost solely on whether the difference between the treatments attained or failed to attain statistical significance.

3.1.1. Participants Participants were the authors of articles published in the 2013 volume of the *American Journal of Epidemiology* (AJE; issues 177.4 to 178.4). A link to our survey was sent via email to the 1,111 authors; about 110 email addresses were incorrect. 299 authors completed a survey, yielding a completion rate of 30%. Thirty-eight responses could not be used because they were inadvertently diverted to the wrong survey; consequently, we report results from the 261 participants who completed the correct survey.

3.2. Procedure

Participants completed a likelihood judgment question followed by a choice question. Participants were randomly assigned to one of sixteeen conditions following a four by two by two design. The first level of the design varied whether the p-value was set to 0.025, 0.075, 0.125, or 0.175 and the second level of the design varied the magnitude of the treatment difference (52% and 44% versus 57% and 39%). The third level of the design applied only to the choice question and varied whether participants were asked to make a choice for a close versus distant other (see below). Participants saw the same p-value and magnitude of the treatment difference in the choice question as they saw in the preceding judgment question.

The judgment question was:

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 **0.175**.

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.

For the choice question, participants were presented with the same study summary but were instead asked to make a hypothetical choice. Moreover, participants were randomized into one of two conditions: they were asked to choose a treatment for either a close other (i.e., a loved one) or a distant other (i.e., physicians treating patients). We predicted that participants would be more likely to choose a superior alternative for a close other than for a distant other when the superior alternative was not statistically significantly different from the inferior alternative. The basis for this prediction was our hypothesis that choice tends to shift the focus away from statistical significance and towards whether an option is superior combined with the logic that this shift would be greater the more consequential the choice for the individual. Participants in the close other condition saw the following wording:

> If you were to advise a loved one who was a patient from the same population as those in the study, what drug would you advise him or her to take?

Participants in the distant other condition saw the following wording:

> If you were to advise physicians treating patients from the same population as those in the study, what drug would you advise these physicians prescribe for their patients?

All participants then saw the following response options:

A. I would advise Drug A.
B. I would advise Drug B.
C. I would advise that there is no difference between Drug A and Drug B.

Option	Judgment				Choice			
	$p=0.025$	$p=0.075$	$p=0.125$	$p=0.175$	$p=0.025$	$p=0.075$	$p=0.125$	$p=0.175$
A	70	16	25	16	87	50	53	41
B	0	0	0	0	0	0	0	0
C	10	22	34	38	13	50	47	59
D	20	62	41	47				
n	30	32	32	32	30	32	32	32

(a) Small Treatmeant Difference

Option	Judgment				Choice			
	$p=0.025$	$p=0.075$	$p=0.125$	$p=0.175$	$p=0.025$	$p=0.075$	$p=0.125$	$p=0.175$
A	81	21	24	22	94	53	52	49
B	0	0	3	0	0	0	0	0
C	3	35	15	16	6	47	48	51
D	16	44	58	62				
n	31	34	33	37	31	34	33	37

(b) Large Treatmeant Difference

Table 2 Study 2 Results. Each cell gives either the percentage of participants who gave the given response option or the sample size. Participants are much more likely to answer the questions correctly when $p<0.05$ though this effect is attenuated for the choice question relative to the judgment question.

In addition to asking participants to make judgments and choices, we also sought to gain insight into participants' reasoning by asking them to explain why they chose the option they chose in free response form both after the judgment question and after the choice question. Participants were provided with a text box to provide their response.

After these questions, participants were asked a multiple choice question about their primary area of expertise (epidemiology, medicine, statistics/biostatistics, or other) and a free response question asking at what p-value statistical significance is conventionally defined ($p<0.05$; 99% of participants answered correctly).

3.3. Results and Discussion

As the effect of making a choice for a close versus distant other was a secondary hypothesis, we collapse over both "other" (i.e., close versus distant) conditions in the principal presentation of our results and return to the analysis of the effect of a close versus distant other below. We present our results in Table 2 (for full details, see the online supplementary materials).

While the p-value quantifies the strength of the evidence regarding the likelihood that the efficacy of Drug A is higher than that of Drug B (and thus the likelihood of hypothetical new patients recovering under Drug A versus Drug B), again the level of p-value does not alter the correct response option. The correct answer is again option A as Drug A is more likely to have higher efficacy than Drug B regardless of the p-value. The share of participants who correctly answered

the judgment question when the p-value was set to 0.025 was 70% and 81% in the small and large treatment difference conditions respectively. However, the share of participants who correctly answered the judgment question was substantially lower when the p-value was set to 0.075, 0.125, and 0.175, with no substantial variation in the share answering correctly across these three conditions (16%, 25%, and 16% respectively when the treatment difference was small and 21%, 24%, and 22% respectively when the treatment difference was large).

An argument might be made that there is a sense in which option D is the correct option for the judgment question because, as discussed above, at no p is the null definitively overturned. More specifically, under a classical frequentist interpretation of the question, which drug is 'more likely' to result in recovery depends upon the parameters governing the probability of recovery for each drug. As these parameters are unknown and unknowable, option D could be construed as the correct answer under this interpretation. We note that no such difficulty arises under a Bayesian interpretation of the question and for which option A is definitively the correct response.

Were participants approaching the judgment question with the classical interpretation, they would select option D both when presented with the $p < 0.05$ and the $p > 0.05$ versions of the question. However, participants overwhelmingly selected option A when presented with the $p < 0.05$ version of the question whereas they chose option D predominantly when presented with $p > 0.05$ versions. Thus, it seems highly improbable that participants approached the question classically.

We next examined participants' responses to the choice question. The share of participants choosing Drug A in the choice question when the p-value was set to 0.025 was 87% and 94% in the small and large treatment difference conditions respectively. This dropped substantially when the p-value was set to 0.075, 0.125, and 0.175, with no substantial variation across the three (50%, 53%, and 41% respectively when the treatment difference was small and 53%, 52%, and 49% respectively when the treatment difference was large).

In sum, the share of participants who correctly answer each question drops steeply once the p-value falls below 0.05 but is stable thereafter and the magnitude of the treatment difference has no substantial impact on the fraction answering correctly. This dichotomization of responses around the conventional threshold for statistical significance is consistent with the notion that dichotomization of evidence into statistically significant and not statistically significant biases researchers' judgments. Moreover, the lack of any substantial effect of the magnitude of the treatment difference suggests that, within the range of magnitudes we examined (i.e., a more than doubling of the magnitude), whether a result attains or fails to attain statistical significance has a far greater impact on the response than the magnitude of the treatment difference.

We further examined whether the choices made by participants varied by whether the choice was made on behalf of a close other or a distant other. When the p-value was set to 0.075, 0.125, and

0.175, the results showed that participants were more likely to choose Drug A when making a choice for a close other than when making a choice for a distant other (64%, 58%, and 53% respectively versus 39%, 47%, and 36% respectively). This finding supports our proposition that making a choice shifts participants' focus from whether a result attains or fails to attain statistical significance to the available evidence, and that this effect is greater the more consequential the choice for the participant. Nonetheless, we find it striking that–even when faced with a consequential choice–only 50% of participants across the two other conditions chose Drug A when the difference between the two drugs failed to attain statistical significance whereas 90% chose it when the difference attained statistical significance.

We next examined participants' explanations for their answers. Of the 159 participants who incorrectly answered the judgment question when the p-value was above 0.05, 115 suggested that they chose the answer they did because the difference in treatment outcomes failed to attain statistical significance. Many of these responses alluded to the idea that they could not label as evidence differences that did not reach the threshold for statistical significance. Some representative responses were "test for statistical significance was 0.07, which is above the well-established standard of p-value < 0.05."; "H_0 is not rejected"; "p-value is > 0.05, indicating no statistical difference between groups."; and "because the p-value indicated that there was not a significant difference between groups and thus no detectable difference between drug A or B." Other responses that alluded to statistical significance indicated that the lack of statistical significance impacted participants' confidence: "Although the relative difference appears large, statistically the diff is not signif and not knowing more about the sample size and disease pathology or presumably drug mechanism I would not feel confident about 'prescribing' one drug over the other." A small minority of the responses among those assigned to the small treatment difference condition also expressed that the lack of statistical significance combined with the small magnitude of the treatment difference made any difference practically unimportant: "p-value > 0.05 plus from an intuitive standpoint both drugs essentially gave a 50-50 chance of recovery." Such explanations are consistent with our account that researchers' perceptions of evidence are dichotomized around the threshold for statistical significance and that this can manifest itself either as a total disregard for evidence for which $p > 0.05$ or a sharp change in confidence or perceptions of practical significance around $p = 0.05$.

3.4. Robustness

As with Study 1, we tested the robustness of the effect demonstrated in Study 2 by using different wordings for the choice question (i.e., self rather than other) and response options, different question orders, adding additional information (i.e., a posterior probability based on a Bayesian calculation), and different subject populations including researchers in cognitive science (members

of the editorial board of *Cognition*), psychology (members of the editorial board of *Social Psychology and Personality Science*), and economics (recent authors in the *American Economic Review*, the *Quarterly Journal of Economics*, and the *Journal of Political Economy*). To briefly summarize the results of these studies, the pattern of results was not substantially affected by the wording, the question order, or the subject population. However, the inclusion of the Bayesian posterior probability substantially attenuated the proportion of participants answering the likelihood judgment question incorrectly. For additional details, see the online supplementary materials.

4. Discussion

4.1. Summary and Implications

We have shown that researchers across a variety of fields are likely to make erroneous statements and judgments when presented with evidence that fails to attain statistical significance (while undergraduates who lack statistical training are less likely to make these errors; for full details, see the online supplementary materials). These errors pertain to descriptions of data, evaluations of evidence via likelihood judgments, and choices; they do not appear to be moderated by either the magnitude of the treatment difference or the precise size of the p-value associated with the effect. Indeed our quantitative results in tandem with researchers' own explanations of their reasoning suggest that the preponderance of researchers focus primarily or even exclusively simply on whether or not the p-value is below or above the "magic number" of 0.05 when evaluating evidence. This suggests that the dominant NHST paradigm and the rote and recipe-like manner in which it is typically taught and practiced can impair reasoning.

Such errors in interpretation of evidence are likely to have important implications for researchers and clinicians. In the case of the former, as in the illustration presented in the introduction (Messori et al. 1993), researchers may draw incorrect conclusions about findings and their implications when evidence fails to attain statistical significance. In the case of the latter, clinicians may provide inferior treatments or fail to provide superior treatments when evidence fails to attain statistical significance. For example, consider a drug-comparison study lacks adequate power to detect serious side effects; clinicians may dismiss evidence of a doubling of fatalities in one treatment arm if the difference in fatalities between the treatment arms fails to attain statistical significance even though the difference in fatalities constitutes evidence of a difference in risk (Hauer 2004, Healy 2006). On the other hand, spurious findings, for example those based on poor quality data or lacking a plausible mechanism, may be published simply because they attain statistical significance.

4.2. What Can Be Done?

Where does the fault for these errors in the evaluation of evidence lie? We do not believe it lies with statistical training *per se*. It is well-established that evaluating evidence under uncertainty

is difficult and fraught with biases (Tversky and Kahneman 1974), and that, in the aggregate, statistical training is likely to reduce rather than increase such biases (Fong et al. 1986). We also do not believe the fault lies with researchers. Instead, we believe the problem lies with the dichotomization of evidence intrinsic to the NHST paradigm. As noted in our introduction, one should not be surprised that the discrete categorization of evidence leads to categorical thinking. That said, we also do not believe the problem is with the NHST paradigm universally: in many settings null hypothesis significance tests and the concomitant p-values can be useful (though, in the social sciences, the NHST paradigm is much more limited in application due to such issues as the falsity of the nil hypothesis (Cohen 1994, Tukey 1991) and treatment effects that vary at the study- and individual-level (Gelman 2015, McShane and Böckenholt 2014)).

What can be done to ameliorate this potentially deleterious problem? While the tendency of researchers to think of results that attain statistical significance as "true" or "there" and results that fail to attain statistical significance as "zero" or "not there" is clearly problematic, it is not clear that other frequently-proposed approaches such as confidence intervals and Bayesian modeling with non-informative priors do not face the same dichotomization problem as they "can be viewed as mere re-expressions of p-value information in different forms" (Gelman (2015); see also Greenland and Poole (2013a,b)). For instance, were confidence intervals adopted with the stipulation that only 95% confidence intervals that did not overlap zero constituted evidence or were Bayesian modeling adopted with the stipulation that only 95% and higher posterior probabilities constituted evidence, researchers would still suffer from the problems associated with dichotomous thinking.

Rather, a greater focus on effect sizes, their variability, and the uncertainty in estimates of them will naturally lead researchers to think of evidence as lying on a continuum. Instead of thinking of effects as being "there" or "not there," careful consideration of study-level and individual-level variation as well as moderators of this variation can lead researchers to develop deeper and richer theories. These concerns are, as mentioned above, particularly important in the social sciences where the assumption of constant effects is generally untenable (Gelman 2015) thereby making sharp point null hypothesis significance tests questionably meaningful and making the consideration of the generalizability of an effect in other subject populations, at other times, and in different contexts even more important. Finally, even in situations where researchers are not interested in generalizability and thus issues concerning variability are not at play, the uncertainty inherent in statistical estimation and inference can often lead researchers astray particularly under the NHST paradigm. For instance, a large fraction of effects that attain statistical significance are either of the wrong sign or of a greatly biased magnitude when the underlying effect size is small, a problem known as the statistical significance filter (Gelman and Tuerlinckx 2000, Gelman and Weakliem 2009).

Further, researchers should also move away from focusing solely on statistical considerations. Careful attention must be paid to real world costs and benefits in many settings (e.g., pharmaceutical testing and adoption) and these considerations are relevant whether or not underlying differences attain or fail to attain statistical significance (Gelman 2013). Similarly, researchers should pay heed to the size and scientific importance of their results (Sawyer and Peter 1983). They should also focus on issues pertaining to the quality of the data and propriety of the statistical analysis.

In sum, we propose a more holistic and integrative view of evidence that includes consideration of prior and related evidence, the type of problem being evaluated, the quality of data, the effect size, and other considerations.

A counterargument to our position is that there are advantages to objective standards or rules for what constitutes evidence since such standards and rules ostensibly remove personal biases involved in the evaluation of evidence. Such standards might be particularly important when approving a costly but potentially life-saving drug or determining a verdict in a multi-billion dollar court case. However, it should be noted that the p-value is not a purely objective standard: different model specifications and statistical tests for the same data and null hypothesis yield different p-values, and, in many settings, one specification or test is not necessarily superior to the other. More importantly, the same rules of evidence cannot apply to every problem. For example, while it is commonly acknowledged that correlation does not imply causation, a strong correlation can provide evidence of causation–particularly given other evidence. For instance, no randomized controlled trial has ever been performed to show that cigarette smoking leads to lung cancer but the strong correlation between cigarette smoking and lung cancer combined with a plausible biological mechanism constitutes strong evidence that cigarette smoking does indeed cause lung cancer (Mukherjee 2010). On the other hand, a weak correlation–even if highly statistically significant–would be unlikely to constitute strong evidence for a causal relation, particularly in the absence of a plausible mechanism.

4.3. Replicability and Research Practices in Psychology

Our work is relevant to the recent controversy over replicating prior results that some have labeled the replicability crisis (Ioannidis 2005, Brodeur et al. 2012, Yong 2012, Francis 2013) and the increased interest in research practices that has ensued (Asendorpf et al. 2013, Brandt et al. 2014, McShane and Böckenholt 2015 (forthcoming). This interest has been particularly pronounced in psychology as, for example, *Perspectives on Psychological Science* has published several special sections on replicability and research practices (Pashler and Wagenmakers 2012, Spellman 2012, 2013, Ledgerwood 2014) and effective January 2014 the *Psychological Science* Submissions Guidelines

for authors recommends they use "effect sizes, confidence intervals, and meta-analysis to avoid problems associated with null hypothesis significance testing" and points them towards Cumming (2012) and Cumming (2014).

Cumming (2014) concludes "that best research practice is not to use NHST at all", calls for a focus on "estimation, based on effect sizes, confidence intervals, and meta-analysis," and notes that these "techniques are not new, but adopting them widely would be new for many researchers, as well as highly beneficial." Cumming (2012) also calls for a movement away from the dichotomous thinking encouraged by the NHST paradigm towards "estimation thinking [that] focuses on 'how much' " and "meta-analytic thinking [that] focuses on the cumulation of evidence over studies" as exemplified by the following passage of Cumming (2014) that discusses the difference between two means:

> In Figure 3 the difference is 54.0 (95% CI: 7.2 - 100.8), which suggests that our experiment has low precision and is perhaps of little value–although it might still make a useful contribution to a meta-analysis. That is a much better approach than declaring the result 'statistically significant, $p = 0.024$.'

We concur with this call for a greater focus on both estimation and effect sizes (estimation thinking) as well as prior and related evidence (meta-analytic thinking) as these are key to having a more holistic and integrative view of evidence. We also believe that, if heeded, the recommendations offered in Cumming (2012) and Cumming (2014) will generally prove useful.

We also question whether the perception of a replicability crisis is, at least so some degree, yet another ill effect of the NHST paradigm. In particular, consider the standard for what constitutes a successful replication: a subsequent study successfully replicates a prior study if either both fail to attain statistical significance or both attain statistical significance and match direction. This standard is intrinsically tied to the NHST paradigm. However, alternative standards for replication that involve, for example, comparing estimates of effect sizes and their variability from subsequent studies with those from prior studies for consistency in a more holistic sense are possible. Under these alternative standards–which are consistent with the recommendations offered in Cumming (2012) and Cumming (2014)–the replicability crisis may seem like no such thing at all.

4.4. Can It Be Different?

We are certainly not the first researchers to critique the NHST paradigm (Rozenboom 1960, Bakan 1966, Morrison and Henkel 1970, Meehl 1978, Gigerenzer 1987, Rosnow and Rosenthal 1989, Cohen 1994, Loftus 1996, Schmidt 1996, Hunter 1997, Krantz 1999, Hubbard 2004, Gigerenzer 2004,

Gigerenzer et al. 2004, Schwab et al. 2011, Cumming 2014). Nor, we must admit, are our proposed remedies particularly new or original. Yet, despite past criticisms, the dominance of the NHST paradigm appears, at least on the surface, as unassailable as ever.

Part of its persistence, no doubt, is, as noted above, due to the inherent appeal of a seemingly objective standard for evaluating evidence. However, we believe the impact of criticisms of the NHST paradigm has also been blunted because researchers do not realize or acknowledge that their judgments are in fact clouded by considerations of statistical significance (e.g., Hoover and Siegler (2008)).

While researchers may be formally aware that statistical significance at the 0.05 level is a mere convention, our findings highlight that what started as a rule of thumb has evolved into an ironclad principle that indeed affects the interpretation of evidence. We hope that our findings will raise awareness of this phenomenon and thereby lead researchers to adopt a more holistic and integrative view of evidence and to correspondingly reduce their reliance on whether a result attains or fails to attain statistical significance in their interpretation evidence.

Acknowledgments

References

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

Asendorpf, Jens B., Mark Conner, Filip De Fruyt, Jan De Houwer, Jaap J. A. Denissen, Klaus Fiedler, Susann Fiedler, David C. Funder, Reinhold Kliegl, Brian A. Nosek, Marco Perugini, Brent W. Roberts, Manfred Schmitt, Marcel A. G. van Aken, Hannelore Weber, Jelte M. Wicherts. 2013. Recommendations for increasing replicability in psychology. *European Journal of Personality* **27**(2) 108–119.

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

Berger, James O., Donald A. Berry. 1988. Statistical analysis and the illusion of objectivity. *American Scientist* **76** 159–165.

Berger, James O., M. Delampady. 1987. Testing precise hypotheses (with comments). *Statistical Science* **2** 317–352.

Berger, James O., Thomas Sellke. 1987. Testing a point null hypothesis: The irreconciliability of p values and evidence. *Journal of the American Statistical Association* **82**(397) 112–122.

Brandt, Mark J, Hans IJzerman, Ap Dijksterhuis, Frank J Farach, Jason Geller, Roger Giner-Sorolla, James A Grange, Marco Perugini, Jeffrey R Spies, Anna Van't Veer. 2014. The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology* **50** 217–224.

Brodeur, Abel, Mathias Le, Marc Sangnier, Yanos Zylberberg. 2012. Star wars: The empirics strike back. Tech. rep., Paris School of Economics.

Casella, George, Roger L. Berger. 1987. Reconciling bayesian and frequentist evidence in the one-sided testing problem (with comments). *Journal of the American Statistical Association* **82** 106–138.

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

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

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

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

Cumming, Geoff. 2012. *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. Routledge, New York.

Cumming, Geoff. 2014. The new statistics: Why and how. *Psychological Science* **25**(1) 7–29.

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

Fong, Geoffrey T., David H. Krantz, Richard E. Nisbett. 1986. The effects of statistical training on thinking about everyday problems. *Cognitive Psychology* **18** 253–292.

Francis, Gregory. 2013. Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology* **57**(5) 153–169.

Freedman, David, Robert Pisani, Roger Purves. 2007. *Statistics*. 4th ed. W. W. Norton and Company, New York.

Freeman, P. R. 1993. The role of p-values in analysing trial results. *Statistics in Medicine* **12** 1443–1452.

Gelman, Andrew. 2013. Interrogating p-values. *Journal of Mathematical Psychology* **57**(5) 188–189.

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

Gelman, Andrew, John B. Carlin, Hal S. Stern, Donald B. Rubin. 2003. *Bayesian Data Analysis*. 2nd ed. Chapman and Hall/CRC, Boca Raton, FL.

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

Gelman, Andrew, Francis Tuerlinckx. 2000. Type s error rates for classical and bayesian single and multiple comparison procedures. *Computational Statistics* **15** 373–390.

Gelman, Andrew, David Weakliem. 2009. Of beauty, sex and power. *American Scientist* **97** 310–316.

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

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

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

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

Greenland, Sander, Charles Poole. 2013a. Living with p-values: resurrecting a bayesian perspective on frequentist statistics. *Epidemiology* **24**(1) 62–68.

Greenland, Sander, Charles Poole. 2013b. Living with statistics in observational research. *Epidemiology* **24**(1) 73–78.

Haller, H., S. Krauss. 2002. Misinterpretations of significance: a problem students share with their teachers? *Methods of Psychological Research* **http://www.mpr-online.de**.

Hauer, E. 2004. The harm done by tests of significance. *Accident Analysis and Prevention* **36**(495-500).

Healy, D. 2006. The antidepressant tale: figures signifying nothing? *Advances in Psychiatric Treatment* **12** 320–328.

Hommes, D. W., A. Bura, L. Mazzolai andH. Buller, J. W. ten Cate. 1992. Subcutaneous heparin compared with continuous intravenous heparin administration in the initial treatment of deep vein thrombosis. *Annals of Internal Medicine* **116** 279–284.

Hoover, Kevin D., Mark V. Siegler. 2008. Sound and fury: Mccloskey and significance testing in economics. *Journal of Economic Methodology* **15**(1) 1–37.

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

Hubbard, Raymond, R. Murray Lindsay. 2008. Why p values are not a useful measure of evidence in statistical significance testing. *Theory and Psychology* **18**(1) 69–88.

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

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

Kramer, W., Gerd Gigerenzer. 2005. How to confuse with statistics or: The use and misuse of conditional probabilities. *Statistical Science* **20** 223–230.

Krantz, David H. 1999. The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association* **94** 1372–1381.

Ledgerwood, Alison. 2014. Introduction to the special section on moving toward a cumulative science: Maximizing what our research can tell us. *Perspectives on Psychological Science* **9** 610–611.

Loftus, G. R. 1996. Psychology will be a much better science when we change the way we analyze our data. *Current Directions in Psychology* **6** 161–171.

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

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

McShane, Blakeley B., Ulf Böckenholt. 2015 (forthcoming). Planning sample sizes when effect sizes are uncertain: The power-calibrated effect size approach. *Psychological Methods* .

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

Messori, A., G. Scrocarro, N. Martini. 1993. Calculation errors in meta-analysis. *Annals of Internal Medicine* **118** 77–78.

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

Mukherjee, Siddhartha. 2010. *The Emperor of All Maladies: A Biography of Cancer*. Scribner, New York.

Nickerson, R. S. 2000. Null hypothesis statistical testing: A review of an old and continuing controversy. *Psychological Methods* **5** 241–301.

Oakes, M. 1986. *Statistical Inference: A Commentary for the Social and Behavioral Sciences.*. Wiley, New York, NY.

Pashler, Harold, Eric-Jan Wagenmakers. 2012. Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science* **7** 528–530.

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

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

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

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

Schwab, A., E. Abrahamson, W. H. Starbuck, F. Fidler. 2011. Researchers should make thoughtful assessments instead of null-hypothesis significance tests. *Organization Science* **22** 1105–1120.

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

Spellman, Barbara A. 2012. Introduction to the special section on research practices. *Perspectives on Psychological Science* **7** 655–656.

Spellman, Barbara A. 2013. Introduction to the special section on advancing science. *Perspectives on Psychological Science* **8** 412–413.

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

Tversky, Amos, Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* **185**(4157) 1124–1131.

Yong, Ed. 2012. Replication studies: Bad copy. *Nature* **485** 298–300.

Yule, George U., Maurice G. Kendall. 1950. *An introduction to the theory of statistics*. 14th ed. Griffin, London.