American Psychological Association

Psychological Methods

© 2019 American Psychological Association
ISSN: 1082-989X

2020, Vol. 25, No. 5, 596–609
http://dx.doi.org/10.1037/met0000248

Misinterpreting *p*: The Discrepancy Between *p* Values and the Probability the Null Hypothesis is True, the Influence of Multiple Testing, and Implications for the Replication Crisis

Samantha F. Anderson
Arizona State University

Abstract

The *p* value is still misinterpreted as the probability that the null hypothesis is true. Even psychologists who correctly understand that *p* values do not provide this probability may not realize the degree to which *p* values differ from the probability that the null hypothesis is true. Importantly, previous research on this topic has not addressed the influence of multiple testing, often a reality in psychological studies, and has not extensively considered the influence of different prior probabilities favoring the null and alternative hypotheses. Simulation studies are presented that emphasize the magnitude by which *p* values are distinct from the posterior probability that the null hypothesis is true, under an extensive set of conditions including multiple testing. Particular emphasis is placed on *p* values just under .05, given the prevalence of these *p* values in the published literature, though *p* values in other intervals are also assessed. In diverse conditions, results indicate that posterior probabilities favoring the null hypothesis are often far removed from .05, and this pattern quickly gets much worse when multiple testing is conducted. Rather than simply telling researchers that *p* values do not reflect the probability favoring the null hypothesis, as has been done previously, the results presented here allow psychologists to see the evidence provided by various *p* values. These results have particularly topical implications for the replication crisis, for how much weight should be placed on a single study, and for how the term statistical significance should be interpreted, particularly in conditions typical in psychological research.

Translational Abstract

Scientific studies often pit two hypotheses against each other: a null hypothesis (typically a claim of no effect) and an alternative hypothesis (which claims the effect of interest exists). Studies often rely heavily on a quantity known as the *p* value to evaluate the results. The *p* value is commonly believed to imply the likelihood that the null hypothesis is true: A small *p* value would imply that it is unlikely that the null hypothesis is true, leading psychologists to find support for the alternative hypothesis instead. However, *p* values do not, in fact, reveal the likelihood that the null hypothesis is true. This article (a) shows how different the *p* value is from the corresponding likelihood favoring the null hypothesis under a variety of important conditions; (b) investigates the influence that multiple testing (conducting multiple statistical tests on the same or similar sets of variables) and the overall likelihood that the null hypothesis is true have on these differences; and (c) pays particular attention to *p* values falling just under .05, the standard threshold for considering a result "statistically significant." Results indicate that *p* values are often very different from the likelihood that the null hypothesis is true, and multiple testing makes these differences even larger. These results have implications for the replication crisis, for relying too much on single studies, and for how the statistical significance should be interpreted.

Keywords: *p* values, multiple testing, statistical significance, replication

Suppose you are flipping (or scrolling) through a recent issue of your favorite journal and a title catches your attention. Perhaps the title is "Paradoxical Effects of Thought Suppression" (Wegner, Schneider, Carter, & White, 1987), "Cognitive Consequences of Forced Compliance" (Festinger & Carlsmith, 1959), or "First Impressions: Making up Your Mind After a 100-ms Exposure to a Face" (Willis & Todorov, 2006). These titles are attention grabbing, in part because they reflect surprising effects. Common sense would not necessarily imply that trying to suppress thoughts of a white bear would increase thoughts of the white bear, that paying a participant less for a job would yield higher reported enjoyment, or that impressions formed after a brief exposure to a stimulus are highly correlated with those formed without time restriction. Further, suppose (hypothetically) that the *p* value reported for the focal effect is .045, indicating statistical significance by common conventions. What should you make of this *p* value? How should

This article was published Online First December 12, 2019.

I thank Scott E. Maxwell and Michael C. Edwards for comments on a draft of this article. This work has not been previously disseminated.

Correspondence concerning this article should be addressed to Samantha F. Anderson, Department of Psychology, Arizona State University, 950 South McAllister Avenue, Tempe, AZ 85281. E-mail: samantha.f.anderson@asu.edu

you interpret the theoretical conclusion? Should you be convinced?

To place the *p* value in context, recall that across a variety of scientific disciplines, researchers develop scientific hypotheses and seek to determine whether their data support these hypotheses. Scientific hypotheses have historically been tested using statistical hypotheses, and the process of testing these hypotheses focuses on determining whether a proposed effect exists rather than estimating its size. Although effect size estimation is on the rise in psychology, for good reason (e.g., Cumming, 2014), the *existence* of the effect is arguably of primary importance in many circumstances, such as the studies cited above (Morey, Rouder, Verhagen, & Wagenmakers, 2014). These types of studies are often selected for large scale replication attempts (e.g., Open Science Collaboration, 2015), failures of which have received wide attention. Thus, the question of whether or not an effect exists is at a minimum a natural question to investigate, particularly in theory-driven research. Within this hypothesis-testing framework, the *p* value has enjoyed a relatively stable position as the ultimate decider of whether or not an effect is believed to exist (i.e., "the illusion of certainty"; Gigerenzer, 2018, p. 206).

Despite its predominance, the *p* value has almost always been a topic for debate (e.g., Bakan, 1966). Most recently, a 2019 supplement to *The American Statistician* included articles relevant to the use and misuse of *p* values from a variety of perspectives, indicating that the fate of the *p* value is still a concern among statisticians. One of the most frequent misinterpretations is that *p* values signify the posterior probability that the null hypothesis is true (i.e., "Bayesian wishful thinking"; Gigerenzer, 2018, p. 206), a decidedly Bayesian quantity. The posterior probability in this context is the probability that the null hypothesis is true, in light of observed data.[1]

Compelling research has assessed the posterior probability that the null hypothesis is true, for different *p* values under various assumptions (e.g., Berger & Sellke, 1987; Ioannidis, 2005; Sellke, Bayarri, & Berger, 2001). However, previous research, often published outside of psychology, is limited in several important ways. First, and perhaps most importantly, even psychologists and methodologists who are aware of the correct definition of the *p* value may not realize the *magnitude* of the disparity between the *p* value and the probability that the null hypothesis is true. Although it is rather common for articles to *tell* researchers about misunderstandings of *p* values, it is far less common to *show* researchers in a compelling manner. Consequently, this article aims to clearly convey the influence of a variety of factors on the size of the discrepancy. Second, prior research has not accounted for researcher degrees of freedom (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011), which are prevalent in psychology. A particularly common, and potentially justifiable, form of researcher degrees of freedom is multiple testing. Multiple testing (i.e., multiplicity, multiple comparisons) refers to multiple analyses conducted on data, assumed to address the same scientific question. A focal aim of this article is to assess the impact of multiple testing on the evidence for the null hypothesis signified by *p* values falling under common thresholds for statistical significance. Third, when calculating the relevant posterior probabilities, the literature has generally assumed a prior probability that the null hypothesis is true of .5, meaning that before incorporating new data, the null and alternative hypotheses are assumed to be equally likely. A tertiary aim is to determine the impact on the posterior probability when the null hypothesis is a priori highly likely, equally likely, and highly unlikely.

Overall, there has been somewhat of a disconnect between the *p* value debate among methodologists and the current realities of psychological research. This investigation aims to begin to close that gap by investigating *p* values in contexts more reflective of typical research (i.e., in situations of multiple testing). The article is formulated as follows. First, the definition of the *p* value is reviewed in the context of null hypothesis significance testing (NHST), along with research that has described the disparity between *p* values and the probability the null hypothesis is true. Second, researcher degrees of freedom are described, with special attention to multiple testing. Third, simulations aimed at evaluating the *p* value under conditions of multiple testing are presented. Finally, promising future lines of inquiry and unique implications with respect to the replication crisis are discussed.

The *p* Value: A Folkway of a More Primitive Past?

In the words of Rozeboom (1960), "the statistical folkways of a more primitive past continue to dominate the local scene" (p. 417). The *p* value has indeed remained in a position of prestige. In NHST, researchers pit two models against each other: a null hypothesis (H_0), which typically entails a claim of no effect (though, see the Nil Hypothesis section) and an alternative hypothesis (H_a), a claim of a nonzero effect. The *p* value is the probability of obtaining data as extreme or more extreme than the data obtained, given that H_0 is true. In other words, the researcher must take on the role of a devil's advocate and assume that in reality, the world operates under H_0. The researcher then considers whether the data obtained are a reasonable byproduct of such a world, or whether the data would be surprising under this assumption. Although *p* values can be anywhere between zero and one, researchers typically ascribe "statistical significance" to *p* values below the tolerable Type I error probability (e.g., $\alpha = .05$).

Most focally, the *p* value is often misinterpreted as the conditional ($P(H_0\ \mathit{true} \mid \mathit{data})$) or unconditional ($P(H_0\ \mathit{true})$) probability that H_0 is true.[2] This means that a statistically significant *p* value of .05 may be interpreted as a 5% chance that H_0 is true, which is

[1] The same quantity can be considered using Frequentist logic. Consider a body of literature, wherein 100 studies on a topic are conducted, all using the same sample size, analysis, and a single dependent variable, for the sake of simplicity (i.e., parallel studies). In reality but unbeknownst to the researchers, half of the null hypotheses are true (this concept of prior probability will be defined later). Now, consider only the studies of the initial 100 that achieve a certain result (e.g., $p < .05$ or $p < .01$, for the present purposes): The proportion of these studies that report a spurious effect is the Frequentist equivalent of the posterior probability that the null hypothesis is true.

[2] To be clear, despite being interpreted as an indicator of evidence favoring H_0, *p* values are not valid measures of evidence favoring H_0 in the statistical sense due to lacking the quality of *consistency* (Rouder et al., 2009; Wagenmakers, 2007). Given that *p* values converge toward zero with increasing sample size, and H_0 is always rejected in the large sample limit when it should be, *p* values do represent a consistent test under H_a. However, under H_0, *p* values follow a uniform (0, 1) distribution and thus provide an inconsistent test, as the *p* value does not converge to a *p* value of 1 with increasing sample size and the researcher will still mistakenly reject H_0 α% of the time (Rouder et al., 2009).

sufficiently low enough to imply strong support for H_a. Despite methodologists "sounding the alarm about these matters for decades" (Wasserstein & Lazar, 2016, p. 130), the misinterpretation persists in psychology almost as much as the *p* value itself (Stern, 2016). High percentages (median = 43.5%) of this type of misinterpretation have been found with academics, even methodologists, from a variety of countries (e.g., Badenes-Ribera, Frias-Navarro, Iotti, Bonilla-Campos, & Longobardi, 2016; see also Gigerenzer, 2018; Haller & Kraus, 2002; Kline, 2013; Laber & Shedden, 2017).

It is not clear why these misinterpretations remain. One possibility is that much of the relevant research has appeared in the statistical and medical literatures, rather than psychology. Previous methodological articles have described that *p* values are different from $Pr(H_0\ true \mid data)$ but have typically not emphasized *how* different under a variety of conditions. Moreover, recent research points to alarming inaccuracies presenting *p* values and statistical significance in a large proportion (89%) of psychology textbooks, implying that some of the misinterpretation may stem from a student's first exposure to the concept of a *p* value (Cassidy, Dimova, Giguère, Spence, & Stanley, 2019). Another possibility is referenced in Cohen's (1994) quote: "It does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does" (p. 997). Psychologists often want the *p* value to provide evidence for the probability of H_0, and this desire is difficult to overcome. Whatever the reason, the tendency to misinterpret has implications for how the field of psychology conducts, builds upon, and evaluates empirical research. One example of this is the tendency to rely on a single study to demonstrate the existence of an effect, despite increased attention to replication. A goal of the present study, then, is to provide a concrete demonstration of the extent to which *p* values do or do not reflect the posterior probability H_0 is true.

What Evidence for H_0 Do *p* Values Reveal?

Once it is appreciated that *p* values are in fact misinterpreted as the posterior probability H_0 is true, it might be asked, what do *p* values imply about the posterior probability favoring H_0? Considering the meaning of a statistically significant result, Ioannidis (2005) focused on the positive predictive value (PPV), the posterior probability that $\mathrm{H_a}$ is true:[3] Given factors such as sample size, effect size, and the prior odds that H_0 is true, PPVs above 50% were rare, leading Ioannidis (2005) to claim "most research findings are false" (p. 699). The direct counterpart to PPV is the posterior probability that H_0 is true, given $p < .05$, or

$$P(H_0\ true \mid p < .05) = \frac{\alpha P(H_0\ true)}{\alpha P(H_0\ true) + (1-\beta) P(H_0\ false)}. \tag{1}$$

Equation (1) was derived from Bayes's theorem, which more generally emphasizes the difference between P(A | B) and P(B | A). It is evident that the posterior probability that H_0 is true is directly dependent on α, statistical power (1 − β), and the prior probabilities that H_0 ($P(H_0\ true)$) and H_a ($P(H_\alpha\ true)$) are true.

Rather than focusing on *p* values falling anywhere in the statistically significant range ($p < .05$) as Ioannidis (2005) did, Berger and Sellke (1987) and Sellke, Bayarri, and Berger (2001) focused on $P(H_0\ true \mid p)$, or the posterior probability that H_0 is true for a specific *p* value. Because *p* is a continuous random variable, the probability of any exact *p* value is zero, so the discussion that follows can be thought of as a limit as *p* approaches a specific value. This posterior probability can also be defined in terms of Bayes's theorem, as

$$P(H_0\ true \mid p) = \frac{P(p \mid H_0\ true) P(H_0\ true)}{P(p \mid H_0\ true) P(H_0\ true) + P(p \mid H_0\ false) P(H_0\ false)}. \tag{2}$$

To summarize their findings, Berger and Sellke (1987) wrote "*p* gives a misleading interpretation as to the validity of H_0, from almost any evidentiary viewpoint" (p. 112). For example, for a prior probability of .5, a sample size of $n = 50$, and a standardized mean difference of δ = 0.392 (roughly a "small-medium" effect size by Cohen's (1988) conventions), a *p* value of .05 is consistent with a posterior probability of .52 that H_0 is true, meaning that despite a statistically significant result, H_0 is more likely than H_a. Berger and Sellke (1987) provided lower bounds for the posterior probabilities of different *p* values under a range of prior distributions for H_a. Sellke et al. (2001) followed up with a less theoretical examination, focusing on the posterior probabilities for *p* values around .05 for one dependent variable (DV). The posterior probability of H_0 was a monotonically increasing function over increasing prior proportions of true nulls, with a lower bound of just over .20 when the prior proportion of true nulls was .5. Importantly, the results indicate that *p* values around .05 do not constitute strong evidence against H_0, and may even constitute evidence *for* H_0. In some cases "the probability of getting a *p* value near .05, when $\mathrm{H_1}$ is true, cannot be much bigger than the probability of getting a *p* value near .05, when $\mathrm{H_0}$ is true" (Sellke et al., 2001, p. 64).

The posterior probabilities for a specific *p* value and for $p < .05$ make very different statements; the former is particularly important when the exact *p* value is available (Sellke et al., 2001). A related posterior probability of scientific interest is $P(H_0\ true \mid LL < p < UL)$, where LL and UL represent the lower and upper limits, respectively, of an interval within which a *p* value falls. This has not been explicitly considered, though Equations (1) and (2) represent this intermediate probability in two extremes. The formulation with respect to Bayes's theorem would simply replace each *p* in Equation (2) with $LL < p < UL$.

This operationalization may be of particular interest due to the "peculiar prevalence" of *p* values just under .05 (e.g., between .045 and .05; Krawczyk, 2015; Masicampo & Lalande, 2012, p. 2271), evidence of both publication bias and researcher degrees of freedom. Whereas publication bias results in inflated effect sizes, bias due to researcher degrees of freedom works in the opposite direction, resulting in "'large' significant *p* values" (Simonsohn, Nelson, & Simmons, 2014a, p. 670). These findings relate broadly to the *p*-curve (Simonsohn, Nelson, & Simmons, 2014b) which depicts the expected percentages of different significant *p* values for various levels of power and sample size. When H_0 is false, small *p* values (i.e., $p < .01$) should be more likely than large significant

[3] From the Frequentist perspective, PPV can be thought of as the proportion of $p < .05$ effects in a research body of literature that reflect true results, again assuming parallel studies.

p values (i.e., *p* just under .05), but it is clear that researcher degrees of freedom alter this pattern.

The Influence of the Prior Probability

Previous research has not provided an extensive assessment of the influence of the prior probability H_0 is true on the corresponding posterior probability. Berger and Sellke (1987) typically assumed H_0 and H_a were equally probable, noting its "obvious intuitive appeal in scientific investigation as being 'objective'" (Berger & Sellke, 1987, p. 115). The authors noted that, under the conditions assessed, a *p* value of .05 only reflected a .05 posterior probability for H_0 when the prior probability of H_0 was .15. Relatedly, Ioannidis (2005) provided a brief example of how low prior odds could influence PPV, but noted that future research should assess and "improve our understanding of the range" (p. 701) of prior odds in various research literatures.

Indeed, there is good reason to study the influence of prior probability for *p* values of interest. Prior information in favor of H_0 varies within and across research areas. For example, in clinical trials, the ethical principle of equipoise states that the experimenter or field should be "in a state of genuine uncertainty regarding the comparative merits of Treatments A and B for population P" (Freedman, 2017, p. 141), and a prior of .5 may be reasonable. Alternatively, there are situations where .5 is less sensible. An extreme example of a high prior probability of H_0 is the extrasensory perception phenomenon (e.g., Bem, 2011) and made more famous by highly publicized failures to replicate (e.g., Galak, LeBoeuf, Nelson, & Simmons, 2012). In order to convince others in the field of such an unexpected effect, the prior would need to be biased toward H_0. Genome wide association studies (GWAS) are a less dramatic example, where many single nucleotide polymorphisms (SNPs) are assessed, most of which will not show a phenotypic effect. More generally, the null hypothesis often represents the current theory and may be seen as more likely than the new theory trying to shake things up. Alternatively, confirmatory studies and late-stage clinical trials have already accrued evidence in favor of H_a, and thus, the prior probability H_0 is true may logically be lower (Ioannidis, 2005).

When the prior probability of H_0 is high, statistically significant results can seem especially surprising, which can make the effect appear more impressive (Prentice & Miller, 1992). In fact, selecting DVs unlikely to be influenced by an independent variable (unless the unexpected theory turns out to be true) was a historically popular technique used to emphasize an effect's importance (see Asch, 1951; and Efran, 1974 for examples). Abelson (1997) wrote "null-hypothesis tests are cogent in scrutinizing surprising results that critics doubt" (p. 14). Yet, it is in these situations that $p < .05$ may be the least convincing evidence in favor of H_a. For example, in the Reproducibility Project Psychology (Open Science Collaboration, 2015), the more surprising the original effect was, the less likely it was to replicate ($r = -.244$). It is important to determine what impact this has on the posterior probability in favor of H_0.

The Potential Pitfall of Multiplicity

Psychologists may be prone to "design, analytic, or reporting practices that have been questioned because of the potential for the practice to be employed with the purpose of presenting biased evidence in favor of an assertion" (Banks, Rogelberg, Woznyj, Landis, & Rupp, 2016, p. 3). These practices have been given many names, from the more benign researcher degrees of freedom (Simmons et al., 2011), to questionable research practices (John et al., 2012), to *p*-hacking (Gelman & Loken, 2014). Researcher degrees of freedom may be incentivized by publication bias. In fact, some have argued that researcher degrees of freedom are a primary reason for the paradox that underpowered studies are published at higher than expected rates (Bakker, Hartgerink, Wicherts, & van der Maas, 2016). However, as Gelman and Loken (2014) noted, these practices may not be motivated out of a desperate desire to dig up anything significant; the term researcher degrees of freedom is used here to emphasize a not necessarily sinister process.

In particular, multiple testing, which includes uncorrected testing of multiple comparisons on the same DV, tests of multiple DVs, and tests of multiple predictors in multiple regression, is extremely common in psychology and medicine (Cribbie, 2017; Vickerstaff, Ambler, King, Nazareth, & Omar, 2015). For example, failing to report all of a study's dependent measures was not only the most common researcher degrees of freedom (63.4% of participants indicated having engaged in the practice, even without additional measures implemented to motivate truth telling) surveyed by John et al. (2012) in a sample of over 2,000 psychologists, but it was also rated as the most defensible. Krawczyk (2015) wrote that in situations of multiple possible DVs, researchers may "report the one that 'works best' where two or more options were initially considered" (p. 3). Other situations of multiple testing are more sincere and may be seen as central to the scientific process, such as when multiple DVs are of interest or multiple predictors are of interest in multiple regression. With regard to the former, multivariate tests are still uncommon, and often are followed by univariate tests (Counsell & Harlow, 2017; Vickerstaff et al., 2015). With regard to the latter, although multiple comparison procedures have become somewhat established for analysis of variance, correction for multiplicity in multiple regression is rare (Cribbie, 2017).

Despite that many potential defensible tests exist and that adjustment is uncommon in some domains, Maxwell, Delaney, and Kelley (2018) wrote "it is especially important to realize that failing to control for multiple testing may play a major role in contributing to the disappointing failure rate in attempts to replicate published studies" (p. 216). Relatedly, Gelman and Loken (2014) noted that unadjusted multiple testing leads to statistically significant, but untrustworthy results. Researchers "faced with multiple reasonable measures can reason (perhaps correctly) that the one that produces a significant result is more likely to be the least noisy measure, but then decide (incorrectly) to draw inferences based on that one only" (p. 461; see Humphreys, de la Sierra, & van der Windt, 2013). Moreover, if α is not corrected, Type I errors are more likely to be published, and these tend to linger, as subsequent replications will likely produce nonsignificant results, which will not be published (Greenwald, 1975). However, there is still disagreement over when and how to control the α level for multiple testing (Cribbie, 2017). Whether or not multiple testing is seen as researcher degrees of freedom or a valid feature of scientific investigations, determining how it interacts with the evidence conveyed by *p* values will add a useful perspective to the current debate surrounding the future of NHST.

Method

A simulation study was conducted to assess the discrepancy between *p* values and the posterior probability in favor of H_0 under a variety of relevant conditions, with the aim to provide evidence of the magnitude of this discrepancy.[4] Most importantly, the influence of unadjusted multiple testing on the posterior probability of H_0 was assessed, given that this has received little attention in the literature. The influence of the prior probability in favor of H_0 was assessed, in addition to the effect size associated with H_a and the sample size. To be thorough, *p* values under recommended α levels were considered, as well as subsets of *p* values in the small interval just under each α. These manipulations serve to make the posterior probability of H_0 transparent under conditions relevant to psychological research.

Manipulated Factors

For simplicity, it is assumed that a researcher is conducting an independent samples *t* test, with H_0: $\delta = 0$ and H_a: $\delta \neq 0$. δ is the population value of the standardized mean difference, or

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}, \tag{3}$$

where μ_1 and μ_2 are the population means on the DV of the two groups and σ is the population standard deviation of scores. Five alternative values of δ were considered: 0.1, 0.3, 0.5, 0.7, and 0.9. Additionally, the sample size was varied between $n = 20$ and $n = 60$ per group, reflecting the lower and higher ends of typical sample sizes for designs of this type in psychology (e.g., Marszalek, Barber, Kohlhart, & Holmes, 2011). The focal manipulation was the number of DVs tested by the researcher, varied between values of 1, 3, and 5. For conditions with multiple DVs, the correlation among the DVs was medium ($\rho = .3$).[5] The prior probability in favor of H_0 was varied between conditions of low (.2), equipoise (.5), and high (.8). The α level was varied between .05, .01, and .005 to represent common thresholds for declaring statistical significance. Finally, one set of conditions explored *p* values falling anywhere under α (i.e., $p < .05$, $p < .01$, and $p < .005$), and a second explored *p* values in the small .005-length interval just under α (i.e., $.045 < p < .05$ and $.005 < p < .01$).

Procedure

Monte Carlo simulations were conducted using R Statistical Software. In each replication, samples based on a null condition ($\delta = 0$) and the five alternative conditions were generated. In the single DV condition, each sample *Y* was drawn from a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$. In the three DVs condition, each sample *Y* was drawn from a multivariate normal distribution, with mean vector μ and covariance matrix $\sum$ as follows:

$$\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \; \sum = \begin{pmatrix} 1 & & \\ .3 & 1 & \\ .3 & .3 & 1 \end{pmatrix}. \tag{4}$$

Samples in the five DVs condition were generated in the same manner. Assuming parallel measures for the DV is appropriate to consider the influence of researchers who conduct multiple tests related to a single general scientific hypothesis. Parallel measures are a clear example of this type of researcher degrees of freedom. The specified effect size δ was then added to each *Y* score in the treatment group. For each DV, an independent samples *t* test was conducted on the data, and the *p* values were saved. The minimum *p* value from each set of tests was used for calculation of posterior probabilities favoring H_0 and Bayes factors. Here, the assumption is that a researcher who conducts multiple testing chooses the smallest *p* value to report (or reviewers may suggest this), if at least one *p* value is below .05. Other decision rules are possible, such as selecting the first statistically significant *p* value. Moreover, a multivariate test could be performed on the DVs jointly or a multiple comparison procedure could be conducted (e.g., Bonferroni correction), but the goal is to illustrate researcher degrees of freedom rather than best practices. To assure that a large enough number of *p* values fell within the ranges under investigation, 50,000 replications were run in each condition.

Quantities Assessed

Two output measures were assessed in the simulation. First, focally, the posterior probability in favor of H_0 for *p* values in the various intervals was calculated, using the variations on Equation (1) described in the What Evidence for H_0 Do *p* Values Reveal? section. Second, Bayes factors were calculated as a different metric to describe the strength of evidence provided by *p* values. The Bayes factor is the ratio of the marginal likelihood of the data, *D*, under H_a to H_0, and describes how much to amend the prior odds of H_a in response to new data:

$$BF_{10} = \frac{P(D \mid H_a)}{P(D \mid H_0)}. \tag{5}$$

Additional Simulation Conditions

In addition to the focal simulations, three additional conditions were run to test additional hypotheses. First, some have argued that the concept of H_0 itself is absurd, given that it is unrealistic to assume any effect is exactly zero, as is done with a point null hypothesis (e.g., Cohen, 1994; Meehl, 1967). However, point null hypotheses are good approximations to small interval hypotheses (nil hypotheses) that claim a negligible effect (e.g., Berger & Sellke, 1987; Zellner, 1984). Although it is unlikely that specifying a nil hypothesis will have a substantial impact on the results (see Berger & Sellke, 1987, p. 119), an additional condition ($n = 60$; prior probability = .5, 3 DVs) was run, where H_0 was specified such that $\delta = 0.01$ rather than $\delta = 0$. Second, a large sample size condition ($n = 150$ per group; prior probability = .5, three DVs), was included, given the push to conduct more powerful and precise studies (e.g., Anderson, Kelley, & Maxwell, 2017; Bakker et al.,

[4] The posterior probability favoring H_0 for conditions in which *p* values under .05, .01, and .005 are evaluated for a single DV can be determined analytically, by using Equation (1). However, all findings are presented via a simulation study for consistency.

[5] Differences in results were negligible when the correlation among the DVs was set at .7 instead of .3.

2016). Third, although the focal simulations address *p* values falling into specific, empirically meaningful intervals, the design may make the functional relationship between *p* values and the posterior probability favoring H_0 less clear among the other varied factors. Thus, a final simulation was conducted to make this correspondence explicit. The posterior probability favoring H_0 was considered for *p* values across the range of statistical significance (falling in the .005-length interval just below .05, .04, .03, .02, and .01), conditioning on three effect sizes (δ = 0.1, δ = 0.5, δ = 0.9; *n* = 60 per group, one DV).

Simulation Results and Discussion

Posterior Probability H_0 True for *p* Values in Interval Just Under α

When a *p* value just under .05 is obtained, does it imply a low degree of evidence for H_0? Results are shown in the bottom panels of Tables 1, 2, and 3, and selected results are depicted in Figures 1 and 2. The posterior probability in favor of H_0 decreased then increased as the effect size increased and did not uniformly decrease with the larger sample size. This nonmonotonicity may seem surprising, particularly in terms of the high posterior probabilities in favor of H_0 for large effects. However, the pattern results from the fact that large effects that result in *p* values just under .05 are less convincing because these effect sizes should end up resulting in smaller *p* values. In other words, the *p* value linked to a larger effect size or sample size should ideally be below the interval just under .05, and *p* values that do not fall below the interval do not provide strong evidence against H_0. This point is elaborated upon in the Discussion section.

Table 1
Posterior Probabilities Favoring H_0 for p Values Less Than or Directly Under α: One Dependent Variable

	n = 20					*n* = 60				
δ	.1	.3	.5	.7	.9	.1	.3	.5	.7	.9
$p < .05$										
Prior = .8	.77	.57	.38	.26	.21	.72	.36	.21	.18	.17
Prior = .5	.46	.25	.13	.08	.06	.39	.12	.06	.05	.05
Prior = .2	.17	.08	.04	.02	.02	.14	.03	.02	.01	.01
BF	*W*	*W*	*P*	*P*	*P*	*W*	*P*	*P*	*P*	*P*
$p < .01$										
Prior = .8	.77	.48	.23	.12	.07	.68	.20	.07	.05	.04
Prior = .5	.45	.19	.07	.03	.02	.34	.06	.02	.01	.01
Prior = .2	.17	.05	.02	.01	.00	.12	.02	.00	.00	.00
BF	*W*	*P*	*P*	*S*	*S*	*W*	*P*	*S*	*S*	*S*
$p < .005$										
Prior = .8	.77	.44	.19	.08	.05	.65	.15	.04	.02	.02
Prior = .5	.46	.17	.05	.02	.01	.32	.04	.01	.01	.01
Prior = .2	.17	.05	.01	.01	.00	.11	.01	.00	.00	.00
BF	*W*	*P*	*P*	*S*	*S*	*W*	*S*	*S*	*VS*	*VS*
$.045 < p < .05$										
Prior = .8	.78	.65	.57	.50	.58	.74	.57	.61	.86	.99
Prior = .5	.47	.33	.25	.20	.26	.42	.25	.28	.62	.96
Prior = .2	.18	.11	.08	.06	.08	.15	.08	.09	.29	.86
BF	*W*	*W*	*P*	*P*	*W*	*W*	*P*	*W*	*	*
$.005 < p < .01$										
Prior = .8	.76	.52	.30	.20	.17	.70	.30	.19	.28	.71
Prior = .5	.44	.21	.10	.06	.05	.37	.10	.05	.09	.38
Prior = .2	.17	.06	.03	.01	.01	.13	.03	.01	.02	.13
BF	*W*	*P*	*P*	*P*	*P*	*W*	*P*	*P*	*P*	*W*

Note. Prior = prior probability favoring H_0; *n* = per-group sample size; BF = Bayes factor indicating evidence in favor of H_a; δ = standardized mean difference effect size (population value of Cohen's *d*). Notation for the Bayes factors is as follows (based on Kass & Raftery, 1995): W = weak evidence (BF < 3); *p* = positive evidence (3 < BF < 20); S = strong evidence (20 < BF < 150); VS = very strong evidence (BF > 150); * = evidence favor H_0.

Table 2
Posterior Probabilities Favoring H_0 for p Values Less Than or Directly Under α: Three Dependent Variables

	n = 20					*n* = 60				
δ	.1	.3	.5	.7	.9	.1	.3	.5	.7	.9
$p < .05$										
Prior = .8	.77	.61	.46	.39	.36	.72	.44	.36	.35	.35
Prior = .5	.46	.28	.18	.14	.12	.39	.17	.12	.12	.12
Prior = .2	.18	.09	.05	.04	.03	.14	.05	.03	.03	.03
BF	*W*	*W*	*P*	*P*	*P*	*W*	*P*	*P*	*P*	*P*
$p < .01$										
Prior = .8	.76	.48	.26	.16	.12	.66	.23	.12	.11	.10
Prior = .5	.44	.19	.08	.04	.03	.33	.07	.03	.03	.03
Prior = .2	.17	.05	.02	.01	.01	.11	.02	.01	.01	.01
BF	*W*	*P*	*P*	*S*	*S*	*W*	*P*	*S*	*S*	*S*
$p < .005$										
Prior = .8	.75	.44	.21	.11	.07	.62	.17	.07	.05	.05
Prior = .5	.43	.16	.06	.03	.02	.29	.05	.02	.01	.01
Prior = .2	.16	.05	.02	.01	.00	.09	.01	.00	.00	.00
BF	*W*	*P*	*P*	*S*	*S*	*W*	*P*	*S*	*S*	*S*
$.045 < p < .05$										
Prior = .8	.79	.70	.70	.90	.92	.75	.72	.91	1.0	1.0
Prior = .5	.48	.37	.37	.51	.74	.43	.39	.73	.99	1.0
Prior = .2	.19	.13	.13	.20	.42	.16	.14	.40	.95	1.0
BF	*W*	*W*	*W*	*	*	*W*	*W*	*	*	*
$.005 < p < .01$										
Prior = .8	.77	.53	.37	.32	.41	.70	.37	.43	.84	1.0
Prior = .5	.45	.22	.13	.11	.15	.36	.13	.16	.57	.98
Prior = .2	.17	.07	.04	.03	.04	.12	.04	.04	.25	.94
BF	*W*	*P*	*P*	*P*	*P*	*W*	*P*	*P*	*	*

Note. Prior = prior probability favoring H_0; *n* = per-group sample size; BF = Bayes factor indicating evidence in favor of H_a; δ = standardized mean difference effect size (population value of Cohen's *d*). Notation for the Bayes factors is as follows (based on Kass & Raftery, 1995): W = weak evidence (BF < 3); *p* = positive evidence (3 < BF < 20); S = strong evidence (20 < BF < 150); VS = very strong evidence (BF > 150); * = evidence favor H_0.

For a single DV, a *p* value just under .05 was not typically associated with a small, convincing, posterior probability, as often assumed, and the posterior probability increased as the prior probability in favor of H_0 increased. Even when H_0 was very unlikely a priori, the minimum probability in favor of H_0 was .06, greater than .05, but could be as high as .86 (median = .10). In the more objective condition with a prior probability of .5, the probability

Table 3
Posterior Probabilities Favoring H_0 for p Values Less Than or Directly Under α: Five Dependent Variables

	$n = 20$					$n = 60$				
δ	.1	.3	.5	.7	.9	.1	.3	.5	.7	.9
$p < .05$										
Prior = .8	.77	.64	.52	.47	.46	.73	.51	.46	.46	.46
Prior = .5	.46	.31	.21	.18	.17	.41	.21	.18	.17	.17
Prior = .2	.18	.10	.06	.05	.05	.15	.06	.05	.05	.05
BF	*W*	*W*	*P*	*P*	*P*	*W*	*P*	*P*	*P*	*P*
$p < .01$										
Prior = .8	.75	.50	.29	.20	.17	.67	.28	.17	.16	.16
Prior = .5	.43	.20	.09	.06	.05	.33	.09	.05	.05	.05
Prior = .2	.16	.06	.02	.02	.01	.11	.02	.01	.01	.01
BF	*W*	*P*	*P*	*P*	*S*	*W*	*P*	*P*	*S*	*S*
$p < .005$										
Prior = .8	.75	.45	.23	.13	.10	.64	.21	.10	.09	.09
Prior = .5	.43	.17	.07	.04	.03	.31	.06	.03	.02	.02
Prior = .2	.16	.05	.02	.01	.01	.10	.02	.01	.01	.01
BF	*W*	*P*	*P*	*S*	*S*	*W*	*P*	*S*	*S*	*S*
$.045 < p < .05$										
Prior = .8	.78	.75	.78	.89	.97	.78	.81	.97	1.0	1.0
Prior = .5	.48	.43	.47	.67	.90	.47	.51	.90	1.0	1.0
Prior = .2	.19	.16	.18	.34	.70	.18	.21	.70	1.0	1.0
BF	*W*	*W*	*W*	*	*	*W*	*	*	*	*
$.005 < p < .01$										
Prior = .8	.76	.56	.43	.44	.60	.70	.44	.62	.96	1.0
Prior = .5	.44	.24	.16	.16	.28	.36	.16	.29	.87	1.0
Prior = .2	.16	.07	.04	.05	.09	.13	.05	.09	.62	1.0
BF	*W*	*P*	*P*	*P*	*W*	*W*	*P*	*W*	*	*

Note. Prior = prior probability favoring H_0; n = per-group sample size; BF = Bayes factor indicating evidence in favor of H_a; δ = standardized mean difference effect size (population value of Cohen's d). Notation for the Bayes factors is as follows (based on Kass & Raftery, 1995): W = weak evidence (BF < 3); p = positive evidence (3 < BF < 20); S = strong evidence (20 < BF < 150); VS = very strong evidence (BF > 150); * = evidence favor H_0.

favoring H_0 ranged from .20 to .96 (median = .31). When H_0 was rather likely a priori, the minimum probability favoring H_0 was .50, meaning that, at best, a p value just under .05 in that condition provides equal evidence for H_0 as H_a (median = .63, maximum = .99).

Importantly, the posterior probability generally increased as the number of DVs increased, often substantially. When three DVs were tested, the minimum probability favoring H_0 was .13, .37, and .70 for priors of .2, .5, and .8, respectively. When five DVs were tested, the corresponding minima rose to .18, .43, and .75. To emphasize, when H_0 and H_a were equiprobable a priori and three DVs were tested, a p value of just under .05 at best indicated a 37% chance that H_0 is true, much less convincing than a 5% chance. These patterns, and particularly the dramatic influence of multiple testing, are evident in Figure 1, which shows posterior probabilities favoring H_0 for p values just under .05 with one, three, and five DVs, at a more precise gradation of effect sizes (δ = 0.1 − 0.9 by increments of 0.1; n = 60).

For p values just under .01, the ranges were lower, but still high. Assuming a prior of .5, the ranges of prior probabilities favoring H_0 were .05 to .44 (one DV), .11 to .98 (three DVs), and .16 to 1.00 (five DVs). It is interesting to note that, under a prior of .5, the minimum posterior probability favoring H_0 for a p value just under .01 in the single DV condition was .05. Thus, from one perspective, a p value just under .01 can align with the misinterpretation of a p value of .05. If the field really does want the words "statistically significant" to convey a 5% chance that H_0 is true, this is at least possible with a p value just under .01. After so many years of focusing on an imaginary dividing line between .049 and .051, researchers may not typically attribute much of a difference to p values of .01 versus .05, given that conventionally, both are members of a statistically significant category. However, these results indicate that it may be meaningful to differentiate .05 from .01 (and from .001, e.g.), which is in line with efforts to encourage reporting exact p values and removing mention of statistical significance (Wasserstein, Schirm, & Lazar, 2019). Despite this good news, many conditions still resulted in a p value of just under .01 yielding more evidence for H_0 than H_a, particularly when multiple testing was conducted, which is evident in Figure 2. In fact, when multiple DVs were tested, the results became more similar to what a p value of just under .05 conveys for a single DV.

Posterior Probability H_0 True for p Values Under α

In addition to p values just under .05 and .01, ranges encompassing p values anywhere under .05, .01, and .005 were investigated (shown in the top panels of Table 1–Table 3). The first two of these are diluted forms of the interval conditions, given that p values under .05 (or .01) include p values directly below and

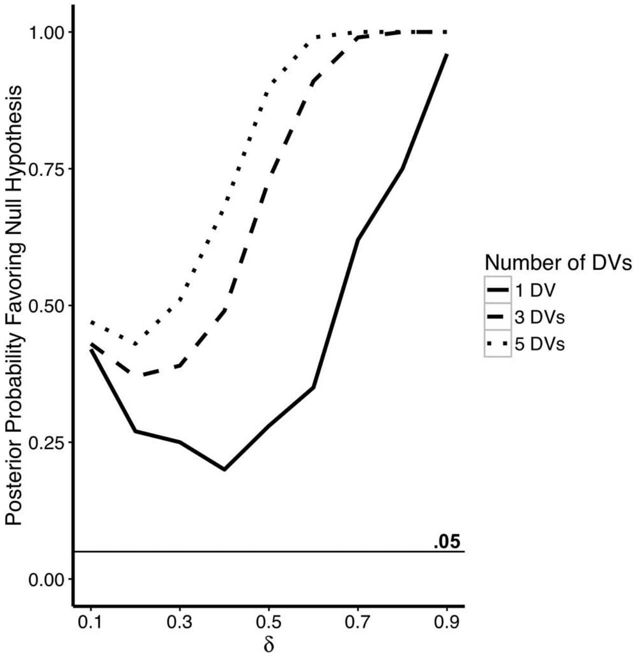

Figure 1. Plot of posterior probabilities that the null hypothesis is true for an independent samples t test for p values just under .05 with a sample size of n = 60 per group. The posterior probability is calculated for each effect size δ from 0.1 to 0.9 in increments of 0.1. The horizontal line indicates a posterior probability of .05.

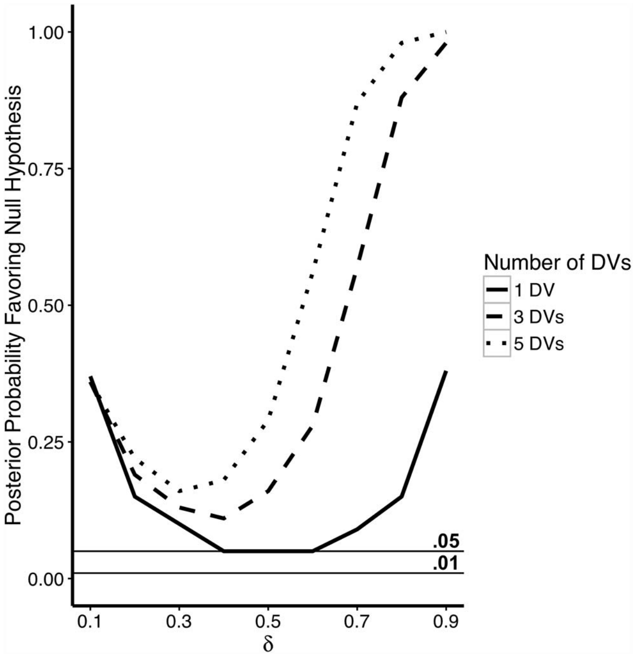

Figure 2. Plot of posterior probabilities that the null hypothesis is true for an independent samples *t* test for *p* values just under .01 with a sample size of $n = 60$ per group. The posterior probability is calculated for each effect size δ from 0.1 to 0.9 in increments of 0.1. The horizontal lines indicate posterior probabilities of .05 and .01.

smaller *p* values. In terms of prominent patterns, the posterior probability in favor of H_0 increased as effect size and sample size decreased, and as α became less stringent, as expected. When considering all *p* values under the typical α of .05, the only conditions where the minimum posterior probability was close to .05 were for a single DV or a highly improbable H_0. The posterior probability correspondingly increased as the prior probability in favor of H_0 increased (e.g., range .01–.17 for an unlikely H_0 vs. .17–.77 for a more likely H_0, assuming a single DV and $p < .05$). Moreover, similarly to the other conditions, the posterior probability increased substantially as the number of DVs increased. For one DV, the minimum posterior probabilities were .05, .01, and .01 for $p < .05$, .01, and .005, respectively, assuming a prior probability of .5 (the maximum values were much higher). The corresponding minima for three DVs were .12, .03, and .01, and for five DVs were .17, .05, and .02. Considering a more stringent α of .005, the posterior probability that H_0 is true was .05 or below in more than half the conditions, assuming a single DV and a prior of .5. When three and five DVs were tested, the corresponding ranges were .01 to .43, and .02 to .43, respectively.

Bayes Factors

Bayes factor interpretations (based on Kass & Raftery, 1995) are presented in the bottom row of each condition in Table 1–Table 3.[6] Out of all of the conditions, only two times did the Bayes factor indicate very strong evidence in favor of H_a (both when considering *p* values below .005; $\delta \geq 0.7$). When focusing on *p* values just under .05, even when only a single DV was tested, the largest Bayes factor was 4.04, barely crossing the threshold for positive evidence in favor of H_a. Over half of the conditions (70%) resulted in Bayes factors indicating weak evidence in favor of H_a or even favoring H_0. When three DVs were tested, no Bayes factors indicated positive evidence in favor of H_a, and half of the conditions yielded a Bayes factor favoring H_0. In other words, a conventionally statistically significant *p* value often provides more evidence in favor of H_0 than against H_0. This phenomenon also occurred for *p* values just under .01, but less frequently. Finally, when five DVs were tested, the number of conditions resulting in a Bayes factor favoring H_0 rose to 60%. Thus, almost always, *p* values just under .05 represent weak evidence in favor of H_a and this gets dramatically worse as the number of tests increases. These findings are in line with prior research finding that, for *p* values between .01 and .05, Bayes factors often yielded only anecdotal evidence for H_a (Wetzels et al., 2011).

Additional Conditions

Results of the nil hypothesis condition are presented in Table 4. Defining the null hypothesis instead as a nil hypothesis did not meaningfully alter results. Results for the large sample size condition ($n = 150$) are also shown in Table 4, compared with the moderate sample size ($n = 60$) condition. In most conditions, the influence of the large sample size was small. In conditions where the posterior probabilities differed, the results for the large sample size were sometimes larger and sometimes smaller than those for the small sample size. Finally, a simulation assessed the posterior probabilities favoring H_0 for *p* values across the range of standard statistical significance, to more explicitly address the relationship between *p* values and the posterior probability favoring H_0. For three effect sizes, Figure 3 depicts this relationship, for a single DV a moderate sample size ($n = 60$), and under an equiprobable set of priors. Thus, when other factors are held constant, *p* values are directly related to the posterior probability favoring H_0, such that as the *p* value approaches zero, the posterior probability favoring H_0 correspondingly decreases, though the nature of this decrease depends on the other factors explored.

General Discussion

In the present study, the correspondence between *p* values and the posterior probability that H_0 is true was investigated, under varying population effect sizes, sample sizes, α levels, and prior probabilities. The focal factor was the number of DVs, to assess the influence of multiple testing. In addition to the posterior probabilities signifying support for H_0, Bayes factors were also calculated. The goal was to not only assess how multiple testing and other factors influence the strength of evidence offered for H_0 and H_a, but also to make the degree of evidence explicit on a larger scale than previous research.

[6] Note that this is only one interpretation scheme for the Bayes factor, and relying too heavily on these guidelines can also be problematic (see the Discussion section).

Table 4
Comparing Null Versus Nil Results and Moderate and Large Sample Size

δ	.1	.3	.5	.7	.9
		$p < .05$			
n = 60 null	.41	.21	.18	.17	.17
n = 60 nil	.39	.17	.12	.12	.12
n = 150 null	.30	.13	.12	.12	.12
		$p < .01$			
n = 60 null	.33	.07	.03	.03	.03
n = 60 nil	.32	.07	.03	.03	.03
n = 150 null	.21	.04	.03	.03	.03
		$p < .005$			
n = 60 null	.29	.05	.02	.01	.01
n = 60 nil	.29	.05	.02	.01	.01
n = 150 null	.17	.02	.01	.01	.01
		$.045 < p < .05$			
n = 60 null	.43	.39	.73	.99	1.0
n = 60 nil	.44	.41	.72	.99	1.0
n = 150 null	.39	.69	1.0	1.0	1.0
		$.005 < p < .01$			
n = 60 null	.36	.13	.16	.57	.98
n = 60 nil	.35	.13	.15	.57	.99
n = 150 null	.25	.14	.86	1.0	1.0

Note. n = per-group sample size.

The Magnitude of the Problem and the Influence of Multiple Testing

This is certainly not the first article to reveal that the p value may be wearing a clever disguise, so good that researchers may forget what the p value means, despite knowledge to the contrary. However, extensive evidence was provided as to the degree to which statistically significant p values may be off from popular perceptions. Moreover, the results provide a new contribution of how much worse the problem gets with multiple testing. Before describing the influence of various manipulated factors, it is important to emphasize explicitly the size of the discrepancy between commonly reported p values and the probability that H_0 is true.

Recall that it is still common to associate a p value of just under .05 with only a 5% probability that H_0 is true, which, of course, is interpreted as strong evidence against H_0. But, how should a p value slightly less than .05 be interpreted? How slim is the evidence that H_0 is true? In the conditions investigated, the absolute minimum probability that H_0 is true was .20, assuming that H_0 and H_a were equiprobable a priori. This means that even under what some might call ideal or optimistic circumstances (no researcher degrees of freedom, a single DV), one of every five p values just below .05 may be linked to a null effect, rather than the supposed one in 20. Importantly, multiple testing quickly and substantially amplifies the discrepancy between the p value and the probability H_0 is true. When three and five DVs are tested, the minimum probability that H_0 is true increases to .37 and .43, respectively. Moreover, recall that these reflect the cases where the posterior probability was closest to the p value. Notably, the only cases in which this minimum was around .05 were when the prior probability strongly favored H_a. The respective maximum posterior probabilities favoring H_0 were close to one, indicating that statistically significant p values can correspond to extreme probabilities in favor of H_0, the opposite of what researchers expect from a significant finding. Moreover, this is not a finding only encountered in extreme situations. For three DVs, half of the conditions studied result in a posterior probability favoring H_0, and the proportion rises to 60% for five DVs.

Overall these results make the evidence p values provide much more tangible. A p value of just under .05 may be adequate for publication by current implicit standards but does not typically provide the standard of evidence against H_0 that is commonly perceived. Moreover, the importance of these results is not dependent on having fallen victim to a misinterpretation of the p value. It is important, even for psychologists who correctly interpret p values, to know not only that p values and posterior probabilities are not the same construct, but also how divergent the two are in conditions plausible in typical research. The discrepancy can sometimes be by orders of magnitude. Furthermore, this is the first extensive evaluation of the influence of multiple testing on this discrepancy: p values that yield acceptable levels of evidence for H_0 and H_a for a single DV become much less convincing under multiple testing.

Although not common practice in multiple regression (Cribbie, 2017), a potential solution specific to multiple testing is to use a Bonferroni correction for each dependent variable or test answering a similar scientific question. This strategy was implicitly

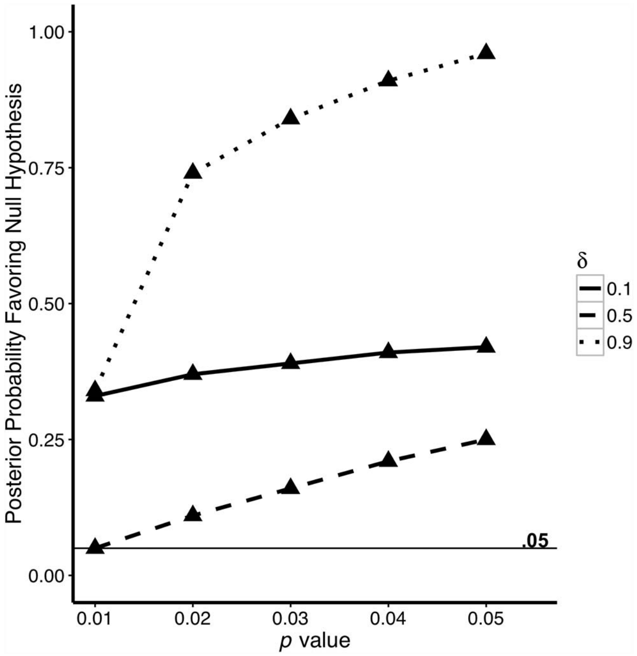

Figure 3. Plot showing the linear relationship between p values and the corresponding posterior probabilities favoring H_0, holding other factors constant. For this plot, δ = 0.1 (solid line), 0.5 (dashed line), and 0.9 (dotted line), n = 60 per group, and there is a single dependent variable. The p values plotted indicate p values falling in the interval of length .005 just under the specified p value. The horizontal line indicates a posterior probability of .05.

assessed in the present study, given that a Bonferroni correction with five DVs is equivalent to using .01 as α. Thus, Table 4 can be consulted, focusing on *p* values under .01 or just under .01. It is evident that this solution represents an improvement compared with no adjustment when multiple tests are conducted, although Bonferroni adjusted *p* values just under .01 still are not particularly convincing in several conditions and would require larger sample sizes to attain the same level of statistical power.

Influence of Prior Probability, Sample Size, and Effect Size

Consider a surprising result, such as the study described in the introduction on making rapid judgments (Willis & Todorov, 2006). Assuming that the common-sense theory is that it takes time to make a first impression, a correlation reaching statistical significance may seem especially convincing. After all, it would seem that the effect would have to be robust for the *p* value to "beat the odds" and end up positive and statistically significant. However, the present results emphasize that especially in situations of surprising results, where the prior probability in favor of H_0 is high to begin with, the burden of proof should not rest upon a single study. The importance of replication is elaborated upon in the Implications for Replication section.

When considering the influence of sample size and effect size, it is important to be aware of whether the conditioning is on a *p* value in a specific range or any statistically significant *p* value. When considering statistically significant *p* values broadly, increasing sample size and effect size serve only to lower the posterior probability that H_0 is true, which is in line with the notion of a consistent test (Rouder, Speckman, Sun, Morey, & Iverson, 2009). Thus, improvements to statistical power not only increase a researcher's chances of detecting a statistically significant effect, but they strengthen the evidence against H_0 when that significant effect is detected, because the resulting *p* value from a more powerful study will be smaller, and smaller *p* values are linked with stronger evidence against H_0. However, when considering *p* values in the interval just under .05 (or .01), the posterior probability no longer uniformly decreases with improvements to statistical power. This is particularly clear when comparing the moderate with the large sample size of $n = 150$. In other words, *p* values resulting from large sample sizes are not immune from the issues described. Note that this should not be taken to imply that increasing statistical power is not advantageous. This paradox occurs because, as aforementioned, the *p* value associated with a more powerful study should often be smaller than barely under .05. Thus, when the *p* value does only barely cross the threshold, when it should have been smaller, this now yields much weaker evidence for the effect of interest.

Limitations and Possible Solutions

Despite considering the impact of multiple testing and a variety of additional factors, in many ways, these results are still not a worst-case scenario. No additional researcher degrees of freedom were considered, which would further decrease the strength of evidence the *p* value provides for H_a. Furthermore, the simulation proceeded under the assumption that the effect sizes of all DVs were constant. In cases of multiple testing, it may be the case that the effect differs depending on how the DV is conceptualized. However, the present results provide a window into the actual evidence provided by statistically significant *p* values and the role of a common form of researcher degrees of freedom on this evidence.

Every several years, a proposed solution is presented to combat the crimes of the *p* value, such as *p*-rep (Killeen, 2005), Bayesian procedures (e.g., Ruberg et al., 2019), confidence intervals and effect sizes (e.g., the "new statistics"; Cumming, 2014), acceptance prior to data analysis (e.g., Locascio, 2019), and banning the *p* value (e.g., Trafimow & Marks, 2015), among others (see Wasserstein et al., 2019). Despite these proposals, the field has held onto NHST, and there have been compelling defenses of NHST and *p* values (e.g., Abelson, 1997; "the *p*-value is innocent"; Kuffner & Walker, 2019, p. 1; "one cheer for null hypothesis significance testing"; Wainer, 1999, p. 212). Of course, these solutions, and the solutions elaborated upon below, are not a panacea. For example, the Bayesian framework is not immune from poor methodological practices and researcher degrees of freedom (Savalei & Dunn, 2015; Simmons et al., 2011; Wilcox & Serang, 2017). Savalei and Dunn (2015) wrote "one can 'b-hack' just as one can 'p-hack'" (p. 2). Importantly, any proposal should be evaluated from both a broad and nuanced perspective to determine its potential impact and consequences (Campbell & Gustafson, 2019). The results of this investigation should not be seen as prioritizing one solution in particular, but represent a piece of evidence to aid psychologists in more effectively interpreting *p* values in research and deciding whether and how the *p* value should be involved in theory testing.

The present study implies some support for the recent proposal to adopt .005 as the new standard for statistical significance (Benjamin & Berger, 2019; Benjamin et al., 2018), given that even amid multiple testing, the posterior probabilities were much lower than in other conditions. However, although lowering α improves the rate of false discoveries within a research literature, the statistical cost is in larger sample sizes needed to achieve the same statistical power, which is a difficult payment to make in some research areas (see Maxwell, 2004). Moreover, Crane (2018) described how lower statistical power could lead to even higher levels of researcher degrees of freedom. Researcher degrees of freedom in the form of multiple testing limit the evidentiary value of a *p* value just under .05, potentially creating a need for a more stringent α, a solution which, somewhat ironically, may prompt a vicious cycle of more researcher degrees of freedom. Researchers have noted additional concerns regarding proposals of this nature, such as the one-size-fits-all approach that caters to certain research areas but alienates others, the failure to consider false negatives, and the potentially detrimental impact on early career researchers (Finkel, Eastwick, & Reis, 2015; Rodgers & Shrout, 2018). Finally, enforcing a more stringent α would still create a dividing line demarcating statistical significance.

Related to the arbitrary statistical significance distinction, some have suggested that *p* values should be reported as equalities rather than inequalities and that the term "statistically significant" no longer be used (e.g., Amrhein, Trafimow, & Greenland, 2019; Greenland, 2019; Ioannidis, 2019). The present results support this advice. It is important to realize that the *p* value as used today emerged from the pairing of two complementary ideas (Lehmann, 1993). Fisher did not interpret the *p* value against a benchmark level of significance nor require H_a, while Neyman required two competing hypotheses and calculating Type I and Type II errors, rather than the p value (Berger, 2003). Calling p values statistically significant when they fall below α indeed is "an incoherent mishmash" of two distinct frameworks (Gigerenzer, 1993, p. 314).

One compelling solution is to see p values as one of many indicators of an effect, rather than the sole determinant (McShane, Gal, Gelman, Robert, & Tackett, 2019). Researchers could provide a suite of evidence for the effect, in the form of, for example, the p value (how surprising the results are if H_0 was true), effect size (tangible information regarding the magnitude of the effect), confidence interval (precision of the effect), and Bayes factor (strength of evidence in favor of the H_0 and/or H_a). Bayes factors are relatively easy to compute for a variety of designs (e.g., Rouder, Morey, Speckman, & Province, 2012). However, interpretation of Bayes factors often relies on cutoffs (as was done here), not unlike interpreting p values as statistically significant. Amrhein, Greenland, and McShane (2019) referred to cutoffs such as these as "dichotomania" (p. 306) and encouraged greater tolerance for uncertainty in evaluating continuous measures. Another potential inclusion is posterior odds or probabilities favoring H_0 or H_a (e.g., Benjamin & Berger, 2019). The present results suggest that this proposal could be expanded to include a sensitivity analysis of posterior probabilities under different priors, or evidence to support the selected prior, though this may become more difficult for more complicated analyses. Importantly, many factors should ideally influence interpretation of the p value, as the simulation results have shown, such as the prior probability and whether multiple testing was conducted. In other words, the p value should be *context dependent* (Betensky, 2019).

Implications for Replication

These results have implications for the dangers of overreliance on a single study, the replication crisis, and the importance of incentivizing cumulative science. The present results imply that a single p value from a single study, no matter how large the sample size, may not offer enough evidence in favor of H_a to be anywhere close to definitive, and this is particularly true for surprising effects and when multiple testing is conducted. Multiple testing is often done within the context of exploratory studies, which are a valuable part of the scientific process, but should be clearly marketed as such (Tukey, 1969), and should not be expected to single-handedly define an effect. Beyond what was shown here, Kenny and Judd (2019) noted another reason to be cautious of a single study, even one with a large sample size. Because of overlooked systematic heterogeneity in effect sizes within a research domain that cannot be explained by sampling error and may be due to hidden moderators, researchers "are better served by a number of studies that permit one to examine the existing variability of effect sizes in a domain" (p. 9). Thus, rather than limiting the p value debate to how to best convey results in a single study, it may be especially advantageous to work with multiple studies.

In terms of the replication crisis, the present results provide a reason that the maligned low replicability may not be so surprising, in addition to concerns already raised (see Shrout & Rodgers, 2018, for a review), including statistical power (Anderson & Maxwell, 2017) and effect size heterogeneity (Kenny & Judd, 2019). Importantly, $1 - p$ is not equivalent to the probability that an effect will be replicated (Gigerenzer, 2018), meaning that a statistically significant result will not necessarily replicate ($p = .05$ does not imply a 95% chance of replication). However, this does not mean that p values are unrelated to the probability of successful replication (Greenwald, Gonzalez, Harris, & Guthrie, 1996). In fact, p values are valuable in providing information about replication: Studies with barely significant p values have lower rates of replication compared with studies with smaller p values (Open Science Collaboration, 2015 $r = -.327$).[7] There are many reasons why an effect may fail to replicate, but if H_0 is true, the probability of replication will assuredly be low, equal to α if other factors are held constant. It might be expected that if an effect does not exist, it would be rare for the study to appear in the literature. However, the present results show that even when p values fall under the conventional threshold for statistical significance, it may be quite likely that H_0 is true. The probability that H_0 was true approached 1.00 in several conditions, particularly for p values just under .05 and when H_0 was likely a priori to be true or multiple testing was conducted. The result is that many effects may not replicate because they do not meet the minimum expectation for replication, that the effect is real. In this way, relying on a cutoff of $p < .05$ may contribute to the replication crisis.

On a more positive note, the results support the current movement toward cumulative and transparent science, including multisite replication studies (e.g., Many Labs, Registered Replication Reports), preregistration, and clear reporting of researcher degrees of freedom. Given the volume of potential solutions offered in relation to problems with p values and NHST, it is clear that determining the role (or fate) of the p value is not a simple task. But, perhaps, rather than elevating the p value to a lofty position or denigrating the p value for failing at a task for which it was never intended to excel at, it is time for the p value to take on a more circumscribed role as a single chapter in a much longer scientific story about an effect. Certainly, this choice will have important implications for how psychological research is interpreted, replicated, and trusted.

[7] Note that this assumes replication as defined by a statistically significant replication effect, though there are many other ways to define replication (e.g., Anderson & Maxwell, 2016).

References

Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8,* 12–15. http://dx.doi.org/10.1111/j.1467-9280.1997.tb00536.x

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature, 567,* 305–307. http://dx.doi.org/10.1038/d41586-019-00857-9

Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician, 73,* 262–270. http://dx.doi.org/10.1080/00031305.2018.1543137

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science, 28,* 1547–1562. http://dx.doi.org/10.1177/0956797617723724

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods, 21,* 1–12. http://dx.doi.org/10.1037/met0000051

Anderson, S. F., & Maxwell, S. E. (2017). Addressing the "replication crisis": Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research, 52,* 305–324. http://dx.doi.org/10.1080/00273171.2017.1289361

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgment. In H. Guetzkow (Ed.), *Groups, leadership and men* (pp. 222–236). Pittsburgh, PA: Carnegie Press.

Badenes-Ribera, L., Frias-Navarro, D., Iotti, B., Bonilla-Campos, A., & Longobardi, C. (2016). Misconceptions of the p-value among Chilean and Italian Academic Psychologists. *Frontiers in Psychology*. Advance online publication. http://dx.doi.org/10.3389/fpsyg.2016.01247

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423–437. http://dx.doi.org/10.1037/h0020412

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. J. (2016). Researchers' intuitions about power in psychological research. *Psychological Science, 27,* 1069–1077. http://dx.doi.org/10.1177/0956797616647519

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. (2016). Editorial: Evidence on questionable research practices: The good, the bad, and the ugly. *Journal of Business and Psychology, 31,* 323–338. http://dx.doi.org/10.1007/s10869-016-9456-7

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. http://dx.doi.org/10.1037/a0021524

Benjamin, D. J., & Berger, J. O. (2019). Three recommendations for improving the use of p-values. *The American Statistician, 73,* 186–191. http://dx.doi.org/10.1080/00031305.2018.1543135

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., . . . Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2,* 6–10. http://dx.doi.org/10.1038/s41562-017-0189-z

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing. *Statistical Science, 18,* 1–32.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association, 82,* 112–122. http://dx.doi.org/10.2307/2289131

Betensky, R. A. (2019). The p-value requires context, not a threshold. *The American Statistician, 73,* 115–117. http://dx.doi.org/10.1080/00031305.2018.1529624

Campbell, H., & Gustafson, P. (2019). The world of research has gone berserk: Modeling the consequences of requiring "greater statistical stringency" for scientific publication. *The American Statistician, 73,* 358–373. http://dx.doi.org/10.1080/00031305.2018.1555101

Cassidy, S. A., Dimova, R., Giguère, B., Spence, J. R., & Stanley, D. J. (2019). Failing grade: 89% of introduction-to-psychology textbooks that define or explain statistical significance do so incorrectly. *Advances in Methods and Practices in Psychological Science*. Advance online publication. http://dx.doi.org/10.1177/2515245919858072

Cohen, J. (1988). *Statistical power analysis for the behavioral science* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003. http://dx.doi.org/10.1037/0003-066X.49.12.997

Counsell, A., & Harlow, L. L. (2017). Reporting practices and use of quantitative methods in Canadian journal articles in psychology. *Canadian Psychology, 58,* 140–147. http://dx.doi.org/10.1037/cap0000074

Crane, H. (2018). The impact of p-hacking on "redefine statistical significance." *Basic and Applied Social Psychology, 40,* 219–235. http://dx.doi.org/10.1080/01973533.2018.1474111

Cribbie, R. A. (2017). Multiplicity control, school uniforms, and other perplexing debates. *Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement, 49,* 159–165. http://dx.doi.org/10.1037/cbs0000075

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25,* 7–29. http://dx.doi.org/10.1177/0956797613504966

Efran, M. G. (1974). The effect of physical appearance on the judgment of guilt, interpersonal attraction, and severity of recommended punishment in a simulated jury task. *Journal of Research in Personality, 8,* 45–54. http://dx.doi.org/10.1016/0092-6566(74)90044-0

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology, 58,* 203–210. http://dx.doi.org/10.1037/h0041593

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. *Journal of Personality and Social Psychology, 108,* 275–297. http://dx.doi.org/10.1037/pspi0000007

Freedman, B. (2017). Equipose and the ethics of clinical research. *New England Journal of Medicine, 317,* 141–145. http://dx.doi.org/10.4324/9781315244426-17

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate ψ. *Journal of Personality and Social Psychology, 103,* 933–948. http://dx.doi.org/10.1037/a0029709

Gelman, A., & Loken, E. (2014). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *American Scientist, 102,* 460–465. http://dx.doi.org/10.1511/2014.111.460

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Hillsdale, NJ: Erlbaum, Inc.

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science, 1,* 198–218. http://dx.doi.org/10.1177/2515245918771329

Greenland, S. (2019). Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with s-values. *The American Statistician, 73,* 106–114. http://dx.doi.org/10.1080/00031305.2018.1529625

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1–20. http://dx.doi.org/10.1037/h0076157

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and p values: What should be reported and what should be replicated? *Psychophysiology, 33,* 175–183. http://dx.doi.org/10.1111/j.1469-8986.1996.tb02121.x

Haller, H., & Kraus, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online, 7,* 1–20.

Humphreys, M., de la Sierra, R. S., & van der Windt, P. (2013). Fishing, commitment, and communication: A Proposal for Comprehensive Nonbinding Research registration. *Political Analysis, 21,* 1–20. http://dx.doi.org/10.1093/pan/mps021

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2,* e124. http://dx.doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2019). What have we (not) learnt from millions of scientific papers with p values? *The American Statistician, 73,* 20–25. http://dx.doi.org/10.1080/00031305.2018.1447512

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23,* 524–532. http://dx.doi.org/10.1177/0956797611430953

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795. http://dx.doi.org/10.1080/01621459.1995.10476572

Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods, 24,* 578–589. http://dx.doi.org/10.1037/met0000209

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science, 16,* 345–353. http://dx.doi.org/10.1111/j.0956-7976.2005.01538.x

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/14136-000

Krawczyk, M. (2015). The search for significance: A few peculiarities in the distribution of p values in experimental psychology literature. *PLoS ONE, 10,* e0127872. http://dx.doi.org/10.1371/journal.pone.0127872

Kuffner, T. A., & Walker, S. G. (2019). Why are p-values controversial? *The American Statistician, 73,* 1–3. http://dx.doi.org/10.1080/00031305.2016.1277161

Laber, E. B., & Shedden, K. (2017). Statistical significance and the dichotomization of evidence: The relevance of the ASA statement on statistical significance and p-values for statisticians. *Journal of the American Statistical Association, 112,* 902–904. http://dx.doi.org/10.1080/01621459.2017.1311265

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association, 88,* 1242–1249. http://dx.doi.org/10.1080/01621459.1993.10476404

Locascio, J. J. (2019). The impact of results blind science publishing on statistical consultation and collaboration. *The American Statistician, 73,* 346–351. http://dx.doi.org/10.1080/00031305.2018.1505658

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills, 112,* 331–348. http://dx.doi.org/10.2466/03.11.PMS.112.2.331-348

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below. 05. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 65,* 2271–2279. http://dx.doi.org/10.1080/17470218.2012.711335

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9,* 147–163. http://dx.doi.org/10.1037/1082-989X.9.2.147

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experiments and analyzing data: A model comparison perspective* (3rd ed.). New York, NY: Routledge.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon Statistical Significance. *The American Statistician, 73,* 235–245. http://dx.doi.org/10.1080/00031305.2018.1527253

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103–115. http://dx.doi.org/10.1086/288135

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: A comment on Cumming (2014). *Psychological Science, 25,* 1289–1290. http://dx.doi.org/10.1177/0956797614525969

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349,* aac4716. http://dx.doi.org/10.1126/science.aac4716

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112,* 160–164. http://dx.doi.org/10.1037/0033-2909.112.1.160

Rodgers, J. L., & Shrout, P. E. (2018). Psychology's replication crisis as scientific opportunity: A précis for policymakers. *Policy Insights from the Behavioral and Brain Sciences, 5,* 134–141. http://dx.doi.org/10.1177/2372732217749254

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225–237. http://dx.doi.org/10.3758/PBR.16.2.225

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56,* 356–374. http://dx.doi.org/10.1016/j.jmp.2012.08.001

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57,* 416–428. http://dx.doi.org/10.1037/h0042040

Ruberg, S. J., Harrell, F. E., Gamalo-Siebers, M., LaVange, L., Jack Lee, J., Price, K., & Peck, C. (2019). Inference and decision making for 21st-century drug development and approval. *The American Statistician, 73,* 319–327. http://dx.doi.org/10.1080/00031305.2019.1566091

Savalei, V., & Dunn, E. (2015). Is the call to abandon p-values the red herring of the replicability crisis? *Frontiers in Psychology*. Advance online publication. http://dx.doi.org/10.3389/fpsyg.2015.00245

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of ρ values for testing precise null hypotheses. *The American Statistician, 55,* 62–71. http://dx.doi.org/10.1198/000313001300339950

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology, 69,* 487–510. http://dx.doi.org/10.1146/annurev-psych-122216-011845

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. http://dx.doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). *p*-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science, 9,* 666–681. http://dx.doi.org/10.1177/1745691614553988

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). *p*-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143,* 534–547. http://dx.doi.org/10.1037/a0033242

Stern, H. S. (2016). A test by any other name: *p* values, Bayes factors, and statistical inference. *Multivariate Behavioral Research, 51,* 23–29. http://dx.doi.org/10.1080/00273171.2015.1099032

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37,* 1–2. http://dx.doi.org/10.1080/01973533.2015.1012991

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24,* 83–91. http://dx.doi.org/10.1037/h0027108

Vickerstaff, V., Ambler, G., King, M., Nazareth, I., & Omar, R. Z. (2015). Are multiple primary outcomes analysed appropriately in randomised controlled trials? A review. *Contemporary Clinical Trials, 45,* 8–12. http://dx.doi.org/10.1016/j.cct.2015.07.016

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14,* 779–804. http://dx.doi.org/10.3758/BF03194105

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods, 4,* 212–213. http://dx.doi.org/10.1037/1082-989X.4.2.212

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician, 70,* 129–133. http://dx.doi.org/10.1080/00031305.2016.1154108

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p < 0.05." *The American Statistician, 73,* 1–19. http://dx.doi.org/10.1080/00031305.2019.1583913

Wegner, D. M., Schneider, D. J., Carter, S. R., III, & White, T. L. (1987). Paradoxical effects of thought suppression. *Journal of Personality and Social Psychology, 53,* 5–13. http://dx.doi.org/10.1037/0022-3514.53.1.5

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science, 6,* 291–298. http://dx.doi.org/10.1177/1745691611406923

Wilcox, R. R., & Serang, S. (2017). Hypothesis testing, *p* values, confidence intervals, measures of effect size, and Bayesian methods in light of modern robust techniques. *Educational and Psychological Measurement, 77,* 673–689. http://dx.doi.org/10.1177/0013164416667983

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17,* 592–598. http://dx.doi.org/10.1111/j.1467-9280.2006.01750.x

Zellner, A. (1984). *Basic issues in econometrics* (pp. 151–152). Chicago, IL: University of Chicago Press.

Received July 23, 2019
Revision received October 28, 2019
Accepted October 30, 2019 ■