UNIVERSITY of CALIFORNIA PRESS | Collabra

Krueger, J. I. and Heck, P. R. (2018). Testing Significance Testing. *Collabra: Psychology,* 4(1): 11. pp. 1–13, DOI: https://doi.org/10.1525/collabra.108

ORIGINAL RESEARCH REPORT

Testing Significance Testing

Joachim I. Krueger* and Patrick R. Heck†

The practice of Significance Testing (ST) remains widespread in psychological science despite continual criticism of its flaws and abuses. Using simulation experiments, we address four concerns about ST and for two of these we compare ST's performance with prominent alternatives. We find the following: First, the *p* values delivered by ST predict the posterior probability of the tested hypothesis well under many research conditions. Second, low *p* values support inductive inferences because they are most likely to occur when the tested hypothesis is false. Third, *p* values track likelihood ratios without raising the uncertainties of relative inference. Fourth, *p* values predict the replicability of research findings better than confidence intervals do. Given these results, we conclude that *p* values may be used judiciously as a heuristic tool for inductive inference. Yet, *p* values cannot bear the full burden of inference. We encourage researchers to be flexible in their selection and use of statistical methods.

Keywords: statistical significance testing; null hypotheses; Bayes' Theorem; NHST; *p* values

The *Zeitgeist* in psychological science is rife with self-doubt and criticism (Lilienfeld & Waldman, 2017). Theories are said to be shallow and narrow (Fiedler, 2017; Gigerenzer, 1998; Gigerenzer & Marewski, 2015), research practices to lack professional and ethical rigor (Simmons, Nelson, & Simonsohn, 2011), and statistical methods to be low in power and validity (Button et al., 2013). Significance Testing (ST), and particularly its null hypothesis variant (NHST), is a prominent target of criticism, in part because of the many documented misconceptions regarding its limitations and its proper use. Although ST continues to be widely used in research practice (Krueger, 2001; Nickerson, 2000), its popularity says little about its validity. A method may be popular simply because of tradition and habit, or because of false beliefs regarding its validity. In recent years, the use of alternative methods, which are sometimes used in combination with ST, has increased. From the researcher's point of view, an eclectic approach is reasonable and pragmatic, given the unsettled nature of the critical debate. Indeed, many prominent commentators have endorsed eclecticism (Abelson, 1995; Cohen, 1990; Dawes, 1991; Senn, 2001, 2017; Wilkinson & the APA Task Force on Statistical Inference, 1999).

Recently, the *American Statistical Association* (ASA) weighed in with a "statement on statistical significance and *P*-values." Noting that "the *p*-value can be a useful statistical measure" (Wasserstein & Lazar, 2016, p. 131), the authors of the statement cautioned that *p* is of limited value on its own. They also express a strong concern about misuses and misinterpretations of *p* values, as did the authors of similar statements in the past. For the purposes of the present article, we consider misconceptions about the properties of ST as the purview of education and ethics (Lilienfeld, Sauvigné, Lynn, Cautin, Latzman, & Waldman, 2015). Here, we are primarily concerned with the method's conceptual and technical properties (Greenland, 2017; Greenland et al., 2016; Perezgonzalez, 2015). A focus on these properties is timely because some authors have concluded that the inferential value of ST is so slim that other methods should take its place (Eich, 2014; McShane, Gal, Gelman, Robert, Tackett, 2017; Trafimow & Marks, 2015). But which of these alternatives shall be favored? Confidence intervals, parameter estimation, descriptive analysis, and various Bayesian methods are available, and each method has its own strengths and weaknesses. Here, we hope to contribute to this conversation by assessing the performance of the *p* value and by comparing it, where possible, with specific alternatives.

The controversy over best statistical practices emerged, in part, from historical accident. The long-standing prominence of ST made it a salient object of critical discussion. Individual alternative methods are sometimes seen as gaining in credibility inasmuch as a particular shortcoming of ST is demonstrated. Direct comparisons between ST and alternatives are rare, as are comparisons between or among those alternatives.

* Department of Cognitive, Linguistic and Psychological Sciences, Brown University, Providence, RI, US

† Geisinger Health System, Danville, PA, US

Corresponding authors: Joachim I. Krueger (Joachim@brown.edu), Patrick R. Heck (Pheck1000@gmail.com)

Our approach is to pose four questions regarding inductive inference, and then to assess ST's performance – where possible in direct comparison with an alternative. We first address each question at the conceptual level and then seek quantitative answers in simulation experiments. The four questions are: [1] How well does the *p* value predict the posterior probability of the tested hypothesis? [2] Is there a relationship between the variability of the *p* value over studies and its inductive value? [3] How well does the *p* value perform compared with the likelihood ratio or Bayes factor? [4] How well does the *p* value perform compared with confidence intervals in predicting the replicability of empirical results? In answering these questions we do not attempt to insulate the *p* value from critique, but rather attempt to evaluate its performance under balanced assumptions and conditions.

1. Does the *p* value predict the probability of a hypothesis given the evidence?

The *p* value refers to the probability of the data at least as extreme as the observed data given the statistical (often the null) hypothesis, p(D|H), and assuming that underlying assumptions are met (Greenland et al., 2016; Wasserstein & Lazar, 2016). In ST, the test statistic (e.g., *z*, *t*, or *F*) represents the data as it is computed from the central tendency of the observed data and the standard error. We use the terms *p* value and p(D|H) interchangeably. As a probability that refers to the size of an area under a density curve, the *p* value is conceptually distinct from the likelihood of the data, which refers the value of the density function at a particular point. In our simulation experiments, we find that the log-transforms of *p* values are nearly perfectly correlated with their associated likelihoods. Consider a continuous distribution under the null hypothesis of $\mu = 0$. As sample observations increase in magnitude (for example, from a range of .01 to 2.0 standard units) when moving from the peak of this distribution toward the positive (right) tail, *p* values and likelihoods both decrease monotonically. In this article, we only report the findings obtained with likelihoods.

A key concern about the *p* value is that it does not speak to the strength of the evidence against the tested hypothesis, that is, that it does not predict the posterior probability of the tested hypothesis (Cohen, 1994; Gelman, 2013; Lykken, 1968). The ASA warns that that "*p*-values do not measure the probability that the studied hypothesis is true" (Wasserstein & Lazar, 2016, p. 131), although "researchers often wish to turn a *p*-value into a statement about the truth of a null hypothesis" (p. 131). In other words, finding that the data are unlikely under the hypothesis is not the same as finding that the hypothesis is unlikely under the data. The question of whether there is any relationship, and how strong it might be, is the crux of inductive inference. All inductive inference is essentially "reverse inference," and reverse inference demands vigilance (Krueger, 2017).

We sought to quantify how much p(D|H) reveals about p(H|D). Bayes' Theorem, which expresses the mathematical relationship between the two inverse conditional probabilities, provides the first clues. The theorem

$$p(H \mid D) = \frac{p(H) \times p(D \mid H)}{p(H) \times p(D \mid H) + p(\sim H) \times p(D \mid \sim H)}$$

shows that as p(D|H) decreases, *ceteris paribus*, so does p(H|D). If the tested hypothesis, H, is a null hypothesis, a low *p* value suggests a comparatively high probability that the alternative hypothesis, ~H, is true. Yet, the association between p(D|H) and p(H|D) is perfect only if the prior probability of the hypothesis, p(H), is the same as the cumulative probability of the data, p(D), that is, the denominator of the ratio in the above formula. This identity may be rare in research practice; so how strongly is p(D|H) related to p(H|D) in practice?

We studied the results for a variety of settings in simulation experiments (Krueger & Heck, 2017). We began by sampling the elements of Bayes' Theorem, p(H), p(D|H), and p(D|~H) from uniform distributions that were independent of one another. These simple settings produced a correlation of $r = .38$ between p(D|H) and p(H|D) (see also Krueger, 2001; Trafimow & Rice, 2009). The size of this correlation may raise questions about the inductive power of the *p* value. Note, however, that this correlation emerges for a set of minimal, and as we shall see unrealistic, assumptions and thus represents a lower bound of possible results. Consider the relationship between p(D|H) and p(D|~H) over studies. Inasmuch as the null hypothesis H and the alternative hypothesis ~H are distinctive, one may expect a negative correlation between p(D|H) and p(D|~H) over studies. The limiting case is given by a daring ~H predicting a large effect, δ, and a set of experiments yielding estimated effects *d* that are greater than 0 but smaller than δ (García-Pérez, 2016). Here, the correlation between p(D|H) and p(D|~H) is perfectly negative.

We sampled values for p(H), p(D|H), and p(D|~H) and varied the size of the negative correlation between p(D|H) and p(D|~H), with the result of interest being the correlation between p(D|H) and p(H|D), that is, the correlation indicating the predictive power of *p* for the posterior probability of the null hypothesis. We found that as the correlation between p(D|H) and p(D|~H) becomes more negative, the correlation between p(D|H) and p(H|D) becomes more positive. For example, when setting the correlation between p(D|H) and p(D|~H) to $r = -.9$, the outcome correlation between p(D|H) p(H|D) is $r = .49$, which is moderately greater than the baseline correlation of .38 obtained under the assumption of independence. Nevertheless, when a research program provides bold hypotheses, that is, hypotheses that overestimate empirical effect sizes, the *p* value becomes an incrementally stronger predictor of the posterior probability of H (and thereby of ~H).

Turning to the effect of researchers' prior knowledge on the inductive power of *p*, we varied the correlation between p(D|H) and the prior probability of a hypothesis p(H). Here, positive correlations reflect the researchers' sense of the riskiness of the tested hypothesis. At one end of the spectrum, consider an experiment in parapsychology, where the prior probability of the null hypothesis (e.g., "Psychokinesis cannot occur") is high – at least

among skeptics. A low *p* value is improbable, that is, the (meta-)probability of a low *p* value is low. Thus, both p(~H) and p($p < .05$) are low.[1] At the other end of the spectrum, consider a social categorization experiment, for example, on ingroup-favoritism. Ingroup-favoritism is a robust empirical finding (Brewer, 2007), and thus the prior probability of the null hypothesis of no favoritism is low. Now, both p(~H) and p($p < .05$) are high. When multiple scenarios across this spectrum are considered, the positive correlation between p(H) and p(D|H) is evident.

When raising the correlation between p(H) and p(D|H) to .5 and to .9, we respectively observe correlations of .628 and .891 between p(D|H) and p(H|D). This result suggests that as a research program matures, the *p* value becomes more closely related to both the prior probability of the tested hypothesis and its updated posterior probability. Interestingly, ST yields diminishing returns within a line of study, as reflected in shrinking differences between p(H) and p(H|D). To review, the distribution of the prior probability of the likelihood of a hypothesis tends to be flat and uncoupled from the obtained *p* value in the early stages of a research program. At this stage, *p* values predict p(H|D) rather poorly. As theory and experience mature, however, the probabilities assigned to hypotheses begin to fall into a bimodal distribution; the researcher's experience allows more informed guesses as to which hypotheses are true and which are false. When a null hypothesis is tested that has already been rejected several times, its probability prior to the next study is low and so is the expected *p* value.

Consider research on the self-enhancement bias as another example for the use of ST in a mature research domain. After years of confirmatory findings, the researcher can predict that most respondents will regard themselves as above average when rating themselves and the average person on dimensions of personal importance (Krueger, Heck, & Asendorpf, 2017). The prior probability of the null hypothesis of no self-enhancement is low and the meta-probability of a low *p* value is high. When *p* values are closely linked to the priors, their surprise value is low; they do not afford much belief updating. In light of this consideration, a desire for a strong correlation between p(D|H) and p(H|D) must be balanced against the desire to maximize learning from the data, that is, the difference between p(H) and p(H|D). A certain hypothesis requires no additional data to increase this certainty. ST is most valuable when the researcher's theory and experience call for tests of novel and *somewhat* risky hypotheses. If the hypothesis is neither novel nor risky, little can be learned; if, in contrast, the hypothesis is too risky, the effort of testing it is likely wasted.

2. Is the variability of the *p* value related to its inductive value?

A second concern about *p* values is their variability (Cumming, 2008; Gelman & Stern, 2006; Halsey, Curran-Everett, Vowler, & Drummond, 2015). Cumming (2014, p. 13) observes "that *p* can take almost any value! The dance of the *p* values is astonishingly wide!" The implied reverse inference is that variable statistics are of limited value because statistics of limited value tend to be variable. How valid is this reverse inference? On the one hand, it is well known that the density distribution of the *p* value is uniform when the null hypothesis is true (Murdoch, Tsai, & Adcock, 2008). Any particular value for *p*, from 0 to 1, has the same chance of appearing *when there is no effect*. On the other hand, it is clear that *p* values are less variable when the alternative hypothesis is true, and their variability shrinks further as true effects and samples get larger (see the tool available at http://rpsychologist.com/d3/pdist/ to visualize this property, Magnusson, 2015). When the null hypothesis is false (i.e., p(H) = 0), the density function of *p* is right-skewed such that its smallest values are the most likely (Cumming, 2008; Murdoch et al., 2008). Now, the probability of finding a particular *p* value or smaller is greater than that particular *p* value. This second-order probability is the probability of a *p* value, or, *pp* (Simonsohn, Nelson, & Simmons, 2014). The critical regularity is the negative correlation between *p* and *pp if the null hypothesis is false*. A small *p* value has a high probability of being observed. If the null hypothesis is true, however, the probability of obtaining $p < .05$ is exactly .05. As the null hypothesis becomes less likely to be true, the probability of obtaining $p < .05$ becomes greater than .05.

Consider – like Jonathan Swift – two islands, one in which the Lilliputians are much shorter than the Blesfucians, and another in which there is no difference. Sampling heights from the no-effect island produces a uniform distribution of *p* values; by chance alone, *p* values of .05 or less will be drawn 5% of the time. Sampling from the population with a large effect positively skews the distribution of *p* values. When most Blesfucians are taller than most Lilliputians, detecting the difference is easy: randomly sampling from each group is likely to show a difference in average height and a correspondingly low *p* value. This case illustrates the logic of *pp*, where there is a high probability (the former *p* in the term) of obtaining a low *p* value (the latter *p*). Here, *failing* to detect a difference is unlikely and would be the surprising outcome.

We ask whether the relationship between *p* and *pp* can shed light on the *p* value's inductive value. We use the term 'value' in the sense of 'utility' or 'worth.' How 'valuable' is any given *p* value in making an inductive inference? The value of an outcome is a multiplicative function of its unconditional value (or 'worth' to the user) and its probability of being obtained (Bernoulli, 1954/1738). We add the psychological assumption that researchers using ST 'prefer' low *p* values to large ones (cf. Benjamin, Berger et al., 2017). Consider a set of studies, some of which are 'safe' and others are 'risky.' In safe studies, substantive effects are confidently predicted on the basis of theory and prior research; in risky studies, researchers rather throw darts in the dark, hoping to capture a novel phenomenon. The safe studies are more likely to yield low *p* values than the risky studies. In other words, when moving from risky to safe research, there is an increasingly negative correlation between *p* values and their probability of occurring (*pp*). Low *p* values become more probable (i.e., low *p* with high *pp*) as a research program becomes

safer (i.e., more predictable). This regularity reflects a general phenomenon found in nature and in the world of human artifacts and services; value increases with scarcity (Pleskac & Hertwig, 2014).[2] In the context of ST, finding a low *p* value in risky research (low p(~H)) is the most prized result, precisely because it is hard to get.

We now quantify the concept of inductive value of the statistical *p* value by taking the complement of the ratio of *p* over its probability of being observed, *pp*. This index of 1–*p*/*pp* is 0 if *p* = *pp*, that is, if null hypothesis is true. If there is no effect, then the *p* value has no inductive value. If, as skeptics of the paranormal, we tested for the presence of psychokinesis in multiple studies, for example, we would expect *a priori* to find $p \leq .05$ in 5% of these studies. In contrast, when *p* is *lower* than its second-order probability of occurring (i.e., if $p < pp$), there is added inductive value. For this to be true, the prior probability of the null hypothesis must be less than 1, that is, there must be point in doing the experiment. Suppose two experiments yield *p* = .01, but *pp* = .02 in one experiment (the distribution of *p* is positively skewed) and .03 in the other (*more* positive skew). The latter case suggests the presence of a larger effect. Therefore, the inductive value of *p* is greater in the second experiment (.667) than in the first (.5). The higher number suggests a stronger case against the null hypothesis because a low *p* value is obtained in a condition where it is *more* likely to be observed, that is, in a condition in which the null is false.

Recall that *p* values tend to decrease as true effects become larger, as null hypotheses become less probable, and as samples become larger. Our prediction is that *p* values decrease *faster* than their probability of occurring (i.e., *pp*). If so, a decrease in *p* increases its inductive value (or 'worth' to the researcher). Our prediction amounts to reverse inference from a low *p* value to an inductively valuable finding. For each setting in a series of simulations, we drew 10,000 samples each from a normal null distribution ($\mu = 50$, $\sigma = 10$) and an alternative distribution ($\mu = 50 + \delta$ and $\sigma = 10$), and conducted one-sample *t*-tests comparing the observed means with the theoretical one. We modeled three different research environments by varying the number of samples drawn from each distribution. In the *sure-thing* environment, all samples were drawn from the alternative distribution (i.e., p(H) = 0). In the *uncertain* environment, half of samples were drawn from the null distribution (p(H) = .5). In the *risky* environment, 80% of the samples were drawn from the null distribution (p(H) = .8). In each environment, we studied the effect sizes of $\delta = .2, .5, .8$ under three sampling conditions ($N = 20, 50, 100$). We obtained the *p* value for each test, and then estimated its corresponding *pp* by asking how many of the 10,000 samples resulted in that *p* value or below it. Finally, we estimated inductive value as 1–*p*/*pp*.

Table 1 shows the results for the *p* value thresholds of .05, .01, .005, and .001, which were set up so that a cumulative probability of *p* values equal or lower was possible. As predicted, inductive value increases with the probability of the alternative hypothesis (~H) being true, the size of that effect (δ), and the sample size (*N*). This pattern is the inverse of the pattern characteristic of the simple *p* value. We thus observe the predicted negative correlation between *p* and (1–*p*/*pp*). When correlating the median *p* values obtained from the 27 simulation experiments with their corresponding inductive-value indices, the result is $r = -.898$. This correlation is negative because the probability of the non-null effect, p(~H), the size of that effect, δ, and the size of the sample, *N*, affect *p* more strongly than they affect its meta-probability, *pp*. When p(H) = 0, the *p* value has great expected value regardless of the size of the effect or the sample. In contrast, when research is risky, p(H) = .8, the expected value of *p* is variable; it depends on the size of the effect and the sample. For these risky environments in particular, researchers must take care to adequately power experiments based on precise effect size estimates.

The index of inductive value is similar to other efforts to place *p* values in the context of their probability of occurring. Mayo (2016) noted that "looking at the *p*-value distribution under various discrepancies from *H*0: $\mu = \mu 0$ allows inferring those that are well or poorly indicated. If you very probably would have observed a more impressive (smaller) *p*-value than you did, if $\mu > \mu 1$ (where $\mu 1 = \mu 0 + \gamma$), then the data are good evidence that $\mu \leq \mu 1$" (see also Simonsohn et al., 2014). Our index integrates first- and second-order probability to reveal the *p* value's inductive value. Researchers may calculate their own inductive value index (1 – *p*/*pp*) with the caveat that, similar to the p-curve, they must have access to the unique underlying distribution of *p* values in their experimental environment (given p(H), *d*, and *n*). **Table 1** provides estimates of this index under several discrete settings of the research environment. Future work may explore the possibility of creating a simulation tool that takes in the necessary inputs to generate a robust distribution of *p* values and computes

Table 1: The "Inductive Value" of *p*.

	Sure-thing environment			Uncertain environment			Risky environment		
	δ = .2	.5	.8	δ = .2	.5	.8	δ = .2	.5	.8
N = 20	0.768	0.973	0.990	0.378	0.814	0.933	–0.019	0.122	0.336
N = 50	0.924	0.990	0.992	0.671	0.948	0.983	0.133	0.543	0.775
N = 100	0.971	0.992	0.993	0.831	0.980	0.991	0.356	0.811	0.924

Note. Table entries are the medians for (1 – (*p*/*pp*)) computed over *p* values of .05, .01, .005, and .001. 'Sure-thing,' 'Uncertain,' and 'Risky' specify that, respectively, 100%, 50%, or 20% of samples were drawn from the alternative distribution.

the inductive value index based on the *p* value obtained from an experiment. This metric, computed from only a single experiment, is a departure from p-curve analysis, which takes in a series of observed *p* values to measure their apparent fit with a theoretical distribution of *p* values (Simmons et al., 2015). The inductive value index is not meant to replace *p* as a metric of evidence and inference, or p-curve analysis as an inference tool capable of evaluating a series of results. Instead, this index and our demonstrations serve to corroborate the intuitive notion that a low *p* value from a single experiment sends a heuristic signal suggesting the presence of a non-null effect.[3]

3. Do likelihood ratios perform better than *p* values?

Two elements of Bayes' Theorem combine to form the likelihod ratio, LR. The numerator of the formula presented earlier contains p(D|H) and thus the *p* value. The denominator contains the term p(D|~H), which is the likelihood of data under the alternative hypothesis. Another way of writing the theorem is to note that the ratio of the posterior probabilities (i.e., the posterior odds that the [null] hypothesis is true) is equal to the prior odds times the LR, i.e.,

$$\frac{p(H \mid D)}{p(\sim H \mid D)} = \frac{p(H)}{p(\sim H)} \times \frac{p(D \mid H)}{p(D \mid \sim H)}$$

When the alternative hypotheis refers to a specific point, the LR may also be referred to as the Bayes factor, BF (e.g., Lee & Wagenmakers, 2005; Ortega & Navarrete, 2017). Many authors recommend the LR (or the posterior odds) as alternatives to ST and its *p* value (Goodman & Royall, 1988; Kruschke & Lidell, 2017; Lindley, 1975). Setting aside the complexity introduced by the selection of priors (van der Linden & Chryst, 2017), we ask how well the *p* value performs compared to the LR.

The estimation of a LR requires a specific alternative hypothesis, ~H, in addition to the null hypothesis, H. Having to make this selection explicit is thought to eliminate the illusion of scooping up a "free lunch" (Rouder et al., 2016). If the likelihood of the null hypothesis is the denominator, finding that LR > 1 favors the alternative. For there to be relative evidence against the null hypothesis, the data must not be only unlikely under the null hypothesis, they must be less likely under the null hypothesis than under the alternative. The emphasis on relative evidentiary value is a critical conceptual shift away from the routine of ST. The other shift is that the location of the alternative hypothesis may vary, thereby allowing multiple LR to be computed.

We now assume that a specific alternative hypothesis has been chosen, and that multiple experiments can be performed. From this perspective, we see the close correspondence between the *p* value and the LR. Indeed, when both indices are log transformed, the correlation between the two is nearly perfect. As the *p* value drops, so does the numerator of the LR and the LR itself.[4] This correlation is, of course, not sufficient to equate the two; LR are often evaluated in absolute terms as exceeding or failing to meet some threshold of evidentiary value (Wetzels, Matzke, Lee, Rouder, Iverson, & Wagenmakers, 2011). Over studies, however, the relationship between *p* and the LR is an important one. We explore when and how LR covary with *p.*

One way to show that the LR can capture variation that is ignored by the *p* value is to hold *p* constant (i.e., by looking only at the data from one sample) and to vary the alternative hypothesis. As the effect δ predicted by ~H varies, so does p(D|~H) and therefore the LR. Wagenmakers et al. (2017) described three hypothetical experiments, each producing $t(48) = 2.11$, $p = .04$, and considering a different ~H for each. The three predicted effect sizes are δ = .15, .60, and 2.0. The corresponding LR [here: p(D|~H)/p(D|H)] are 2.56, 8.61, and 1/13,867.

The point of this illustration is that the same *p* value can weaken or strengthen the null hypothesis depending on the location of the alternative hypothesis. A strong conclusion is that since Fisherian ST does not provide an alternative, the inductive value of *p* is indeterminate. While we agree with this conclusion, we hasten to note that the present illustration generalizes to a single experiment whose data are evaluated in light of different alternative hypotheses. Unless the selection of these hypotheses is made with the greatest theoretical rigor and restraint, a danger of hypothesis hacking lurks (Kerr, 1998).

We now return to the question of how the *p* value is related to the LR. So far, we have considered two scenarios, both of which are incomplete. In the first scenario, both hypotheses were fixed while the data varied. Here, the LR was perfectly redundant to the *p* value. In the second scenario, we followed Wagenmakers and colleagues to show that when the null hypothesis and the data are fixed while the alternative hypothesis varies, the relationship between *p* and the LR is undefined. A more comprehensive view involves variation in both *p* and the LR, where the latter is accomplished by variations of ~H or δ.

We therefore conducted simulation experiments, in which we varied both p(D|H) and p(D|~H). We set the null distribution to $\mu = 10$ and $\sigma = 5$, and chose a series of mean values for ~H (11, 12.5, 14, 15, 17.5, 20, 22.5, 30, and 40) to represent alternatives with a spread of effect sizes (δ = .2, .5, .8, 1, 1.5, 2, 2.5, and 4). Next, we chose the three *p* values of .05, .01, and .001, which we respectively obtained with sample data that had means of 19.8, 22.9, and 26.5. **Table 2** shows these settings along with the probabilities of the data under the alternative hypotheses, p(D|~H) as well as the LR. With these settings, we find that as the *p* value gets smaller, so does the average p(D|~H), $r = .38$. The crucial result is that the LR also decreases. It does so because the value of p(D|H) drops more sharply (from .05 to .001) than does the value of p(D|~H) (from .381 to .152 averaged over variation in δ). Across experiments, the correlation between the *p* value and the LR is $r = .46$. The size of this correlation supports our view that the *p* value is a useful heuristic for induction; it is neither useless nor can it do all the required work.

We have considered some low values for *p* to see what happens within the range of results conventionally considered significant, and we used only specific alternative

Table 2: Likelihood Ratio to *p* Value Comparisons.

Simulation Parameters	δ	$M_{\sim H}$	$Z_{\sim H}$	p(D|~H)	LR
m = 19.8	0.2	11	1.76	0.079	0.689
H = 10	0.5	12.5	1.46	0.145	0.425
z = 1.96	0.8	14	1.16	0.246	0.287
p = .05	1	15	0.96	0.337	0.232
	1.5	17.5	0.46	0.646	0.163
	2	20	0.04	0.968	0.147
	2.5	22.5	0.54	0.589	0.169
	4	30	2.04	0.042	1.174
	Average	*17.81*	*1.05*	*0.381*	*0.159*
m = 22.9	0.2	11	2.38	0.017	0.609
H = 10	0.5	12.5	2.08	0.038	0.312
z = 2.58	0.8	14	1.78	0.075	0.175
p = .01	1	15	1.58	0.114	0.125
	1.5	17.5	1.08	0.280	0.064
	2	20	0.58	0.562	0.042
	2.5	22.5	0.08	0.936	0.036
	4	30	1.42	0.156	0.098
	Average	*17.81*	*1.37*	*0.272*	*0.060*
m = 26.5	0.2	11	3.10	0.002	0.527
H = 10	0.5	12.5	2.80	0.005	0.218
z = 3.30	0.8	14	2.50	0.013	0.098
p = .001	1	15	2.30	0.022	0.061
	1.5	17.5	1.80	0.072	0.022
	2	20	1.30	0.194	0.010
	2.5	22.5	0.80	0.424	0.006
	4	30	0.70	0.484	0.006
	Average	*17.81*	*1.91*	*0.152*	*0.020*

Note. m = observed mean, H = Null hypothesis mean, *z* = test statistic observed, *p* = *p* value resulting from test, δ = effect size for alternative hypothesis, LR = Likelihood ratio.

hypotheses. When researchers do not wish to commit to specific alternative hypotheses, they can aggregate over a distribution of alternatives to the null, and compute 'diffuse' or 'default' tests. The results of this approach often resemble ST. Wetzels et al. (2011) computed diffuse-alternative LR (they preferred the label BF) for a large set of published results, most of which significant by the lights of ST. Plotting LR against *p*, the authors found a strong relationship (see Wetzels et al., Figure 2, where the LR and *p* are nearly perfectly correlated). Yet, they considered the LR to be superior, noting that the range of *p* from .05 to .01 merely qualifies as "anecdotal evidence" by Bayesian standards (Jeffreys, 1961). We submit that this discrepancy reflects different conventions in labeling segments of a scale rather than substantive differences in evidentiary power (Krueger & Heck, 2017).

To conclude this section, we observe that the relationship between the LR and the *p* value is a special case of the general relationship between a ratio and its own numerator. In a set of simulations, we randomly sampled values for variables X and Y drawn from a uniform distribution ranging from 0 to 1, and varied the correlation between X and Y. When using the raw values of X and X/Y, the correlation between the two hovers around zero because very small denominators can produce enormous ratios. These outliers skew the distributions to the point at which Pearson's *r* becomes meaningless. This observation may explain – and guard against – the claim that the LR is independent of the *p* value. With log transformation (or Spearman's Rho), however, high correlations emerge and they increase as r_{XY} becomes more negative. For example, $r_{X,X/Y}$ = .34 and .83 respectively for r_{XY} = .5 and –.5. Small

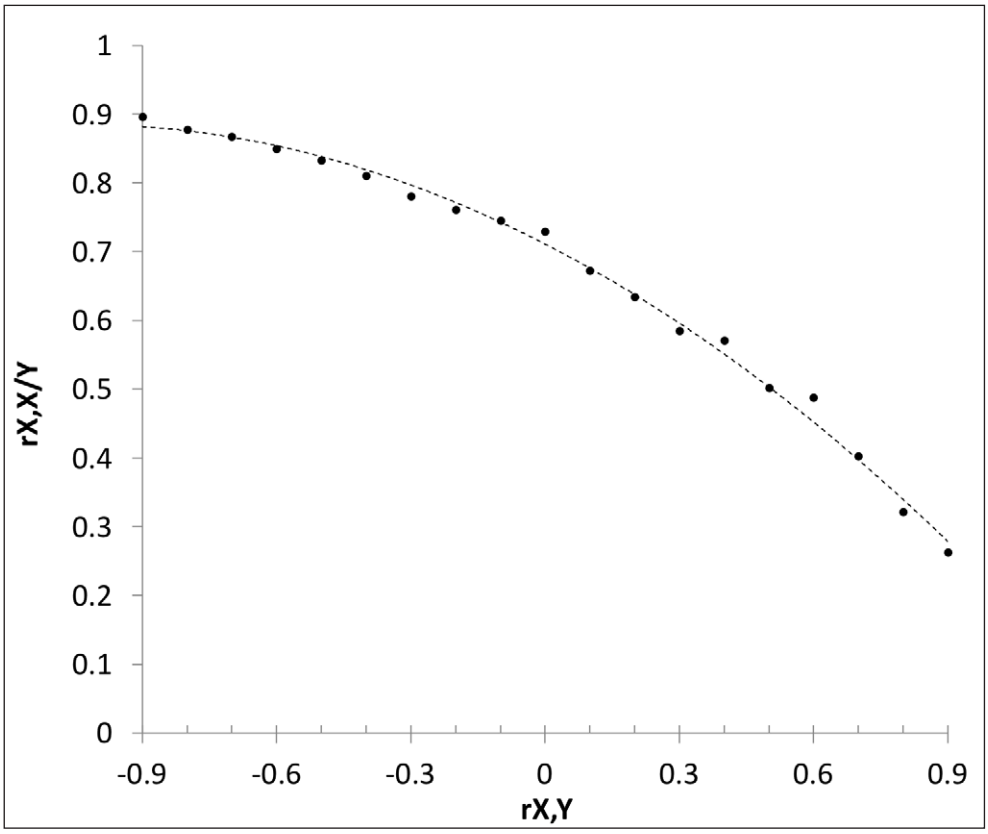

Figure 1: The correlation between a (log-transformed) ratio and its (log-transformed) numerator for different input correlations between numerator and denominator.

nonlinearities remain so that the best-fitting associations are even stronger. **Figure 1** shows these correlations for a full range of r_{XY}. The values for $r_{X,X/Y}$ remain positive even under the least favorable conditions (i.e., when X and Y are increasingly redundant). The reason why the correlations are less than perfect is simply the researcher's ignorance of what the research hypothesis (~H) might be. It is neither a specific prediction nor a default-diffuse one.

We have seen that the LR can improve inductive inferences if a well-reasoned alternative hypothesis is available. A researcher who wishes to estimate the posterior probability of the null hypothesis, p(H|D), is better served by knowing p(D|H) and p(D|~H) than by knowing only the former. Yet, we also saw that the *p* value is a useful heuristic for predicting LR and p(H|D). Researchers may consider using this heuristic when available theory is too imprecise to make point predictions.

4. Are confidence intervals better predictors of replicability?

One long-standing alternative to *p* values are confidence intervals, CI, which provide more information than do *p* values (Loftus, 1993; 1996). The lower bound of a CI reveals whether statistical significance would have been obtained had one computed *p*. The upper bound reveals at what magnitude effect sizes become unlikely given the current data. The main focus of the CI approach is the estimation of the size of the latent effect rather than the question of whether that effect exists.

The Open Science Collaboration (2015) reported that the replicability of empirical results is low when evaluated in terms of ST or in terms of CI. However, Cumming (2014) proposed that the CI approach can yield stronger results (see Lindsay, 2015, for an enthusiastic endorsement of this method). To illustrate his claim, Cumming simulated 25 experiments each with a true effect of $\delta = .50$, and $N = 32$ for each of two independent samples. Eleven of these experiments yielded $p < .05$. In other words, this simulation is designed to model an environment where successful replication within ST is improbable. The statistical power, that is, the probability of rejecting the null hypothesis upon re-experimentation and assuming that the null hypothesis is false, is $1 - \beta = .53$, a virtual coin flip. With higher power, all tests eventually become significant. Because many experiments in psychological science are underpowered, Cumming's simulation has some ecological validity. He then showed that "in 20 of the 24 cases, the 95% CI includes the mean next above it in the figure" (p. 13; see also Cumming & Maillardet, 2006, for a similar and more precise result).

ST and the CI approaches use different definitions of replication (see also Open Science Collaboration, 2015). ST asks whether the point-value of the tested hypothesis (i.e., 0 in the case of NHST) lies outside of the CI in a second study if it lay outside of the CI in an original study, whereas the CI approach asks whether the mean of a first study lies within the CI of the second (where the order of the two is arbitrary). The CI approach is more liberal when empirical effects have the same sign (Simonsohn, 2015). Trivially, one might make ST equally liberal by, for example, simply asking whether, after an initial rejection of the null hypothesis, the mean of the second study has

the same sign. Researchers might balk at this suggestion, but imagine a scenario in which each of a large number of experiments yields a small nonsignificant effect of the same sign. Meta-analysis will assert the existence of a small effect regardless of whether the ST or the CI approach is used.

Inspection of Cumming's CI criterion reveals potentially awkward patterns. The second mean might lie within the CI of the first mean but have a different sign. This would be consistent with the ST view that a null finding was replicated, but the CI approach does not refer to a null hypothesis. So what has been replicated? Another concern involves sample size. As *N* increases, so does the precision of the parameter estimates. The shrinkage of the CI signals this greater precision, but it also reduces the probability that one sample mean will lie within the CI of another. The goal of maximizing the probability of replicating an inference therefore conflicts with the goal of maximizing the precision of measurement. If replication is defined by Cumming's criterion, small-sample experiments with wide CIs will yield the greatest number of successful replications. Imprecise and error-prone experimental designs produce CIs large enough to contain the means of similarly noisy follow-up studies. When sample size is increased, or other means are deployed to increase precision, the probability of a successful replication should go down. As CIs become smaller, it will be less likely that they contain the means of other studies. In short, increases in sample size produce conflicting consequences. On the one hand, they shrink the CI of an individual study, but on the other hand, they also shrink the variation of the means across studies. The question is whether this greater homogeneity in the means will offset the reduction in the width of the CI.

We sampled observations from a distribution with $\mu = 55$ and $\sigma = 10$ (i.e., $\delta = .5$ relative to the null distribution of $\mu = 50$ and $\sigma = 10$), computed a 95% CI around each observed mean, and conducted a one-sample *t*-test against a null hypothesis. We repeated this process 1,000 times for each of 10 steps of sample size ($N = 10$ to 100). **Table 3** shows the mean and the median widths of these CI, the SD of these means, as well as the mean and the median *p* values. Notice that the correlation between the width of the CI and the SD of the means was nearly perfect, $r = .997$, which suggests that, increasing the statistical power of a study does not increase the probability of a successful replication by the CI criterion.

We then estimated of the probability of a successful replication using both the CI and the ST frameworks. Assuming a false null hypothesis (i.e., p(H) = 0), we simulated the probability with which the mean obtained in one simulated experiment would fall within the CI of another experiment. The results in **Table 4** show that the CI measure predicts a high probability of replication regardless of sample size. In contrast, the *p* value is sensitive to sample size; as samples get larger, the probability of significance (i.e., *pp*) increases sharply. The joint probabilities in the rightmost column of the table show that with increasing sample size, the probability of finding significance in the first and the second study (i.e., pp^2) approaches 1. Over this series of simulations, the median probability of replication is remarkably similar for both the CI ($M = .841$) and ST ($M = .830$) approaches.

If the replicability of research findings is in question, the CI measure ignores the power of large studies to repeatedly yield the same result. When $N = 100$, our simulation yields a probability of replication of .85 and .98 respectively for the CI and the ST approach. In contrast, when samples are small ($N = 10$) and power is low, the probability of replication remains high (.85) for the CI measure, whereas ST yields a total probability of obtaining the same result of .60 (the probability of two significant results plus the probability of two non-significant results). The modest probabilities of replication offered by ST for small samples are not a flaw of method, but the sign of healthy skepticism. By casting the net for acceptable replications too wide, the CI approach inflates both false positive and false negative error rates – again, especially for small samples.[5]

Table 3: CI and *p* Values as a Function of Sample Size when p(H) = 0.

N	*M* CI Width	*Mdn* CI Width	SD of *M*	*M* of *p*	*Mdn p*
10	14.13	14.04	3.24	0.2467	0.1455
20	9.33	9.24	2.31	0.1249	0.0381
30	7.41	7.39	1.85	0.0635	0.0114
40	6.32	6.28	1.61	0.0292	0.0025
50	5.68	5.67	1.35	0.0125	0.0010
60	5.13	5.10	1.27	0.0068	0.0002
70	4.76	4.75	1.21	0.0039	0.0001
80	4.45	4.45	1.08	0.0015	0.0000
90	4.20	4.18	1.10	0.0011	0.0000
100	3.97	3.96	0.97	0.0005	0.0000

Note. M = mean; *Mdn* = median; SD = standard deviation.

Table 4: Probability of replication with CI and NHST.

	Confidence Interval Approach		NHST Approach	
N	***M* p(rep)**	**SD p(rep)**	**p (p <= .05)**	**p (sign.^{2})**
10	0.854	0.129	0.279	0.078
20	0.834	0.143	0.545	0.296
30	0.836	0.145	0.732	0.535
40	0.829	0.148	0.85	0.722
50	0.856	0.138	0.95	0.902
60	0.844	0.148	0.971	0.942
70	0.833	0.151	0.983	0.965
80	0.849	0.140	0.994	0.987
90	0.820	0.158	0.996	0.991
100	0.851	0.146	0.999	0.997
M	*0.841*	*0.145*	*0.830*	*0.742*

Note. p(rep) = probability of replication, defined as drawing a CI containing the mean of the previous experiment. p(sign.^{2}) = probability of finding significance ($p \leq .05$) twice in a row.

5. Review

Significance testing, ST, is meant to support statistical inference under uncertainty. As any method of inductive inference, ST faces many challenges, and it has been difficult to find a balanced evaluation of its strengths and weaknesses. Recently, diverse proposals have been made to reform statistical practice, such as lowering the threshold for statistical significance, adding alternative methods, and even abandoning ST altogether. As some of the uncertainty raised by questions of induction are irreducible, it is necessary to explore not only the strengths and weakness of a particular method, but to also ask how the balance of strengths and weaknesses compares with the strengths and weaknesses of other available methods. A comprehensive review of all methods along all possible criteria of validity is beyond the scope of any particular investigation. We therefore focused on four questions in the critical literature.

Using simulation experiments, we reproduced the statistical patterns associated with each concern. Then we showed that each concern is valid in the context of specific assumptions. By making assumptions more flexibly, we sought a broader evaluation of ST. In each of four areas of concern, we found that *p* values can help researchers reduce the uncertainty they face regarding the hypotheses at stake. Using *p* values as heuristic cues instead of rigid decision criteria, researchers can make informed inferences that outperform guessing and that predict inferences made with methods requiring more assumptions or informational input. We have also seen that some of the alternatives to ST are susceptible to shortcomings and limitations of their own. All told, we agree with Wasserstein (2017) that the greatest psychological challenge in the context of statistical induction is the mistaken belief that there may be a perfect way to do it.

To review, we first addressed the concern that *p* values are poor predictors of the posterior probability of the tested hypothesis. This concern goes to the heart of induction, which requires a leap from the known (evidence) to the unknown (the reality creating these data). We found that in plausible settings, the correlation between *p* values and the post-study probability of the tested hypothesis can be quite large. Perhaps these correlations are large enough to allow ST to play a continued role in scientific inference, especially when researchers carefully estimate a range of credible correlations between of *p* values and posterior probabilities in their area of study.

Second, we addressed the concern that *p* values are too variable to be useful. We showed that this concern is biased by its limited focus on a true null hypothesis. This restriction begs the question of *whether* the null is true. The probability distribution of *p* should be studied for different alternative hypotheses, a strategy encouraged by recent developments in *p*-curve analysis and related methods (Simonsohn et al., 2014). We introduced an 'inductive value' index, which considers both the obtained *p* value and its prior probability of occurring for a single study. Our guiding assumption is that lower *p* values are often more 'valuable' to the researcher than large *p* values because they are associated with large true effects. By this metric, ST performs well unless the null hypothesis is highly probable *a priori* and the to-be-detected effect is small.

Third, we compared likelihood ratios with *p* values. LRs can vary widely for a given *p* value depending on the alternative hypothesis considered, but a limited focus on this independent variation obscures the close log-linear relationship between the two indices when both are allowed to vary. When *p* values are held constant, their sensitivity to variations in the relative evidence against the statistical hypothesis is defined away. Our analysis suggests that the

p value should not be dismissed lightly because it (or its corresponding likelihood) determines the numerator (or the denominator depending on the analyst's preference) of the LR. If the LR performs well, it does so partly because the *p* value performs well. This is not to say that there is no place for the LR: when specific alternatives are defined, this can be a useful tool for evaluating the relative evidentiary strength of the evidence.

Fourth, we addressed the claim that confidence intervals provide better estimates of the replicability of an empirical result. We find that CI overlaps are uniformly large and that this is not a useful feature for the estimation of replicability. The width of the CI for a particular sample mean is highly correlated with the variability of means over different sample sizes. Therefore, estimates of replicability performed with CI are insensitive to statistical power. If the results of studies with high power are to be regarded as more predictive of replication than the results of studies with low power, ST should be preferred. The blindspot of ST lies elsewhere. When two studies yield significant results but very different effect size estimates, a CI analysis takes note, whereas ST does not. We urge researchers to pay close attention to effect sizes and CI as well as *p* values.

6. Conclusion

One of David Hume's lasting legacies is to have shown that no method of inductive inference can be justified deductively (Hume, 1978/1739). The few who tried to do this for ST (e.g., Chow, 1998) reaped a storm of criticism (see commentaries on Chow's article). Researchers have to accept the necessity of evaluating methods of inductive inference by indirect, inductive means. In this investigation, we used simulation experiments to find the conditions under which the *p* value performs well or poorly. Our findings are a partial response to a set of challenges to the validity of *p*. We find that these critiques carry their own burden of incompleteness. The proposed alternative methods have considerable strengths, which researchers can exploit under suitable conditions, but we do not share the view that ST is fatally flawed. The goal of any method of inductive reasoning and inference is to model and manage uncertainty. Often, the selection of inductive methods is itself a choice made under uncertainty. We hope that our investigation will help researchers discern when to include ST in their quest to understand their data.

We began this article with a note of how the debate over statistical analysis is often framed as an inquisition into the flaws of ST. The present article too is cast in the mold of this ongoing controversy. There have been many critical assessments (e.g., Lykken, 1968; Meehl, 1978; Rozeboom, 1960, are famous early examples, though far from the earliest) and occasional efforts at defense (Chow, 1998; García-Pérez, 2016; Hagen, 1997). The current climate is characterized by the question of why ST is still alive after so many fatal blows. A seemingly inescapable conclusion is that researchers who continue to use the *p* value are to be accused of inertia or mental denseness. Our approach is different (Krueger & Heck, 2017; see also Cohen, 1994; Krueger, 2001; Nickerson, 2000). We cast ST as a heuristic tool of induction, thereby trying to avoid what many critics condemn as ST's most fatal flaw: its encouragement of dichotomous decision-making (Amrhein, Korner-Nievergelt, & Roth, 2017; Greenland, 2017; McShane & Gal, 2016). We cannot conclude that ST, or any other method, is categorically good or bad. If we did, we might ourselves be guilty of dichotomania.

There is a lesson for future comparative efforts and intervention of institutional task forces. Instead of setting up ST as a defendant awaiting a verdict, it might be useful to articulate the mission of inductive inference and specific questions and challenges arising from it. We encourage careful consideration of when a statistical test might be necessary, and when estimation methods, unrestricted by dichotomania (Greenland, 2017), may be better suited for describing a set of results. Against these questions and challenges the various tools in the statistical box can then be measured and evaluated. Assuming that the perspective of ecological rationality can be extended to ecological induction, we may then identify the conditions under which individual tools perform better than their alternatives. Once this goal has been achieved – if it is achievable – statisticians and working researchers can move beyond the well-intended but vague advice to use their judgment and deploy statistical tools wisely.

Data Accessibility Statement

Matlab code for simulations [1] [2] and [4] can be found on Patrick Heck's website http://www.patrickrheck.com/data--materials.html. Simulation [3] was run in Excel.

Notes

[1] This does not mean that a low *p* value will *never* be produced in this scenario (see Bem, 2011; Replicability-Index, 2018)

[2] More generally, it is extremity that is inversely related to probability. Catastrophes are rarer than mishaps much like great joy is rarer than a pleasant mood.

[3] The argument presented in this section may also be stated as a probabilistic reverse inference. If there is an effect, *p* values will tend to be low. The *p* value is low. We conclude that there is probably an effect. This inference is logically invalid. Its degree of probabilistic validity depends, in part, on the factors discussed in this section.

[4] The same holds true for the upper-Bayes-factor-bound proposed by Bayarri et al. (2016), i.e., $\frac{1}{-e \times p \times \ln p}$.

[5] Simonsohn (2015) offered an innovative use of ST to answer one version of the replication riddle. His method finds the effect size δ that would have been detected by the original study with the probability of .33. Then, a replication study is conducted with a power of .80 to reject that δ. If the observed effect size, d, in the replication study is smaller than δ and if that difference is significant, the researcher concludes that the original study was so underpowered that the obtained significant result may not be trusted.

Acknowledgements

We thank Joe Austerweil, Dan Balliet, Dan Benjamin, Tony Evans, Florian Kutzner, and Jan Rummel, for sharing their insightful ideas about NHST and its limitations.

Competing Interests

The authors have no competing interests to declare.

Authors Contributions

JIK and PRH contributed equally to this work. JIK drafted the manuscript. PRH wrote the simulations and analyzed the simulated data. JIK and PRH revised the manuscript.

References

Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Erlbaum.

Amrhein, V., Korner-Nievergelt, F., & **Roth, T.** (2017). The earth is flat ($p > 0.05$): significance thresholds and the crisis of unreplicable research. *PeerJ*: e3544. DOI: https://doi.org/10.7717/peerj.3544

Bayarri, M. J., Benjamin, D. J., Berger, J. O., & **Sellke, T. M.** (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology, 72*, 90–103. DOI: https://doi.org/10.1016/j.jmp.2015.12.007

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. DOI: https://doi.org/10.1037/a0021524

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., . . ., & **Cesarini, D.** (2017). Redefine statistical significance. *Nature Human Behaviour*, 1. DOI: https://doi.org/10.1038/s41562-017-0189-z

Bernoulli, D. (1954/1738). Exposition of a new theory on the measurement of risk. *Econometrika, 22,* 23–36. DOI: https://doi.org/10.2307/1909829

Brewer, M. B. (2007). The social psychology of intergroup relations: Social categorization, ingroup bias, and outgroup prejudice. In Kruglanski, A. W., & Higgins, E. T. (Eds.), *Social psychology: Handbook of basic principles*. New York, NY, US: Guilford Press, pp. 695–715.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & **Munafò, M. R.** (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 365–376. DOI: https://doi.org/10.1038/nrn3475

Chow, S. L. (1998). Précis of statistical significance: Rationale, validity, and utility. *Behavioral and Brain Sciences, 21*, 169–239. DOI: https://doi.org/10.1017/S0140525X98001162

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304–1312. DOI: https://doi.org/10.1037/0003-066X.45.12.1304

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

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, 3, 286–300. DOI: https://doi.org/10.1111/j.1745-6924.2008.00079.x

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

Cumming, G., & **Maillardet, R.** (2006). Confidence intervals and replication: where will the next mean fall?. *Psychological Methods*, *11*(3), 217–227. DOI: https://doi.org/10.1037/1082-989X.11.3.217

Dawes, R. M. (1991). Probabilistic versus causal thinking. In Cicchetti, D., & Grove, W. M. (Eds.), *Thinking clearly about psychology: Vol. 1. Matters of public interest: Essays in honor of Paul Everett Meehl*. Minneapolis: University of Minnesota Press, pp. 235–264.

Eich, E. (2014). Business not as usual. *Psychological Science*, *25*, 3–6. DOI: https://doi.org/10.1177/0956797613512465

Fiedler, K. (2017). What constitutes strong psychological science? The (neglected) role of diagnosticity and a-priori theorizing. *Perspectives on Psychological Science*, 12, 46–61. DOI: https://doi.org/10.1177/1745691616654458

García-Pérez, M. A. (2016). Thou shalt not bear false witness against null hypothesis significance testing. *Educational and Psychological Measurement*, 1–32. Online first. DOI: https://doi.org/10.1177/0013164416668232

Gelman, A. (2013). *P* values and statistical practice. *Epidemiology, 24*, 69–72. DOI: https://doi.org/10.1097/EDE.0b013e31827886f7

Gelman, A., & **Stern, H.** (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician, 60*, 328–331. DOI: https://doi.org/10.1198/000313006X152649

Gigerenzer, G. (1998). Surrogates for theories. *Theory & Psychology, 8*, 195–204. DOI: https://doi.org/10.1177/0959354398082006

Gigerenzer, G., & **Marewski, J.** (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management, 41*, 421–440. DOI: https://doi.org/10.1177/0149206314547522

Goodman, S. N., & **Royall, R.** (1988). Evidence and scientific research. *American Journal of Public Health, 78*, 1568–1574. DOI: https://doi.org/10.2105/AJPH.78.12.1568

Greenland, S. (2017). The need for cognitive science in methodology. *American Journal of Epidemiology*. *6*, 639–645. DOI: https://doi.org/10.1093/aje/kwx259

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & **Altman, D. G.** (2016). Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology, 31*, 337–350. DOI: https://doi.org/10.1007/s10654-016-0149-3

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52*, 15–24. DOI: https://doi.org/10.1037/0003-066X.52.1.15

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & **Drummond, G. B.** (2015). The fickle *P* value generates irreproducible results. *Nature Methods, 12*, 179–185. DOI: https://doi.org/10.1038/nmeth.3288

Hume, D. (1978). *A treatise of human nature.* Glasgow, Scotland: William Collins (Original work published 1739).

Jeffreys, H. (1961). *Theory of probability,* (3rd ed.). Oxford, UK: Oxford University Press.

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review, 2*, 196–217. DOI: https://doi.org/10.1207/s15327957pspr0203_4

Krueger, J. I. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist, 56*, 16–26. DOI: https://doi.org/10.1037//0003-066X.56.1.16

Krueger, J. I. (2017). Reverse inference. In Lilienfeld, S. O., & Waldman, I. D. (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions*. New York, NY: Wiley, pp. 110–124. DOI: https://doi.org/10.1002/9781119095910.ch7

Krueger, J. I., & **Heck, P. R.** (2017). The heuristic value of *p* in inductive statistical inference. *Frontiers in Psychology, 8*, 108–116. DOI: https://doi.org/10.3389/fpsyg.2017.00908

Krueger, J. I., Heck, P. R., & **Asendorpf, J. B.** (2017). Self-enhancement: Conceptualization and assessment. *Collabra: Psychology: Psychology, 3*(1), 28. DOI: https://doi.org/10.1525/collabra.91

Kruschke, J. K., & **Lidell, T. M.** (2017). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*. Online first. DOI: https://doi.org/10.3758/s13423-016-1221-4

Lee, M. D., & **Wagenmakers, E.-J.** (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review, 112*, 662–668. DOI: https://doi.org/10.1037/0033-295X.112.3.662

Lilienfeld, S. O., Sauvigné, K. C., Lynn, S. J., Cautin, R. L., Latzman, R. D., & **Waldman, I. D.** (2015). Fifty psychological and psychiatric terms to avoid: a list of inaccurate, misleading, misused, ambiguous, and logically confused words and phrases. *Frontiers in Psychology, 6*. DOI: https://doi.org/10.3389/fpsyg.2015.01100

Lilienfeld, S. O., & **Waldman, I. D.** (2017). *Psychological science under scrutiny: Recent challenges and proposed solutions.* New York, NY: Wiley. DOI: https://doi.org/10.1002/9781119095910

Lindley, D. V. (1975). The future of statistics: a Bayesian 21st century. *Advances in Applied Probability (Suppl.) 7*, 106–115. DOI: https://doi.org/10.2307/1426315

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science, 26*, 1827–1832. DOI: https://doi.org/10.1177/0956797615616374

Loftus, G. R. (1993). Editorial comment. *Memory & Cognition, 23*, 1–3.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5*, 161–171. DOI: https://doi.org/10.1111/1467-8721.ep11512376

Lykken, D. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70*, 151–159. DOI: https://doi.org/10.1037/h0026141

Magnusson, K. (2015, August 13) *P-curve visualization updated with log x-axis.* Retrieved from: http://rpsychologist.com/updated-d3-js-visualization-p-curve-distribution.

Mayo, D. G. (2016). Don't throw out the error control baby with the bad statistics bathwater: A commentary. *The American Statistician,* Online Discussion. Retrieved from: http://amstat.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108?scroll=top.

McShane, B. B., & **Gal, D.** (2016). Blinding us to the obvious? The effect of statistical training on the evaluation of evidence. *Management Science, 62*, 1707–1718. DOI: https://doi.org/10.1287/mnsc.2015.2212

McShane, B. B., Gal, D., Gelman, A., Robert, C., & **Tackett, J. L.** (2017). Abandon statistical significance. *arXiv preprint*. arXiv:1709.07588.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806–834. DOI: https://doi.org/10.1037/0022-006X.46.4.806

Murdoch, D. J., Tsai, Y.-L., & **Adcock, J.** (2008). P-Values are random variables. *The American Statistician, 62*, 242–245. DOI: https://doi.org/10.1198/000313008X332421

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 5*, 241–301. DOI: https://doi.org/10.1037/1082-989X.5.2.241

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716–aac4716. DOI: https://doi.org/10.1126/science.aac4716

Ortega, A., & **Navarrete, G.** (2017). Bayesian hypothesis testing: An alternative to Null Hypothesis Signficance Testing (NHST) in psychology. In Tejedor, J. P. (Ed.), *Bayesian inference*. Open access. DOI: https://doi.org/10.5772/intechopen.70230

Perezgonzalez, J. D. (2015). The meaning of significance in data testing. *Frontiers in Psychology, 6*, 1293. DOI: https://doi.org/10.3389/fpsyg.2015.01293

Pleskac, T. J., & **Hertwig, R.** (2014). Ecologically rational choice and the structure of the environment. *Journal of Experimental Psychology: General, 143*, 2000–2019. DOI: https://doi.org/10.1037/xge0000013

Replicability-Index. (2018, January 5). Why the journal of personality and social psychology should retract article "Feeling the Future: Experimental evidence for anomalous retroactive influences on cognition and affect" by Daryl J. Bem [Blog post].

Retrieved from: https://replicationindex.wordpress.com/2018/01/05/why-the-journal-of-personality-and-social-psychology-should-retract-article-doi-10-1037-a0021524-feeling-the-future-experimental-evidence-for-anomalous-retroactive-influences-on-cognition-a/. DOI: https://doi.org/10.1037/a0021524

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & **Wagenmakers, J.-E.** (2016). Is there a free lunch in inference? *Topics in Cognitive Science, 8*, 520–547. DOI: https://doi.org/10.1111/tops.12214

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

Senn, S. J. (2001). Two cheers for *p*-values. *Journal of Epidemiology and Biostatistics, 6*, 193–204. DOI: https://doi.org/10.1080/135952201753172953

Senn, S. J. (2017). Contribution to the discussion of "A critical evaluation of the current *p*-value controversy." *Biometrical Journal*. DOI: https://doi.org/10.1002/bimj.201700032

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. DOI: https://doi.org/10.1177/09567976114 17632

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*, 559–569. DOI: https://doi.org/10.1177/0956797614567341

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

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

Trafimow, D., & **Rice, S.** (2009). A test of the null hypothesis significance testing procedure correlation argument. *The Journal of General Psychology, 136*, 261–269. DOI: https://doi.org/10.3200/GENP.136.3.261-270

van der Linden, S., & **Chryst, B.** (2017). No need for Bayes Factors: A fully Bayesian evidence synthesis. *Frontiers in Applied Mathematics and Statistics, 3*. http://journal.frontiersin.org/article/10.3389/fams.2017.00012/full.

Wagenmakers, E.-J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., & **Morey, R.** (2017). The need for Bayesian hypothesis testing in psychological science. In Lilienfeld, S. O., & Waldman, I. D. (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions*. New York, NY: Wiley, pp. 123–138. DOI: https://doi.org/10.1002/9781119095910.ch8

Wasserstein, R. L. (2017, April). The game is still afoot. Comment on Matthews, 2017. *Significance*, 40–41.

Wasserstein, R. L., & **Lazar, N. A.** (2016). The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician, 70*, 129–133. Available at: http://amstat.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108?scroll=top. DOI: https://doi.org/10.1080/00031305.2016.1154108

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 of 855 t tests. *Perspectives on Psychological Science, 6*, 291–298. DOI: https://doi.org/10.1177/1745691611406923

Wilkinson, L., & **the APA Task Force on Statistical Inference.** (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604. DOI: https://doi.org/10.1037/0003-066X.54.8.594

Peer review comments
The author(s) of this paper chose the Open Review option, and the peer review comments are available at: http://doi.org/10.1525/collabra.108.pr

How to cite this article: Krueger, J. I. and Heck, P. R. (2018). Testing Significance Testing. *Collabra: Psychology,* 4(1): 11. pp. 1–13, DOI: https://doi.org/10.1525/collabra.108

Senior Editor: Simine Vazire

Editor: Victoria Savalei

Submitted: 22 August 2017 **Accepted:** 30 March 2018 **Published:** 26 April 2018