



The Practical Alternative to the p Value Is the Correctly Used p Value

Daniël Lakens

Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology

Abstract

Because of the strong overreliance on p values in the scientific literature, some researchers have argued that we need to move beyond p values and embrace practical alternatives. When proposing alternatives to p values statisticians often commit the “statistician’s fallacy,” whereby they declare which statistic researchers really “want to know.” Instead of telling researchers what they want to know, statisticians should teach researchers which questions they can ask. In some situations, the answer to the question they are most interested in will be the p value. As long as null-hypothesis tests have been criticized, researchers have suggested including minimum-effect tests and equivalence tests in our statistical toolbox, and these tests have the potential to greatly improve the questions researchers ask. If anyone believes p values affect the quality of scientific research, preventing the misinterpretation of p values by developing better evidence-based education and user-centered statistical software should be a top priority. Polarized discussions about which statistic scientists should use has distracted us from examining more important questions, such as asking researchers what they want to know when they conduct scientific research. Before we can improve our statistical inferences, we need to improve our statistical questions.

Keywords

p values, null-hypothesis testing, equivalence tests, statistical inferences

Scientific progress requires answering a highly diverse set of questions. Sometimes researchers try to answer questions by specifying a model of the world. To examine whether these models have predictive power, researchers collect data with the aim to test hypotheses derived from these models. Statistical inferences can then be used to interpret the data that have been collected. Researchers can choose to use a wide range of statistical tools to make inferences, including p values, effect sizes, confidence intervals, likelihood ratios, Bayes factors, and posterior distributions. It is rare to find an article in the statistical literature that presents all of these approaches to statistical inferences as valid answers to questions a researcher might be interested in. In particular, p values are often dismissed as a useful tool for answering scientific questions. In this article I evaluate whether p values provide an answer to a question researchers would want to know, whether alternatives to p values would fare any better in the hands of researchers, and how we can improve the use of p values in practice.

Researchers have criticized the overreliance on null-hypothesis significance testing (NHST) and common

misconceptions about p values for more than half a century (e.g., Bakan, 1966; Nunnally, 1960; Rozeboom, 1960). The correct definition of a p value is the probability of observing the sample data, or more extreme data, assuming the null hypothesis is true. The interpretation of a p value depends on the statistical philosophy one subscribes to. In a Fisherian framework a p value is interpreted as a continuous measure of compatibility between the observed data and the null hypothesis (Greenland et al., 2016). The compatibility of observed data with the null model falls between 1 (perfectly compatible) and 0 (extremely incompatible), and every individual can interpret the p value with “statistical thoughtfulness” (Wasserstein, Schirm, & Lazar, 2019). In a Neyman-Pearson framework, the goal of statistical tests is to guide the behavior of researchers with respect to a hypothesis. On the basis of the results of a statistical test, and without ever knowing whether

Corresponding Author:

Daniël Lakens, Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology
E-mail: d.lakens@tue.nl

the hypothesis is true or not, researchers choose to tentatively *act* as if the null hypothesis or the alternative hypothesis is true. In psychology, researchers often use an imperfect hybrid of the Fisherian and Neyman-Pearson frameworks, but the latter is, according to Dienes (2008), “the logic underlying all the statistics you see in the professional journals of psychology” (p. 55).

The widespread use of p values is criticized for two main reasons. First, researchers often misinterpret p values or mindlessly apply hypothesis testing. Second, in many situations the point null hypothesis of an effect of exactly zero is unlikely to be true, in which case asking if it can be rejected is a relatively uninteresting question. Some journals, such as *Basic and Applied Psychology*, *Epidemiology*, and *Political Analysis*, have banned the use of p values in an attempt to improve statistical inferences in the articles they publish (Fidler, Thomason, Cumming, Finch, & Leeman, 2004; Gill, 2018; Trafimow & Marks, 2015). There is an overwhelming range of proposed alternatives to p values (see the special issue of *The American Statistician*; Wasserstein et al., 2019). Hubbard (2019) reviews how, although criticisms of NHST and p values have received widespread attention, little has changed in practice. He notes how a possible reason for the lack of change is that statisticians¹ rarely explicitly state the circumstances in which the use of p values is *not* problematic and in which null-hypothesis significance testing provides a useful answer to a question of interest.

When we survey the literature, we rarely see the viewpoint that *all* approaches to statistical inferences, including p values, provide answers to specific questions a researcher might want to ask. Instead, statisticians often engage in what I call the “statistician’s fallacy”—a declaration of what they believe researchers really “want to know” without limiting the usefulness of their preferred statistical question to a specific context. The most well-known example of the statistician’s fallacy is provided by Cohen (1994) when discussing null-hypothesis significance testing:

What’s wrong with NHST? Well, among many other things, 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! What we want to know is ‘Given these data, what is the probability that H_0 is true?’ (p. 997)

Other statisticians have disagreed with Cohen about what it is “we want to know.” Colquhoun (2017) thought that “what you want to know is that when a statistical test of significance comes out positive, what is the probability that you have a false positive” (p. 2). Kirk

(1996) said that “what we want to know is the size of the difference between A and B and the error associated with our estimate” (p. 754). Blume (2011), on the other hand, suggested that “what we really want to know is how likely it is that the observed data are misleading” (p. 509). Bayarri, Benjamin, Berger, and Sellke (2016) believed that “we want to know how strong the evidence is, given that we actually observed the value of the test statistic that we did” (p. 91). Finally, Mayo (2018) argued that “we want to know what the data say about a conjectured solution to a problem: What erroneous interpretations have been well ruled out?” (p. 300). Thus, according to six different (groups of) statisticians, what we want to know is the posterior probability of a hypothesis, the false-positive risk, the effect size and its confidence interval, the likelihood, the Bayes factor, or the severity with which a hypothesis has been tested.

I call these beliefs about what researchers want to know a fallacy, which might sound severe, but I believe the arguments provided by these statisticians for their claims about what we want to know boil down to nothing more than wishful thinking. Some statisticians have used common misconceptions of p values as an argument for their choice of what researchers really want to know. Cohen (1994) explained that a p value does not provide the probability that the null hypothesis is true, but the posterior probability does. Colquhoun (2017) explained that a p value does not provide the probability that the results have occurred by chance, but the false-positive risk does. Kirk (1996) noted how a nonsignificant p value can be incorrectly interpreted as the absence of an effect, even when the size of the effect supports the alternative hypothesis. However, the fact that common misinterpretations correspond to completely different statistical entities, together with the larger context in which these statisticians made their claims,² suggests that all statisticians seem to mean “what I wish you wanted to know,” or more normatively, “what I think you should want to know.” Even if we could define a reference class for “we,” it is doubtful that all people included in this category would unanimously agree. Furthermore, it seems highly unlikely that there is a single thing anyone wants to know at all times or that asking a single statistical question leads to the most efficient empirical progress. Researchers often ask different questions at distinct phases of a research project, and the questions they ask depend on the field, the specific study, the reliability and availability of previous knowledge, and their philosophy of science. The first point I want to make in this article is that we stop teaching researchers that there is something they want to know. There is no room for the statistician’s fallacy in our journals or in our statistics education. I do not think that it is useful to tell researchers what

they want to know. Instead, we should teach them the possible questions they can ask (Hand, 1994).

Are p Values Ever Something Anyone Wants to Know?

Savalei and Dunn (2015) have argued that “the strong NHST-bashing rhetoric common on the ‘reformers’ side of the debate may prevent many substantive researchers from feeling that they can voice legitimate reservations about abandoning the use of p values” (para. 2). Nevertheless, some researchers have argued that p values can provide an interesting answer to a statistical question whenever researchers want to make an ordinal claim about the direction of an effect (Abelson, 1997b; Chow, 1988; Cortina & Dunlap, 1997; Hagen, 1997; Haig, 2017; Miller, 2017; Nickerson, 2000). Although Meehl harshly criticized the overreliance of psychology on NHST (Meehl, 1978), he also noted, “When I was a rat psychologist, I unabashedly employed significance testing in latent-learning experiments; looking back I see no reason to fault myself for having done so in the light of my present methodological views” (Meehl, 1990, p. 138). Abelson (1997a) writes, “Realistically, if the null hypothesis test did not exist, it would have to be (re)invented” (p. 118). In his book *Beyond Significance Testing*, Kline (2004) wrote, “The ability of NHST to address the dichotomous question of whether relations are greater than expected levels of sampling error may be useful in some new research areas” (p. 86). Cohen agreed in a 1995 rejoinder to his 1994 article that rejecting a point null hypothesis in a strictly controlled experiment can be a useful way of establishing the direction of an effect, whenever this question is central to the purpose of the experiment (Cohen, 1995, p. 1103).

When discussing the question a p value can answer, I focus on the use of p values in a Neyman-Pearson approach to statistical inferences, which Hacking (1965) considers “very nearly the received theory on testing statistical hypotheses” (p. 84). A Neyman-Pearson hypothesis test is worth performing if two conditions are met. First, the null hypothesis should be plausible enough so that rejecting it is surprising, at least for some readers. This is typically easier to accomplish in a controlled experiment than in a correlational study because in the latter variables are typically connected through causal structures that result in real nonzero correlations, known as the “crud factor” (Meehl, 1990; Orben & Lakens, 2020). Second, the researcher is interested in applying a methodological procedure that allows him or her to make decisions about how to act while controlling error rates. Neyman and Pearson (1933) were very clear that they did not intend to

develop a method to inform us about the probability that our hypotheses are true, but that, “without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong” (p. 291).

This “act” is not limited to the decision to adopt a treatment, intervention, or government policy. The act can also be the decision to abandon a research line, to change a manipulation, or even, under a slightly broader interpretation of an act, to make a certain type of statement or claim (Cox, 1958; Frick, 1996). On the basis of carefully controlled studies, researchers can use NHST to make ordinal claims, such as the claim that the mean in one condition is larger than the mean in another condition. If we look at articles in the scientific literature, researchers often seem to be interested in making such ordinal claims, especially in the context of theory corroboration (Abelson, 1997a; Chow, 1988). Any time researchers make a claim, they can do so erroneously. The Neyman-Pearson approach to hypothesis testing allows researchers to limit the frequency or erroneous claims in the long run by choosing the α level and designing a study with a desired statistical power for a specified effect size.

Researchers are free to refrain from making claims in their article about whether hypotheses are corroborated or not. Rozeboom (1960) criticized the use of NHST because “the primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one accepts, or believes, the hypothesis or hypotheses being tested.” (p. 420). If this is your philosophy, then a p value is unlikely to provide the answer you are looking for, and you might prefer to draw nondichotomous inferences using a likelihood ratio, Bayes factor, or a Fisherian interpretation of p values. However, if you are interested in establishing claims about ordinal effects, distinguishing signal from noise, and drawing valid conclusions on the basis of data, and if you want to ensure that in the long run you will not be wrong too often, the Neyman-Pearson approach to statistical inferences might, when correctly used, answer a question of interest.

Why Would Alternatives to p Values Fare Any Better?

The suggestion that research practices would improve if we no longer relied on p values and NHST (e.g., Cumming, 2014; Trafimow & Marks, 2015) lacks empirical support. Hanson (1958) examined the replicability of research findings published in anthropology,

psychology, and sociology as a function of whether claims were based on explicit confirmation criteria, such as the rejection of a hypothesis at a 5% significance level, and found that such claims were more replicable than claims made without such an explicit confirmation criterium. He noted that “over 70 per cent of the original propositions advanced with explicit confirmation criteria were later confirmed in independent tests, while less than 46 per cent of the propositions advanced without explicit confirmation criteria were later confirmed” (p. 363). I do not know of any other empirical research that has examined this question, but this finding is in line with qualitative analyses of the null-hypothesis significance ban in the journal *Basic and Applied Social Psychology* (Fricker, Burke, Han, & Woodall, 2019), which revealed that authors will claim that data support their prediction with a higher error rate than an α level of 5%, leading Fricker and colleagues to conclude that “when researchers only employ descriptive statistics we found that they are likely to overinterpret and/or overstate their results compared to a researcher who uses hypothesis testing with the $p < 0.05$ threshold” (p. 380).

Although there is little doubt that complementing p values with other statistics (such as effect sizes and confidence intervals) is often a good idea, as each statistic provides an answer to a different question of interest, some past suggestions to replace p values have not fared particularly well. For example, p_{rep} (Killeen, 2005) was used by the journal *Psychological Science* as a measure that should convey some information about the probability that a finding would replicate (Iverson, Lee, & Wagenmakers, 2009), until it was severely criticized and is now no longer reported. In some research articles in sports science, p values were replaced by magnitude-based inferences (Batterham & Hopkins, 2006), which were recently strongly criticized because of their high error rates (Sainani, 2018). Recently proposed “second-generation p values” (Blume, D’Agostino McGowan, Dupont, & Greevy, 2018) turned out to be highly similar to, but less informative than, equivalence tests (Lakens & Delacre, 2020). Training researchers how to use existing frequentist and Bayesian approaches to estimation and hypothesis testing well (which means with care and while acknowledging the limitations of each approach) might be a more fruitful approach for improving statistical inferences than developing novel statistical approaches. As Cohen (1994) concluded, “don’t look for a magic alternative to NHST, some other objective mechanical ritual to replace it. It doesn’t exist” (p. 1001).

The correct use of established frequentist and Bayesian methods will often lead to similar statistical inferences. In a recent study in the gerontology literature in

which four null effects were evaluated with equivalence tests or Bayes factors (Lakens, McLatchie, Isager, Scheel, & Dienes, 2020), both approaches led to similar inferences in each example. Likewise, four teams of researchers using frequentist or Bayesian hypothesis testing or estimation independently reached similar conclusions when reanalyzing two studies (Dongen et al., 2019). Although one can always find exceptions if one searches long enough, in most cases Bayes factors and p values will strongly agree (Tendeiro & Kiers, 2019). Jeffreys, who developed a Bayesian hypothesis test, noted the following when comparing the inferences using his procedure against frequentist methods proposed by Fisher:

I have in fact been struck repeatedly in my own work, after being led on general principles to a solution of a problem, to find that Fisher had already grasped the essentials by some brilliant piece of common sense, and that his results would be either identical with mine or would differ only in cases where we should both be very doubtful. As a matter of fact I have applied my significance tests to numerous applications that have also been worked out by Fisher’s, and have not yet found a disagreement in the actual decisions reached. (Jeffreys, 1939, p. 394)

If alternative approaches largely lead to the same conclusions as a p value when used with care, perhaps we can improve research practices more by focusing on transparency when reporting results, theory development, and measurement instead of extensively debating which statistical test researchers should or should not report.

Although statistical misconceptions are not limited to p values, it is true that NHST and p values are often misunderstood. It is therefore remarkable that there is so very little empirical research that examines how we can train scientists to prevent these misinterpretations (Sotos, Vanhoof, Van den Noortgate, & Onghena, 2007). One exception is research on the mistake to interpret a p value larger than .05 as evidence for the absence of an effect. A nonsignificant result means that an effect size of zero cannot be rejected, but neither can we reject effect sizes in a range around zero. It is therefore never possible to conclude that there is no effect. At best, we can use an equivalence test to examine whether the observed effect falls in a range of values close enough to zero to conclude that any effect that is present is too small to matter (Lakens, Scheel, & Isager, 2018). Indeed, Parkhurst (2001) reports the anecdotal observation that the proportion of students who misinterpret $p > .05$ as the absence of an effect declined

dramatically when students were taught equivalence tests. Research by Fidler and Loftus (2009) shows that presenting a figure with confidence intervals alongside the results of a t test reduces the mistake of interpreting $p > .05$ as the absence of an effect, although confidence intervals themselves are not immune to being misunderstood (Hoekstra, Morey, Rouder, & Wagenmakers, 2014).

In my own work I have also observed that students in a massive open online course made many errors when attempting to correctly interpret p values (Herrera-Bennett, Heene, Lakens, & Ufer, 2020). However, a similar number of errors were made on questions concerning the correct interpretation of confidence intervals and Bayes factors, providing further support that misconceptions are not limited to p values. Most importantly, however, students on average made considerable progress during the course: the percentage of correct responses increased from 8.3 out of 14 to 11.1. This increase highlights the importance of further research on how to best train scientists to prevent statistical misconceptions. We should acknowledge that research on how to prevent the misuse of statistics most likely needs to take the reward structures in academia into account.

Have We Really Tried Hard Enough?

Any statistician who cares about the practical impact of the discipline should be embarrassed by the continued inability of scientists to correctly interpret the meaning of a p value. The problems have been pointed out in hundreds of articles, but very little progress has been made (Gigerenzer, 2018). This problematic situation is not unlike something that happened in experimental psychology in which problems with publication bias, low power, and inflated α levels had been pointed out for decades without any noticeable effect. But even after ignoring important problems for decades, change is possible. Psychologists are embracing Registered Reports as a solution for publication bias (Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014; Nosek & Lakens, 2014), large collaborative research efforts have been started to empirically examine the replicability of psychological findings (Klein et al., 2014; Open Science Collaboration, 2015), and new journals dedicated to training researchers to improve their research practices (Simons, 2018) and publishing meta-scientific work in psychology (Carlsson et al., 2017) have emerged. I see no reason why a similarly collaborative effort to improve the widespread misunderstanding of p values would fail.

When I was taught German, my teacher spent weeks training us to remember “*aus bei mit nach seit von zu*” and “*bis durch für gegen ohne um*.” Nouns and pronouns following the first list of prepositions will always

be in the dative, whereas nouns and pronouns following the second list will always be in the accusative. The teacher expected us to repeat this list on the beat of his wedding ring as he tapped on his desk, and we were not supposed to miss a beat. Today, 25 years after I was taught these prepositions, I can still remember them. How many students leave our university with the ability to repeat and understand the definition of a p value from memory? If anyone seriously believes the misunderstanding of p values affects the quality of scientific research, why are we not investing more effort to ensure that misunderstandings of p values are resolved before young scholars perform their first research project? Although I am sympathetic to statisticians who think that all of the information researchers need to educate themselves on this topic is already available, as an experimental psychologist who works in human-technology interaction, this reminds me too much of the engineer who argues that all of the information for understanding the copy machine is available in the user manual. In essence, the problems we have with how p values are used is a *human-factors* problem (Tryon, 2001). The challenge is to get researchers to improve the way they work.

Looking at the deluge of research published in the past half century that point out how researchers have consistently misunderstood p values, I am left to wonder: Where is the innovative coordinated effort to create world-class educational materials that can freely be used in statistical training to prevent such misunderstandings? It is now relatively straightforward to create online apps that can simulate studies and show the behavior of p values across studies, which can easily be combined with exercises that fit the knowledge level of bachelor's and master's students.

The second point I want to make in this article is that a dedicated attempt to develop evidence-based educational material in a cross-disciplinary team of statisticians, educational scientists, cognitive psychologists, and designers seems worth the effort if we really believe that young scholars should understand p values. I do not think that the effort statisticians have made to complain about p values is matched with a similar effort to improve the way researchers use p values and hypothesis tests. We really have not tried hard enough. Where is the statistical software that does not simply return a p value but provides a misinterpretation-free verbal interpretation of the test? The statistical software package IBM SPSS is 40 years old, but in none of its 26 editions did it occur to the creators that it might be a good idea to provide researchers with the option to compute an effect size when performing a t test. We might need to take it onto ourselves as a research community to create better statistical software that returns results in a way that, for example, prevents us from

interpreting a p value larger than .05 as the absence of an effect. From a human-factors perspective there seems to be room for substantial improvement. Where is the word-processor plug-in that detects incorrect interpretations of p values, akin to how a spellchecker can automatically prevent spelling mistakes? Are we not technically able to flag statements such as “no effect of” in a document that occurs in proximity of $p > .05$? If we do not know how to prevent misinterpretations of a p value, do we know how to prevent misinterpretations of any alternatives that are proposed? To cite just one example, a review by van de Schoot, Winter, Ryan, Zondervan-Zwijenburg, and Depaoli (2017) revealed that 31% of articles in the psychological literature that used Bayesian analyses did not even specify the prior that was used, at least in part because the defaults by the software package were used. Mindless statistics are not limited to p values.

Testing Range Predictions

Most problems attributed to p values are problems with the practice of null-hypothesis significance testing. For example, one misinterpretation in NHST is that people interpret a significant result as an important effect (ignoring that samples that are sufficiently large can make even trivial differences from zero reach statistical significance). One of the most widely suggested improvements of the use of p values is to replace null-hypothesis tests (in which the goal is to reject an effect of exactly zero) with tests of range predictions (in which the goal is to reject effects that fall outside of the range of effects that is predicted or considered practically important). This idea is hardly novel, although the distinction between a null-hypothesis test and the test of a range prediction is worth stating explicitly. One example of a range prediction is a test that aims to reject effects smaller than a Cohen's d of 0.2. Such a test allows one to conclude that the effect is not only different from zero but also large enough to be meaningful. Hodges and Lehmann (1954) wrote, “About the set H_0 we may then distinguish a larger set of H_1 of values, representing situations close enough to H_0 that the difference is not *materially significant* [emphasis added] in the problem at hand,” adding, “It might be objected that there is nothing novel in the point of view just presented” (p. 262). Nunnally (1960) noted that “an alternative to the null hypothesis is the ‘fixed-increment’ hypothesis. In this model, the experimenter must state in advance how much of a difference is an important difference” (p. 644). Serlin and Lapsley (1985) discussed the “good-enough principle,” whereby a statistical test is performed against a “good-enough belt of width Δ ” such that “even with an infinite sample

size, the point-null hypothesis, fortified with a good-enough belt, is not always false” (p. 79).

In practice, researchers often have a smallest effect size of interest that is determined either by theoretical predictions, the practical significance of the effect, or the feasibility of studying a research question with the available resources (Lakens, 2014). Performing statistical tests to reject effects closer to zero than the smallest effect size of interest, known as minimum-effect tests (Murphy & Myers, 1999), or testing whether we can reject the presence of effects as large or larger than the smallest effect size of interest, known as equivalence tests (Lakens, Scheel, & Isager, 2018; Rogers, Howard, & Vessey, 1993), are often more interesting than testing against an effect of exactly zero.

For example, Burriss and colleagues (2015) examined a prediction from evolutionary psychology that a slight increase in redness in the face signals when women are most fertile to attract men. Data from 22 women revealed a statistically significant increase in the redness of their facial skin during their fertile period. If these authors had limited their analysis to a null-hypothesis test, they would have concluded that their prediction was supported. However, their theory predicted not just an increase in redness of the face but an increase in redness of the face *that was noticeable by men*. Their analyses revealed that the increase in redness was not prominent enough to be noticeable by the naked eye. This is a nice example of how a statistically significant effect is not misinterpreted as a meaningful effect by complementing a null-hypothesis test by a minimum-effect test. Likewise, the use of equivalence tests can prevent misinterpreting a nonsignificant effect as the absence of a meaningful effect (Lakens, Scheel, & Isager, 2018; Parkhurst, 2001).

Although minimum-effect tests and equivalence tests will still return a p value as the main result and still answer the question about whether an ordinal claim can be made or not, they also force researchers to ask more interesting questions. One interesting question that researchers rarely consider when making a prediction is what would falsify the hypothesis. An important starting point for answering such a question in experimental research is what the smallest effect size of interest would be. Imagine one theory predicts an effect size of a Cohen's d of 0.3 or larger, and another theory predicts the absence of a meaningful effect, which the researchers define as any effect between $d = -0.1$ and 0.1. We can design a randomized controlled experiment with high statistical power and a low α level that will yield informative results where either one, or the other, or both theories are falsified.

I have explained these alternative approaches to hypothesis tests in some detail because they use the

same machinery as NHST, including the computation of p values, but ask slightly different questions concerning the direction of effects. Tests of range predictions have been proposed as an improvement to NHST for more than half a century but rarely feature in discussions about statistical reform. As Haig (2017) notes,

Relatedly, advocates of alternatives to NHST, including some Bayesians (e.g., Wagenmakers, 2007) and the new statisticians (e.g., Cumming, 2014), have had an easy time of it by pointing out the flaws in NHST and showing how their preferred approach does better. However, I think it is incumbent on them to consider plausible versions of ToSS [tests of statistical significance], such as the neo-Fisherian and error-statistical approaches, when arguing for the superiority of their own positions. (p. 13)

As Hand (1994) has observed, statisticians should focus more on deconstructing different statistical approaches to formulate precisely which question an approach is answering and know which question a researcher wishes to answer. Including range predictions in this deconstruction process will lead to a more interesting discussion when comparing different approaches to statistical inferences.

Unless we examine which questions researchers ask, depending on the goals they have when they perform a study, the phase of the research line, the knowledge that already exists on the topic, and the philosophy of science that researchers subscribe to, it is impossible to draw conclusions about the statistical approach that gives the most useful answer. It may very well be that most researchers cannot precisely formulate the question they want to ask (as most statistical consultants will have experienced). A shift away from the statistician's fallacy and toward teaching people that different statistical approaches answer different questions might push researchers to think more carefully about what it is they want to know.

Conclusion

I believe that pursuing practical alternatives to p values is a form of escapism. Improvements are unlikely to come from telling researchers to calculate a different number but from educating researchers how to ask better questions (see Hand, 1994). Some statisticians have fanatically argued why the alternative statistic they favor (be it confidence intervals, Bayes factors, effect-size estimates, or the false-positive report probability) is what we really want to know. Although these discussions might not reflect the majority viewpoint, they are

extremely visible. However, it is doubtful that there is a single thing anyone wants to know. In certain situations, such as well-controlled experiments in which we want to test ordinal claims, p values can provide an answer to a question of interest. Whenever this is the case, we do not need alternatives to p values—we need correctly used p values.

If we really consider the misinterpretation of p values to be one of the more serious problems affecting the quality of scientific research, we need to seriously reflect on whether we have done enough to prevent misunderstandings. Treating the misinterpretation of p values as a human-factors problem might illuminate ways in which statistics education and statistical software can be improved. We should consider ways in which limitations of null-hypothesis significance testing can be ameliorated with the highest probability of success. Before we dismiss p values, we should examine whether the widespread recommendation to embrace tests of range predictions such as minimum-effect tests and equivalence tests might help to reduce misunderstandings and improve the questions researchers ask. Finally, if we want to know which statistical approach will improve research practices, we need to know which questions researchers want to answer. Polarized discussions about which statistic we should use might have distracted scientists from asking ourselves what it is we actually want to know.

Transparency

Action Editor: Richard Lucas

Editor: Laura A. King

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was funded by Netherlands Organisation for Scientific Research VIDI Grant 452-17-013.

ORCID iD

Daniël Lakens  <https://orcid.org/0000-0002-0247-239X>

Notes

1. I use the term “statistician” to refer broadly to anyone who has weighed in on statistical issues.
2. See the Supplemental Material available online for a more detailed contextual discussion of the quotes used in this article and why I believe they are valid examples of the statistician's fallacy.

References

- Abelson, R. P. (1997a). A retrospective on the significance test ban of 1999 (If there were no significance tests, they

- would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 155–176). New York, NY: Routledge.
- Abelson, R. P. (1997b). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science*, *8*, 12–15. doi:10.1111/j.1467-9280.1997.tb00536.x
- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*, 423–437. doi:10.1037/h0020412
- Batterham, A. M., & Hopkins, W. G. (2006). Making meaningful inferences about magnitudes. *International Journal of Sports Physiology and Performance*, *1*, 50–57. doi:10.1123/ijsp.1.1.50
- 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:10.1016/j.jmp.2015.12.007
- Blume, J. D. (2011). Likelihood and its evidential framework. In P. S. Bandyopadhyay & M. R. Forster (Eds.), *Philosophy of statistics* (pp. 493–511). Amsterdam, The Netherlands: North Holland. doi:10.1016/B978-0-444-51862-0.50014-9
- Blume, J. D., D'Agostino McGowan, L., Dupont, W. D., & Greevy, R. A. (2018). Second-generation p -values: Improved rigor, reproducibility, & transparency in statistical analyses. *PLOS ONE*, *13*(3), Article e0188299. doi:10.1371/journal.pone.0188299
- Burriss, R. P., Troscianko, J., Lovell, P. G., Fulford, A. J. C., Stevens, M., Quigley, R., . . . Rowland, H. M. (2015). Changes in women's facial skin color over the ovulatory cycle are not detectable by the human visual system. *PLOS ONE*, *10*(7), Article e0130093. doi:10.1371/journal.pone.0130093
- Carlsson, R., Danielsson, H., Heene, M., Innes-Ker, A., Lakens, D., Schimmack, U., . . . Weinstein, Y. (2017). Inaugural editorial of meta-psychology. *Meta-Psychology*, *1*(1), Article a1001. doi:10.15626/MP2017.1001
- Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead of “playing the game” it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. *AIMS Neuroscience*, *1*, 4–17. doi:10.3934/Neuroscience.2014.1.4
- Chow, S. L. (1988). Significance test or effect size? *Psychological Bulletin*, *103*, 105–110. <http://doi.org/10.1037/0033-2909.103.1.105>
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003. doi:10.1037/0003-066X.49.12.997
- Cohen, J. (1995). The earth is round ($p < .05$): Rejoinder. *American Psychologist*, *50*(12), 1103. doi:10.1037/0003-066X.50.12.1103
- Colquhoun, D. (2017). The reproducibility of research and the misinterpretation of p -values. *Royal Society Open Science*, *4*(12), Article 171085. doi:10.1098/rsos.171085
- Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, *2*, 161–172. doi:10.1037/1082-989X.2.2.161
- Cox, D. R. (1958). Some problems connected with statistical inference. *Annals of Mathematical Statistics*, *29*, 357–372. doi:10.1214/aoms/1177706618
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29. doi:10.1177/0956797613504966
- Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. London, England: Palgrave Macmillan.
- Dongen, N. N. N., van Doorn, J. B., van Gronau, Q. F., Ravenzwaaij, D., van Hoekstra, R., Haucke, M. N., . . . Wagenmakers, E.-J. (2019). Multiple perspectives on inference for two simple statistical scenarios. *The American Statistician*, *73*, 328–339. doi:10.1080/00031305.2019.1565553
- Fidler, F., & Loftus, G. R. (2009). Why figures with error bars should replace p values: Some conceptual arguments and empirical demonstrations. *Zeitschrift Für Psychologie/ Journal of Psychology*, *217*, 27–37. doi:10.1027/0044-3409.217.1.27
- Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science*, *15*, 119–126. doi:10.1111/j.0963-7214.2004.01502008.x
- Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*, 379–390. doi:10.1037/1082-989X.1.4.379
- Fricke, R. D., Burke, K., Han, X., & Woodall, W. H. (2019). Assessing the statistical analyses used in *Basic and Applied Social Psychology* after their p -value ban. *The American Statistician*, *73*, 374–384. doi:10.1080/00031305.2018.1537892
- Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, *1*, 198–218. doi:10.1177/2515245918771329
- Gill, J. (2018). Comments from the new editor. *Political Analysis*, *26*, 1–2. doi:10.1017/pan.2017.41
- 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:10.1007/s10654-016-0149-3
- Hacking, I. (1965). *Logic of statistical inference*. Cambridge, England: Cambridge University Press.
- Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *The American Psychologist*, *52*, 15–24.
- Haig, B. D. (2017). Tests of statistical significance made sound. *Educational and Psychological Measurement*, *77*, 489–506. doi:10.1177/0013164416667981
- Hand, D. J. (1994). Deconstructing statistical questions. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *157*, 317–356. doi:10.2307/2983526
- Hanson, R. C. (1958). Evidence and procedure characteristics of “reliable” propositions in social science. *American Journal of Sociology*, *63*, 357–370. Retrieved from <https://www.jstor.org/stable/2774136>
- Herrera-Bennett, A. C., Heene, M., Lakens, D., & Ufer, S. (2020). Improving statistical inferences: Can a MOOC

- reduce statistical misconceptions about p -values, confidence intervals, and Bayes factors? *PsyArXiv*. doi:10.31234/osf.io/zt3g9
- Hodges, J. L., Jr., & Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society B: Methodological*, *16*, 261–268. doi:10.1111/j.2517-6161.1954.tb00169.x
- Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*, 1157–1164. doi:10.3758/s13423-013-0572-3
- Hubbard, R. (2019). Will the ASA's efforts to improve statistical practice be successful? Some evidence to the contrary. *The American Statistician*, *73*, 31–35. doi:10.1080/00031305.2018.1497540
- Iverson, G. J., Lee, M. D., & Wagenmakers, E.-J. (2009). P rep misestimates the probability of replication. *Psychonomic Bulletin & Review*, *16*, 424–429. doi:10.3758/PBR.16.2.424
- Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford, England: Oxford University Press.
- Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, *16*, 345–353. doi:10.1111/j.0956-7976.2005.01538.x
- Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746–759. doi:10.1177/0013164496056005002
- Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Bernstein, Š. M. J., Bocian, K., . . . Nosek, B. A. (2014). Investigating variation in replicability: A “Many Labs” replication project. *Social Psychology*, *45*, 142–152. doi:10.1027/1864-9335/a000178
- Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.
- Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses: Sequential analyses. *European Journal of Social Psychology*, *44*, 701–710. doi:10.1002/ejsp.2023
- Lakens, D., & Delacre, M. (2020). Equivalence testing and the second generation p -value. *Meta-Psychology*, *4*, Article MP.2018.933. doi:10.15626/MP.2018.933
- Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving Inferences about null effects with Bayes factors and equivalence tests. *The Journals of Gerontology B*, *75*, 45–57. doi:10.1093/geronb/gby065
- Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*, 259–269. doi:10.1177/2515245918770963
- Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge, England: Cambridge University Press.
- 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:10.1037/0022-006X.46.4.806
- Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*, 108–141. doi:10.1207/s15327965pli0102_1
- Miller, J. (2017). Hypothesis testing in the real world. *Educational and Psychological Measurement*, *77*, 663–672. doi:10.1177/0013164416667984
- Murphy, K. R., & Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. *Journal of Applied Psychology*, *84*, 234–248. doi:10.1037/0021-9010.84.2.234
- Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A: Containing Papers of a Mathematical or Physical Character*, *231*, 289–337. Retrieved from <https://www.jstor.org/stable/91247>
- Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301. doi:10.1037/1082-989X.5.2.241
- Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*, 137–141. doi:10.1027/1864-9335/a000192
- Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement*, *20*, 641–650. doi:10.1177/001316446002000401
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), Article aac4716. doi:10.1126/science.aac4716
- Orben, A., & Lakens, D. (2020). Crud (re)defined. *Advances in Methods and Practices in Psychological Science*, *3*, 238–247. doi:10.1177/2515245920917961
- Parkhurst, D. F. (2001). Statistical significance tests: Equivalence and reverse tests should reduce misinterpretation. *Bioscience*, *51*, 1051–1057. doi:10.1641/0006-3568(2001)051[1051:SSTEAR]2.0.CO;2
- Rogers, J. L., Howard, K. L., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*, 553–565. <http://doi.org/10.1037/0033-2909.113.3.553>
- Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*, 416–428. doi:10.1037/h0042040
- Sainani, K. L. (2018). The problem with “Magnitude-Based Inference.” *Medicine & Science in Sports & Exercise*, *50*, 2166–2176. doi:10.1249/MSS.0000000000001645
- Savalei, V., & Dunn, E. (2015). Is the call to abandon p -values the red herring of the replicability crisis? *Frontiers in Psychology*, *6*, Article 245. doi:10.3389/fpsyg.2015.00245
- Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, *40*, 73–83. doi:10.1037/0003-066X.40.1.73
- Simons, D. J. (2018). Introducing advances in methods and practices in psychological science. *Advances in Methods and Practices in Psychological Science*, *1*, 3–6. doi:10.1177/2515245918757424
- Sotos, A. E. C., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*, *2*, 98–113. doi:10.1016/j.edurev.2007.04.001
- Tendeiro, J. N., & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods*, *24*, 774–795. doi:10.1037/met0000221

- Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*, 1–2. doi:10.1080/01973533.2015.1012991
- Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, *6*, 371–386. doi:10.1037//1082-989X.6.4.371
- van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*, 217–239. doi:10.1037/met0000100
- Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (Eds.). (2019). Statistical inference in the 21st century: A world beyond $p < 0.05$ [Special issue]. *The American Statistician*, *73* (Suppl. 1). <https://www.tandfonline.com/toc/utas20/73/sup1>