The practical alternative to the *p* value is the correctly used *p* value

Daniël Lakens

Eindhoven University of Technology

In Press at Perspectives on Psychological Science

Correspondence can be addressed to Daniël Lakens, Human Technology Interaction Group, ATLAS 9.042, PO Box 513, 5600MB Eindhoven, The Netherlands. E-mail: D.Lakens@tue.nl. This work was funded by VIDI Grant 452-17-013 from the Netherlands Organisation for Scientific Research.

Abstract

Due to 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’, where 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 to include 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 impacts 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 do 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, they collect data with the aim to test hypotheses derived from these models. Statistical inferences can then be used to interpret the data that has 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 these approaches to statistical inferences as valid answers to questions a researcher might be interested in. Especially *p* values are often dismissed as a useful tool to answer 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 tests (NHST) and common misconceptions about *p* values for over 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 et al., 2019). In a Neyman-Pearson framework the goal of statistical tests is to guide the behavior of researchers with respect to a hypothesis. Based on the results of a statistical test, and without ever knowing whether 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, p. 55) “the logic underlying all the statistics you see in the professional journals of psychology”.

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 0 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 et al., 2004; Gill, 2018; Trafimow & Marks, 2015). There is an overwhelming range of proposed alternatives to p values (see the special issue by the American Statistical Association on 'Moving to a World beyond "p < .05"', Wasserstein et al., 2019). Hubbard (2019) reviews how, even though 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 statisticians[1] rarely explicitly state the circumstances in which the use of p values is *not* problematic, and where 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 best known example of the Statistician's Fallacy is provided by Cohen (1994, p. 997) 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 Ho is true?"*

Other statisticians have disagreed with Cohen about what it is we 'want to know'. Colquhoun (2017, p. 2) thinks 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'*. Kirk (1996, p. 754) says "*What we*

[1] I use the term 'statistician' to refer broadly to anyone who has weighed in on statistical issues.

want to know is the size of the difference between A and B and the error associated with our estimate". Blume (2011, p. 509) on the other hand suggests that *'what we really want to know is how likely it is that the observed data are misleading'*. Bayarri, Benjamin, Berger, and Sellke (2016, p. 91) believe that '*we want to know how strong the evidence is, given that we actually observed the value of the test statistic that we did'*. Finally, Mayo (2018, p. 300) argues that '*We want to know what the data say about a conjectured solution to a problem: What erroneous interpretations have been well ruled out?*'. Thus, according to six different (groups of) statisticians, what we 'want to know' is 1) the posterior probability of a hypothesis, 2) the false positive risk, 3) the effect size and its confidence interval, 4) the likelihood, 5) the Bayes factor, or 6) 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) explains that a *p* value does not provide the probability that the null hypothesis is true, but the posterior probability does. Colquhoun (2017) explains that *p* value does not provide the probability that the results have occurred by chance, but the false positive risk does. Kirk (1996) notes how a non-significant *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 all the people included in this category would unanimously agree. Furthermore, it seems highly unlikely 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

[2] A supplement is available in which all quotes are discussed in context, and where I explain why I believe these quotes are valid examples of the Statistician's Fallacy.

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 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 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". 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 has harshly criticized the overreliance of psychology on NHST (Meehl, 1978), he also notes that "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, p. 118) writes that 'Realistically, if the null hypothesis test did not exist, it would have to be (re)invented'. In his book 'Beyond Significance Testing' Kline (2004, p. 86) writes '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.' 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 to establish the direction of an effect, whenever this is question is central to the purpose of the experiment (Cohen, 1995, p. 1103).

When discussing the question a *p* value can answer I will 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'. 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 non-zero correlations, known as the ‘crud factor’ (Meehl, 1990; Orben & Lakens, 2020). Second, the researcher is interested in applying a methodological procedure that allows them to make decisions how to *act*, while controlling error rates. Neyman and Pearson (1933, p. 291) 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.”

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’ the decision to make a certain type of statement or claim (Cox, 1958; Frick, 1996). Based on 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 a researcher makes 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 alpha 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 paper about whether hypotheses are corroborated or not. Rozeboom (1960) criticizes 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.” 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 non-dichotomous inferences using a likelihood ratio, Bayes factor, or a Fisherian interpretation of *p* values. But if you are interested in establishing claims about ordinal effects, distinguishing signal from noise, or if you feel you need to make decisions in research lines based on data, and you want to insure 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 would no longer rely 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 how '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.' I do not know of any other empirical data 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 et al., 2019) which revealed that authors will claim that data supports their prediction with a higher error rate than an alpha level of 5%, leading Fricker and colleagues to conclude "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".

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 until it was severely criticized (Iverson et al., 2009), but it 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 et al., 2018) turn out to highly similar to, but less informative than, equivalence tests (Lakens & Delacre, 2019). 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 to improve statistical inferences than developing novel statistical approaches. As Cohen (1994, p. 1001) concludes: “don't look for a magic alternative to NHST, some other objective mechanical ritual to replace it. It doesn't exist”.

The correct use of established frequentist and Bayesian methods will often lead to similar statistical inferences. In a recent manuscript where we evaluated four null-effects in the gerontology literature with equivalence tests or Bayes factors (Lakens, McLatchie, et al., 2018) both approaches lead to similar inferences in each example. Similarly, four teams of researchers employing frequentist or Bayesian hypothesis testing or estimation independently reached similar conclusions when re-analyzing two studies (Dongen et al., 2019). Although one can always find exceptions if one search long enough, in most cases Bayes factors and *p* values will strongly agree (Tendeiro & Kiers, 2019). Jeffreys, who developed a Bayesian hypothesis test, already noted that when comparing the inferences using his procedure against Frequentist methods proposed by Fisher that: “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 et al., 2007). One exception is research on the mistake to interpret a *p* value larger than 0.05 as evidence for the absence of an effect. A non-significant 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 there is no effect. At best, we can use an equivalence test to examine if 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, et al., 2018). Indeed, Parkhurst (2001) reports the anecdotal observation that the proportion of students who misinterpret $p > 0.05$ as the absence of an effect declined dramatically when students were taught equivalence tests. Research by Fidler and Loftus (2009) show that presenting a figure with confidence intervals alongside the results of a *t*-test reduces the mistake to interpret $p > 0.05$ as the absence of an effect, although confidence intervals themselves are not immune to being misunderstood (Hoekstra et al., 2014).

In our own work we have observed that students in a massive open online course made many errors when attempting to correctly interpret *p* values (Herrera-Bennett et al., 2020). However, a similar amount of errors was 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, with the percentage of correct responses increasing from 8.3 out of 14 to 11.1. This 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 their 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 there has been very little progress (Gigerenzer, 2018). This problematic situation is not unlike something that happened in experimental psychology where problems with publication bias, low power, and inflated alpha 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 et al., 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, while 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 impacts the quality of scientific research, why are we not investing more effort to make sure misunderstandings of *p* values are resolved before young scholars perform their first research project? Although I am sympathetic to statisticians who think all the information researchers need to educate themselves on this topic is already available, as an experimental psychologist who works at a Human-Technology Interaction department this reminds me too much of the engineer who argues all the information to understand 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 papers published in the last 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 nowadays relatively straightforward to create online apps where people can simulate studies and see the behavior of *p* values across studies, which can easily be combined with exercises that fit the knowledge level of bachelor and master 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 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 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 them from interpreting a *p* value larger than 0.05 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 the automatic spell-checker prevents grammatical mistakes? Surely, we are technically able to flag statements such as 'no effect of' in a word document that occur in proximity of a '$p > .05$'? 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? As just one example, a review by van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, 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 large enough samples 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 (where the goal is to reject an effect of exactly 0) with tests of range predictions (where 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 the effect is not just different from zero, but also large enough to be meaningful. Hodges and Lehman (1954) wrote: “About the set H0 we may then distinguish a larger set of H1 of values, representing situations close enough to H0 that the difference is not *materially significant* in the problem at hand” (italics added) and already add “It might be objected that there is nothing novel in the point of view just presented”. Nunnaly (1960) already 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.”. Serlin and Lapsey (1985) discussed the ‘Good-Enough Principle’ where 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”.

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 0 than the smallest effect size of interest, known as a minimum-effect tests (Murphy & Myors, 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, et al., 2018; Rogers et al., 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 in order to attract men. Data from 22 women revealed a statistically significant increase in redness of their facial skin increased during their fertile period. If these authors would have limited their analysis to a null hypothesis test, they would have concluded 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 large 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. Similarly, the use of equivalence tests can prevent misinterpreting a non-significant effect as the absence of a meaningful effect (Lakens, Scheel, et al., 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 whether an ordinal claim can be made or not, they also force researchers to ask more interesting questions. One interesting question that is rarely asked when making a prediction is: "What would falsify my hypothesis?". An important starting point to answer such a question in experimental research is what the smallest effect size of interest would be. Imagine one theory predicts and 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 d = 0.1. We can design a randomized controlled experiment with high statistical power and a low alpha 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 over 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." 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 towards 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 there is a single thing anyone wants to know. In certain situations, such as well-controlled experiments where 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 impacting the quality of scientific research we need to seriously reflect on whether we have done enough to prevent misunderstandings. Treating it 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 reducing 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.

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 *What If There Were No Significance Tests?* (pp. 155–176). Routledge.

Abelson, R. P. (1997b). On the Surprising Longevity of Flogged Horses: Why There Is a Case for the Significance Test. *Psychological Science*, *8*(1), 12–15. https://doi.org/10.1111/j.1467-9280.1997.tb00536.x

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423–437. https://doi.org/10.1037/h0020412

Batterham, A. M., & Hopkins, W. G. (2006). Making Meaningful Inferences About Magnitudes. *International Journal of Sports Physiology and Performance*, *1*(1), 50–57. https://doi.org/10.1123/ijspp.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. https://doi.org/10.1016/j.jmp.2015.12.007

Blume, J. D. (2011). Likelihood and its Evidential Framework. In *Philosophy of Statistics* (pp. 493–511). Elsevier. https://doi.org/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), e0188299. https://doi.org/10.1371/journal.pone.0188299

Burriss, R. P., Troscianko, J., Lovell, P. G., Fulford, A. J. C., Stevens, M., Quigley, R., Payne, J., Saxton, T. K., & 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), e0130093. https://doi.org/10.1371/journal.pone.0130093

Carlsson, R., Danielsson, H., Heene, M., Innes-Ker, A., Lakens, D., Schimmack, U., Schönbrodt, F. D., van Asssen, M., & Weinstein, Y. (2017). Inaugural Editorial of Meta-Psychology. *Meta-Psychology*, *1*, 1–3. https://doi.org/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*(1), 4–17. https://doi.org/10.3934/Neuroscience.2014.1.4

Chow, S. L. (1988). Significance test or effect size? *Psychological Bulletin*, *103*(1), 105–110. http://dx.doi.org/10.1037/0033-2909.103.1.105

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Colquhoun, D. (2017). The reproducibility of research and the misinterpretation of p-values. *Royal Society Open Science*, *4*(12), 171085. https://doi.org/10.1098/rsos.171085

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, *2*(2), 161–172. https://doi.org/10.1037/1082-989X.2.2.161

Cox, D. R. (1958). Some Problems Connected with Statistical Inference. *Annals of Mathematical Statistics*, *29*(2), 357–372. https://doi.org/10.1214/aoms/1177706618

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Palgrave Macmillan.

Dongen, N. N. N. van, Doorn, J. B. van, Gronau, Q. F., Ravenzwaaij, D. van, Hoekstra, R., Haucke, M. N., Lakens, D., Hennig, C., Morey, R. D., Homer, S., Gelman, A., Sprenger, J., & Wagenmakers, E.-J. (2019). Multiple Perspectives on Inference for Two Simple Statistical Scenarios. *The American Statistician*, *73*(sup1), 328–339. https://doi.org/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*(1), 27–37. https://doi.org/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*(2), 119–126. https://doi.org/10.1111/j.0963-7214.2004.01502008.x

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*(4), 379–390. https://doi.org/10.1037/1082-989X.1.4.379

Fricker, 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*(sup1), 374–384. https://doi.org/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*(2), 198–218. https://doi.org/10.1177/2515245918771329

Gill, J. (2018). Comments from the New Editor. *Political Analysis*, *26*(1), 1–2. https://doi.org/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*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Hacking, I. (1965). *Logic of Statistical Inference*. Cambridge University Press.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *The American Psychologist*, *52*(1), 15–24.

Haig, B. D. (2017). Tests of Statistical Significance Made Sound. *Educational and Psychological Measurement*, *77*(3), 489–506. https://doi.org/10.1177/0013164416667981

Hand, D. J. (1994). Deconstructing Statistical Questions. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *157*(3), 317–356. https://doi.org/10.2307/2983526

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?* https://doi.org/10.31234/osf.io/zt3g9

Hodges, J. L., & Lehmann, E. L. (1954). Testing the Approximate Validity of Statistical Hypotheses. *Journal of the Royal Statistical Society. Series B (Methodological)*, *16*(2), 261–268. https://doi.org/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*(5), 1157–1164. https://doi.org/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*(sup1), 31–35. https://doi.org/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*(2), 424–429. https://doi.org/10.3758/PBR.16.2.424

Jeffreys, H. (1939). *Theory of probability* (1st ed). Oxford University Press.

Killeen, P. R. (2005). An Alternative to Null-Hypothesis Significance Tests. *Psychological Science*, *16*(5), 345–353. https://doi.org/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*(5), 746–759. https://doi.org/10.1177/0013164496056005002

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating Variation in Replicability: A "Many Labs" Replication Project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research* (1st ed). American Psychological Association.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses: Sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. https://doi.org/10.1002/ejsp.2023

Lakens, D., & Delacre, M. (2019). Equivalence Testing and the Second Generation P-Value. *Meta-Psychology*. https://doi.org/10.31234/osf.io/7k6ay

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2018). Improving Inferences about Null Effects with Bayes Factors and Equivalence Tests. *The Journals of Gerontology: Series B*. https://doi.org/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*(2), 259–269. https://doi.org/10.1177/2515245918770963

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. 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*(4), 806–834. https://doi.org/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*(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Miller, J. (2017). Hypothesis Testing in the Real World. *Educational and Psychological Measurement*, *77*(4), 663–672. https://doi.org/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*(2), 234–248. https://doi.org/10.1037/0021-9010.84.2.234

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301. https://doi.org/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*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement*, *20*(4), 641–650. https://doi.org/10.1177/001316446002000401

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

Orben, A., & Lakens, D. (2020). Crud (Re)defined. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.31234/osf.io/96dpy

Parkhurst, D. F. (2001). Statistical significance tests: Equivalence and reverse tests should reduce misinterpretation. *Bioscience*, *51*(12), 1051–1057. https://doi.org/10.1641/0006-3568(2001)051[1051:SSTEAR]2.0.CO;2

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*(3), 553–565. http://dx.doi.org/10.1037/0033-2909.113.3.553

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*(5), 416–428. https://doi.org/10.1037/h0042040

Sainani, K. L. (2018). The Problem with "Magnitude-Based Inference." *Medicine & Science in Sports & Exercise*, *Publish Ahead of Print*. https://doi.org/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*. https://doi.org/10.3389/fpsyg.2015.00245

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, *40*(1), 73–83. https://doi.org/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*(1), 3–6. https://doi.org/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*(2), 98–113. https://doi.org/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*. https://doi.org/10.1037/met0000221

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

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminancy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, *6*(4), 371–386. https://doi.org/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*(2), 217–239. https://doi.org/10.1037/met0000100

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond "p < 0.05." *The American Statistician*, *73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913