Review of Philosophy and Psychology
https://doi.org/10.1007/s13164-018-0421-4

Statistical Inference and the Replication Crisis

Lincoln J. Colling[1] · Dénes Szűcs[1]

Published online: 17 November 2018

Abstract
The replication crisis has prompted many to call for statistical reform within the psychological sciences. Here we examine issues within Frequentist statistics that may have led to the replication crisis, and we examine the alternative—Bayesian statistics—that many have suggested as a replacement. The Frequentist approach and the Bayesian approach offer radically different perspectives on evidence and inference with the Frequentist approach prioritising error control and the Bayesian approach offering a formal method for quantifying the relative strength of evidence for hypotheses. We suggest that rather than mere statistical reform, what is needed is a better understanding of the different modes of statistical inference and a better understanding of how statistical inference relates to scientific inference.

1 Introduction

A series of events in the early 2010s, including the publication of Bem's (2011) infamous study on extrasensory perception (or PSI), and data fabrication by Diederik Stapel and others (Stroebe et al. 2012), led some prominent researchers to claim that psychological science was suffering a "crisis of confidence" (Pashler and Wagenmakers 2012). At the same time as these scandals broke, a collective of scientists was formed to undertake a large-scale collaborative attempt to replicate findings published in three prominent psychology journals (Open Science Collaboration 2012). The results of these efforts would strike a further blow to confidence in the field (Yong 2012), and with the replication crisis in full swing old ideas that science was self-correcting seemed to be on shaky ground (Ioannidis 2012).

One of the most commonly cited causes of the replication crisis has been the statistical methods used by scientists, and this has resulted in calls for statistical reform (e.g., Wagenmakers et al. 2011; Dienes 2011; Haig 2016). Specifically, the suite of procedures known as Null Hypothesis Significance Testing (NHST), or simply *significance testing*, and their associated *p* values, and claims of statistical significance, have

✉ Lincoln J. Colling
ljc65@cam.ac.uk

[1] Department of Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, UK

come in most to blame (Nuzzo 2014). The controversy surrounding significance testing and p values is not new (see Nickerson 2000 for a detailed treatment); however, the replication crisis has resulted in renewed interest in the conceptual foundations of significance testing and renewed criticism of the procedures themselves (e.g., Wagenmakers 2007; Dienes 2011; Szűcs and Ioannidis 2017a). Some journals have gone so far as to ban p values from their pages (Trafimow and Marks 2014) while others have suggested that what gets to be called *statistically significant* should be redefined (Benjamin et al. 2017). Some criticism of p values stems from the nature of p values themselves—a position particularly common with those advocating some form of Bayesian statistics—while other criticisms have focused on their use rather than attacking the conceptual grounding of the procedures themselves (Nickerson 2000; García-Pérez 2016). However, one thing that was made clear by the replication crisis, and the ensuing debates about the use of p values, is that few people understood the nature of p values, the basis of the Frequentist statistics that generate them, and what inferences could be warranted on the basis of *statistical significance*. Such was the confusion and misunderstanding among many in the scientific community that the American Statistical Association (ASA) took the unusual step of releasing a statement on statistical significance and p values in the hope of providing some clarity about their meaning and use (Wasserstein and Lazar 2016).

In order to make sense of the criticisms of p values and to make sense of their role in the replication crisis it is important to understand what a p value is (how it is derived) and what conditions underwrite its inferential warrant. We detail this in Section 2. There we also outline what inferences can be made on the basis of p values and introduce a recent framework, the *error statistical approach*, which addresses some of the shortcomings of previous Frequentist approaches. In Section 3 we introduce an alternative to Frequentist statistics—Bayesian statistics. Specifically, in Section 3.1 we examine some of the claimed benefits of the Bayesian approach while in Section 3.2 we introduce the Bayesian notion of statistical evidence, and examine whether the Bayesian approach and the Frequentist approach lead to different conclusions. In Section 4 we compare the two approaches more directly and examine how each approach fits into a system of scientific inference. Finally, we conclude by suggesting that rather than mere *statistical reform* what is needed is a change in how we make scientific inferences from data. And we suggest that there might be benefits in pragmatic pluralism in statistical inference.

2 Frequentist Statistic and p Values

The ASA statement on p values provides an informal definition of a p value as "the probability *under a specified statistical model* that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be *equal to or more extreme* than its observed value" (Wasserstein and Lazar 2016, our emphasis). Probability is an ill-defined concept with no generally agreed definition that meets all the requirements that one would want. In the context of significance testing, however, p values are often interpreted with reference to the long run behaviour of the test procedure (e.g., see Neyman and Pearson 1933). That is, they can be given a *frequency* interpretation (see Morey et al. 2016a for more detail on a frequency interpretation of confidence intervals). Although a frequency interpretation may not be universally accepted (or acknowledged), this interpretation more clearly highlights the link between p values and the long run behaviour of significance tests. When given a frequency interpretation, the p indicates how often under a specified model, considering repeated experiments, a test statistic as large or larger than the one observed would be observed if it was the case that the null hypothesis (for example, the hypothesis that the two groups are drawn from the same population) was true. The p value is calculated from the *sampling distribution*, which describes what is to be expected *over the long run* when samples are tested.

What allows one to draw inferences from p values is the fact that statistical tests should rarely produce small p values if the null model is true, and provided certain conditions are met.[1] It is also this fact that leads to confusion. Specifically, it leads to the confusion that if a small p is obtained then one can be 1 - p sure that the alternative hypothesis is true. This common misunderstanding can result in an interpretation that, for example, $p = 0.01$ indicates a 99% probability that the detected effect is real. However, to conclude this would be to confuse the probability of obtaining the data (or more extreme) given the null hypothesis with the probability that the null hypothesis is true given the data (see Nickerson 2000 for examples of this confusion).

The confusion that p values warrant inferences they do not has similarly led to confusion about the conditions under which p values *do* warrant inferences. We will explain what inferences p values do warrant in Section 2.3, but before this can be done it is important to understand what conditions must be met before they can support *any* inferences at all. For now, however, it is sufficient to know that inferences on the basis of p values rely on the notion of *error control.* As we will see, violations of the conditions that grant these error control properties may be common.

2.1 Controlling False Positives

The first condition under which p values are able to provide information on which to base inferences is that *if* the null hypothesis is true then p values should be uniformly distributed.[2] For instance, if one was to repeatedly draw samples from a standard normal distribution centred on 0, and after each sample test the null hypothesis that $\mu = 0$ (for example, by using a one sample t-test) one would obtain a distribution of p values approximately like the one shown in Fig. 1(a). This fact appears to contradict at least one common misinterpretation of p values, specifically the expectation that routinely obtaining high p values should be common when the null hypothesis is true—for instance the belief that obtaining $p > .90$ should be common when the null is true and $p < .10$ should be rare, when in fact they will occur with equal frequency (see Nickerson 2000 for common misinterpretations of p values). Herein lies the concept of

[1] These conditions are the *assumptions* of the statistical test. These might include things such as equal variance between the two groups in the case of t tests or certain assumptions about the covariance matrix in the case of factorial ANOVA. These are often violated and, therefore, tests can be inaccurate. Correction procedures, tests that are robust to violations, or tests that generate their own sampling distribution from the data (such as randomisation tests) are available. However, we will not discuss these as our focus will primarily be on the *inferences* that statistical tests support.

[2] We should note that this is only generally true when the null model takes the form of a continuous probability distribution, which is common for the statistical procedures used in psychology. This assumption does not necessarily hold for discrete probability distributions.

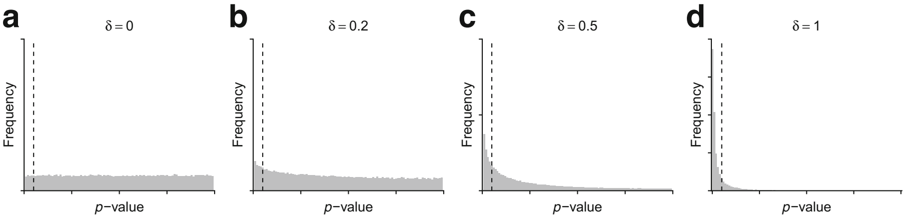

Fig. 1 Examples of p value distributions under different effect sizes. An effect size of $\delta = 0$ indicates that the null hypothesis is true

the *significance threshold*. While, for instance, $p \approx .87$, and $p \approx .02$ will occur with equal frequency if the null is true, p values less than the threshold (defined as α) will only occur with the frequency defined by that threshold. Provided this condition is met, this sets an *upper bound* on how often one will incorrectly infer the presence of an effect when in fact the null is true.

The uniformity of the p value distribution under the null hypothesis is, however, only an ideal. In reality, there are many behaviours that researchers can engage in that can change this distribution. These behaviours, which have been labelled p hacking, QPRs (questionable research practices), data dredging, and significance chasing, therefore threaten to revoke the p value's inferential licence[3] (e.g., Ware and Munafò 2015; Simmons et al. 2011; Szűcs 2016). One of the most common behaviours is optional stopping (also known as data peaking). To illustrate this behaviour, we will introduce an example, which we will return to later in the context of Bayesian alternatives to significance testing. Consider Alice who collects a sample of 10 observations. After collecting her sample, she conducts a significance test to determine whether the mean is significantly different from some null value (this need not be zero, but often this is the case). Upon finding $p = .10$, she decides to add more observations checking after adding each additional observation whether $p < .05$. Eventually, this occurs after she has collected a sample of 20.

On a misunderstanding of the p value this behaviour seems innocuous, so much so that people often express surprise when they are told it is forbidden (e.g., John et al. 2012; Yu et al. 2013). However, it only seems innocuous on the incorrect assumption that large p values should be common if the null is true. After all, Alice checked her p values after every few samples, and while they may have changed as each new sample was added, they weren't *routinely* large. However, optional stopping distorts the distribution of p + values so that it is no longer uniform. Specifically, the probability of obtaining $p < \alpha$, when optional stopping is applied, is no longer equal to α and instead it can be dramatically higher than α.[4] Thus, in the case of optional stopping, the connection between the value of p and the frequency of obtaining a p value of that magnitude (or smaller) is broken.

[3] Concerns over these behaviours is not new. Referring to the practice as "cooking", Charles Babbage (1830) noted that one must be very unlucky if one is unable to select only agreeable observations out of the multitude that one has collected.

[4] To illustrate this, we conducted a simple simulation. We drew samples ($n = 1000$) from a standard normal distribution centred at zero. The values were then tested, using a one sample t-test against the null hypothesis that $\mu = 0$ by first testing the first 10 values, then the first 11, the first 12 and so forth until either obtaining a $p > 0.05$ or exhausting the 1000 samples. After repeating this procedure 10,000 times, we were able to obtain a significant p value approximately 46% of the time. The median sample size for a significant result was 56.

A related issue that can revoke the inferential licence of p values occurs when a researcher treats a collection of p values (also known as a family of tests) in the same way they might treat a single p value. Consider the case where a researcher runs ten independent statistical tests. Given the null, the frequency of finding a significant result ($p < 0.05$) is 5% *for each test*. As a result, the chance of finding *at least one* significant effect in a family of 10 tests is approximately 40%. While most researchers understand the problem of confusing the *chance of finding a significant test* with the *chance of finding at least one significant test in a collection of tests*, in the context of simple tests like *t*-tests, this confusion persists in more complex situations like factorial ANOVA. Consider a two factor ANOVA, which produces three test statistics: Researchers can make this error and confuse the chance of finding *at least one* significant test (for example, a main effect or interaction) with the chance of *a particular* test being significant. In the case of the former, the chance of finding at least one significant main effect or interaction in a two factor ANOVA can be approximately 14%. That a recent survey of the literature, which accompanied a paper pointing out this hidden multiplicity, found that only around 1% of researchers (across 819 papers in six leading psychology journals) took this into account when interpreting their data demonstrates how widespread this confusion is (Cramer et al. 2015). Furthermore, high profile researchers have expressed surprise upon finding this out (Bishop 2014), further suggesting that it was not commonly known. As noted by Bishop (2014), this problem might be particularly acute in fields like event-related potential (ERP) research where researchers regularly analyse their data using large factorial ANOVAs and then interpret whatever results fall out. As many as four factors are not uncommon, and consequently, the chance of finding at least one significant effect can be roughly the same as correctly calling a coin flip. Furthermore, if a theory can be supported by a main effect of one factor, or any interaction involving that factor—that is, if one *substantive hypothesis* can be supported by multiple *statistical hypotheses*—then in the case of a four-way ANOVA that theory will find support as often as 25% of the time even if the null hypothesis is true.

With this in mind, the advice offered by Bem (2009) appears particularly unwise: In speaking about data that might have been collected from an experiment, he suggests "[e]xamine them from every angle. Analyze the sexes separately. Make up new composite indices." (pp. 4–5). That is, add additional factors to the ANOVA to see if anything pops up. However, as we have seen, adding additional factors simply increases the chance of significance even when the null is true. This hidden multiplicity is rarely acknowledged in scientific papers. More generally, any *data dependent* decisions—for example, choosing one composite index over another based on the data—greatly increases the chance of finding significance regardless of whether multiple comparisons *were actually made*.[5] Indeed, Bem (2009 p 6) goes on to state that:

> "Scientific integrity does not require you to lead your readers through all your wrongheaded hunches only to show—voila!—they were wrongheaded. A journal article should not be a personal history of your stillborn thoughts."

[5] In addition to specific data dependent decisions, Steegen et al. (2016) outline how a number of seemingly arbitrary decisions made during the analysis process can give rise to a very wide range of results.

While such a journal article may make for tedious reading, it is only by including all those thoughts, those wrongheaded hunches, those data dependent decisions, that will allow the reader to determine whether the process by which the results were obtained deserve to be awarded any credibility, or whether they are as impressive as correctly calling a coin flip.

2.2 Controlling False Negatives

A second condition that must be met for inferences on the basis of p values to be warranted is that low p values (i.e., $p < \alpha$) should occur more *frequently* when a true or real effect is present. This occurs because when the discrepancy between the null model and the model from which the samples are actually drawn increases (something that can be quantified in terms of *effect size*), the distribution of p values, obtained in the long run, departs from uniformity. This is illustrated in Fig. 1(b–d) by showing the distribution of p values obtained from repeated testing of samples drawn from distributions representing different true effect sizes. When a real effect is present, the frequency with which a p value occurs is inversely proportional to its magnitude. This skewing of the p value distribution in the presence of a real effect illustrates the concept of *statistical power* (e.g., Cohen 1992). The greater the skew observed in the long run distribution of p values the greater the statistical power. That is, power is equal to $1 - \beta$, where β is the proportion of p values $> \alpha$ that occur when a true effect is present. Power, therefore, allows one to place an *upper bound* on how often one will incorrectly conclude the *absence* of an effect (of at least a particular magnitude) when in fact an effect (of that magnitude or greater) is present.

That p values skew towards zero in the presence of a true effect implies that p values near the threshold α should be comparatively rare if a real effect is present. However, near threshold p values are surprisingly common (Masicampo and Lalande 2012; de Winter and Dodou 2015). This suggests that the reported effects may actually accord more with a true null hypothesis. However, they may also imply that statistical power is very low and that the distribution of p values has not departed sufficiently from uniformity. Adequate statistical power—that is, the requirement that experiments are so designed such that in the long run they will produce an extremely skewed distribution of p values—is a fundamental requirement if inferences are to be drawn on the basis of p values. However, empirical studies of the scientific literature suggest that this requirement is not routinely met. For example, studies by Button et al. (2013) and Szűcs and Ioannidis (2017b) suggest that studies with low statistical power are common in the literature. Recall, it is only when the two conditions are met—uniformly distributed p values when the null is true and a heavily skewed p value distribution when a real effect is present—that good inferences on the basis of p values are possible. Neither of these conditions are commonly met and, therefore, the epistemic value of p values is routinely undermined.

What is the cause of low statistical power? In our definition of power, we said that power was determined by the skew of the p value distribution in the presence of a *given* true effect. That is, if samples of a fixed size are repeatedly drawn and tested with a statistical test, and a true effect is present, how often $p < .05$ occurs depends on the *magnitude* of the true effect. To draw valid inferences from p values, in the long run, one needs to know the magnitude of the effect that one is making inferences about. If the magnitude of the effect is small, then one needs more information (larger samples) to reliably detect its presence. When the magnitude of the effect is large, then you can generate reliable decisions using less information (smaller samples). However, it is important to note that basing effect size estimates for a priori power analyses on published results can be very problematic because in the presence of publication bias (only publishing significant results) the published literature will invariably overestimate the real magnitude of the effect. That is, when power is low, statistical significance acts to select only those studies that report effect sizes larger than the true effect. Only through averaging together significant and non-significant effects can one get a good estimate of the actual effect size. Interestingly, an examination of replication attempts by Simonsohn (2015) suggests that in many cases, effect size estimates obtained from high-powered replications imply that the original studies reporting those effects were underpowered and, therefore, could not have reliably studied effects of those magnitudes.

2.3 Frequentist Inferences

Inferences on the basis of p values can be difficult and unintuitive. The problems that we've outlined above are not problems of significance testing per se, rather they are a result of the inferential heuristics that people apply when conducting experiments—heuristics such as, "if it's nearly significant then collect more data" or "if I can obtain significance with a small sample then it's more likely that my hypothesis is true". Part of the reason why people may employ inferential heuristics is that several distinct frameworks exist for drawing inferences on the basis of p values and often these are not clearly distinguished in the statistics textbooks or statistics training. In some cases, researchers may even be unaware that different frameworks exist. The two most prominent frameworks are those of Fisher (e.g., Fisher 1925) and Neyman and Pearson (e.g., Neyman and Pearson 1933). Fisher's view of inference was simply that data must be given an opportunity to disprove (that is, reject or falsify) the null hypothesis (H_0). The innovation of Neyman and Pearson was to introduce the alternative hypothesis (H_1) and with it the concept of false alarms (*errors of the first type*, or inferring the presence of an effect when the null hypothesis is true) and false negatives (*errors of the second type*, or inferring the absence of an effect when the alternative hypothesis is true). They also saw a different role for the p value. Fisher was concerned with the actual magnitude of the p value. Neyman and Pearson, on the other hand, were concerned with whether the p value crossed a threshold (α). If the p value was smaller than α then one could reject H_0 and if the p value was greater than α one could fail to reject H_0.[6] By fixing α and β (that is, by maximising statistical power) at particular levels they could fix the long run error control properties of statistical tests, resulting in rules that, if followed, would lead to inferences that would rarely be wrong. The type of inferences employed in practice, however, appear

[6] Neyman and Pearson (1933) use the terminology *accept* H_0. However, Neyman (1976) uses the terminology *do not reject* H_0. Furthermore, he goes on to state that his preferred terminology is *no evidence against* H_0. We follow Neyman (1976) in preferring the *no evidence against* or *do not reject* phrasing.

in many ways to be a hybrid of the two views (Gigerenzer 1993). A consequence of this is that many of the inferences drawn from significance tests have been muddled and inconsistent.

As a result, some have argued that significance tests need a clearer inferential grounding. One such suggestion has been put forward by Mayo (Mayo 1996; Mayo and Spanos 2006; Mayo and Spanos 2011) in the form of her *error-statistical* philosophy. As the name suggests, it builds on the insight of Neyman and Pearson that Frequentist inference relies on the long run error probabilities of statistical tests. In particular, it argues that for inferences on the basis of p values to be valid (that is, have good long run performance) a researcher cannot simply draw inferences between a null (e.g., no difference) and an alternative which is simply its negation (e.g., a difference). Long run performance of significance tests can only be controlled when inferences are with reference to *a specific alternative hypothesis*. And inferences about these specific alternatives are only well justified if they have passed *severe* tests.

Mayo and Spanos (2011) explains *severity* informally with reference to a math test as a test of a student's math ability. The math test counts as a *severe* test of a student's math ability if it is the case that obtaining a high score would be unlikely unless it was the case that the student actually had a high maths ability. Severity is thus a function of a specific test (the math test), a specific observation (the student's score), and a specific inference (that the student is very good at maths).

More formally, Mayo and Spanos (2011) state the *severity principle* as follows:

> Data x_0 (produced by process G) do not provide good evidence for the hypothesis H if x_0 results from a test procedure with a very low probability or capacity of having uncovered the falsity of H, even if H is incorrect.

Or put positively:

> Data x_0 (produced by process G) provide good evidence for hypothesis H (just) to the extent that test T has severely passed H with x_0.

Severity is, therefore, a property of a specific test with respect to a specific inference (or hypothesis) and some data. It can be assessed qualitatively, as in the math test example above, or quantitatively through the sampling distribution of the test statistic. To illustrate how this works in practice, we can consider the following example (adapted from Mayo and Morey 2017). A researcher is interested in knowing whether the IQ scores of some group are above average. According to the null model, IQ scores are normally distributed with a mean of 100 and a *known* variance of 15^2. After collecting 100 scores ($n = 100$), she tests the sample against the null hypothesis $H_0 : \mu = 100$ with the alternative hypothesis $H_1 : \mu > 100$. If the observed mean ($\bar{x}$) was 103, then a z-test would be significant at $\alpha = .025$. From something like a classic Neyman-Pearson approach, the inference that would be warranted on the basis of this observation would be something like *reject* H_0 and conclude that the mean is greater than 100.

A *severity* assessment, however, allows one to go further. Instead of merely concluding that the group's mean (μ_1) is greater than the null value (μ_0), one can instead use the observed result ($\bar{x}$) to assess specific alternate inferences about discrepancies (γ) from μ_0 of the form $H_1 : \mu > \mu_0 + \gamma$. For example, one might want to use the observation

($\bar{x} = 103$) to assess the hypothesis $H_1 : \mu > \mu_0 + 1$ or the hypothesis $H_1 : \mu > \mu_0 + 3$. The severity associated with the inference $\mu > 101$ would be 0.91,[7] while the severity associated with the inference that $\mu > 103$ is 0.5. Thus, according to the severity principle, the observation that $\bar{x} = 103$ provides us with better grounds to infer that μ_1 is at least 101 relative to an inference that it is at least 103.

Just as one can use severity to test different inferences with respect to a fixed result, one can also use severity to assess a fixed inference with respect to different results. Consider again the inference that $\mu > 103$. The severity associated with this inference and the result $\bar{x} = 103$ is 0.5. However, if one had observed a different result of, for example, $\bar{x} = 105$, then the severity associated with the inference $\mu > 103$ would be 0.91. In order to visualise severity for a range of inferences with reference to a particular test and a particular observation, it is possible to plot severity as a function of the inference. Examples of different inferences about μ for different observations ($\bar{x}$) is shown in Fig. 2(a).

The severity assessment of significant tests has a number of important properties. First, severity assessments license different inferences on the basis of different observed results. Consequently, rather than all statistically significant results being treated as equal, specific inferences may be more or less well justified on the basis of the specific p value obtained. In our above example, the observation of $\bar{x} = 103$ ($n = 100$, $\sigma = 15$) results in $p = .023$, while the observation of $\bar{x} = 105$ results in $p < 0.001$. Thus for a fixed n, lower p values license inferences about larger discrepancies from the null. The severity assessment also highlights the distinction between statistical hypotheses and substantive scientific hypotheses. For example, a test of a scientific hypothesis might require that the data support inferences about some deviation from the null value that is at least of magnitude X. The data might reach the threshold for statistical significance without the inference that $\mu_1 > \mu_0 + X$ passing with high severity. Thus, the statistical hypothesis might find support without the theory being supported.

Severity assessments can also guard against unwarranted inferences in cases where the sample size is very large. Consider the case where one fixes the observed p value (for example, to be just barely significant) and varies the sample size. What inferences can be drawn from these *just* significant findings at these various sample sizes? On a simplistic account, all these significant tests warrant the inference *reject* H_0 and conclude some deviation (of unspecified magnitude) from the null. A severity assessment, however, allows for a more nuanced inference. As sample size increases, one would only be permitted to infer smaller and smaller discrepancies from the null with high severity. Again using our example above, the observation associated with $p = .025$ and $n = 100$, allows one to infer that $\mu_1 > 101$ with a severity of 0.9. However, the same p value obtained with $n = 500$ reduces the severity of the same inference to 0.68. An illustration of the influence of sample size on severity is shown in Fig. 2(b). If one wanted to keep the severity assessment high, one would need to change one's inference to, example, $\mu > 100.5$ (which would now be associated with a severity of 0.89). Or if one wanted to keep the same inference (because that inference is required by the

[7] In the R statistics package, severity for a z-test can be calculated using the command, pnorm(x.bar - (h0 + gamma)/ (sigma / sqrt(n))), where x.bar is the observed mean, h0 is the null value, sigma is the population standard deviation, n is the sample size, and gamma is the deviation from the null value that one wishes to draw an inference about.

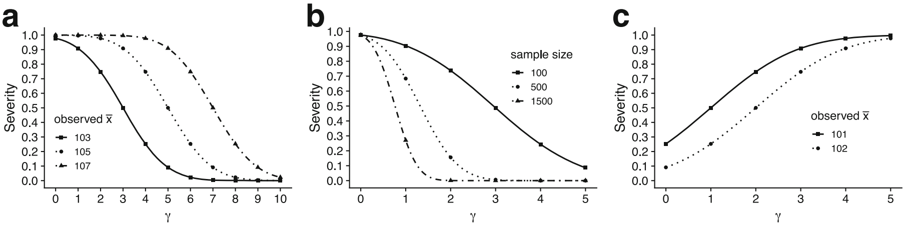

Fig. 2 Examples of severity curves for different statistically significant observations (**a**), barely significant observations with different sample sizes (**b**), and different non-significant observations (**c**)

scientific theory or some background knowledge) at the same severity then one would need to observe a far lower p value before this could occur.[8]

Severity assessments also allow one to draw conclusions about non-significant tests. For instance, when one *fails to reject* H_0, it is possible to ask what specific inferences are warranted on the basis of the observed result. Once again using the IQ testing example above, but with a non-significant observation ($\bar{x} = 102$, $n = 100$, $\sigma = 15$), one can ask what inferences about μ are warranted. For example, one might ask whether an inference that $\mu_1 < 105$ is warranted or whether the inference that $\mu_1 < 103$ is warranted. The severity values associated with each of these inferences (and the observed result) are 0.98 and 0.75, respectively. Therefore, one would have good grounds for inferring that the discrepancy from the null is less than 5, but not good grounds for inferring that it is less than 3. An illustration of severity curves for non-significant observations is shown in Fig. 2(c).

The two examples outlined above are both cases which involve inferences from a single test. But as Mayo (1996) notes, a "procedure of inquiry… may include several tests taken together". The use of multiple tests to probe hypotheses with respect to data may be particularly useful in the case where one has failed to reject the null hypothesis. While it is usual to think of significance testing in terms of a null of no effect and an alternative as departures from this null, any value can be designated the null. For example, one might want to test the null hypotheses $H_0 : \mu \leq B$ and $H_0 : \mu \geq A$ (where usually $B = -A$) as a way to examine whether the data support inferences that μ lies *within* specified bounds (what can be termed *practical equivalence*, see Phillips 1990; Lakens 2017). This procedure can supplement, or be used as an alternative, to severity interpretations so that one can determine precisely what inferences are warranted on the basis of the data. A consequence of this is that Frequentist inference need not come down to a simple binary (for example, *reject* H_0, *fail to reject* H_0/*accept* H_1). Instead, a set of data might lead a researcher to form a far wider range of conclusions. These may include (but are not limited to) inferring: some deviation is present but it is not of sufficient magnitude to support the theory; there are no grounds for inferring that a deviation is present, but neither are there good grounds for inferring any effect lies only

[8] This final suggestion can take the form of calibrating ones α level with reference to the sample size and the effect of (scientific) interest. Typically, however, researchers tend to use a fixed α regardless of context, although recently some have begun to suggest that a single α level may not be appropriate for all contexts (Lakens et al. 2018).

within a narrowly circumscribed region around the null; and, there are good grounds for inferring the presence of a deviation from the null and that the deviation is of sufficient magnitude to support a theory.

We will return to Frequentist inference later. For now, one important point to note is that this kind of Frequentist inference is piecemeal. Claims that are more *severely* tested are given more weight than those claims that are not severely tested. Importantly, severe testing might require more than one statistical test—for example, to test assumptions or to break down a hypothesis into multiple piecemeal statistical hypotheses. The severity principle also encourages replication because having to pass multiple tests is a more severe requirement. Activities such as *p*-hacking, optional stopping, or small samples sizes, all directly affect severity assessments by directly changing the error probabilities of the tests. Unfortunately, error statistical thinking has not been common in the psychological literature. However, its value is now starting to be recognised by some (e.g., Haig 2016), including some working within Bayesian statistics (e.g., Gelman and Shalizi 2013). Although some of the finer details of the error statistical approach are still to be worked out it may provide a good guide for thinking about how to interpret statistical tests.

3 An Alternative to *p* Values

In the preceding section, we showed that the grounds on which *p* values are granted their epistemic licence are easily violated; however, it has also been argued that *p* values are simply not suitable for scientific inferences because they don't provide the information scientists *really want to know* (e.g., see Nickerson 2000; Lindley 2000). On this view, what scientists really want to know is the probability that their hypothesis is true given their data—that is, they want to assign some credence to their hypothesis on the basis of some data they have obtained. Furthermore, *p*-hacking, optional stopping, and similar practices demonstrate the need for procedures that are somehow immune to these behaviours. This alternative, it is claimed, is provided by Bayesian statistics(Wagenmakers 2007; Dienes 2011; Morey et al. 2016b).[9] Bayesian statistics offers a radically different approach to statistical inference, and while largely a niche area in the psychological literature in past decades, events like the replication crisis have sparked renewed interest in these methods.

In offering a solution to what he terms the "pervasive problem of *p* values", Wagenmakers (2007) suggests that Bayesian statistics has the desirable attributes for the ideal statistical procedure. These include: 1) that they are dependent only on the observed data, and not the data that might have been collected, 2) that they are immune to the unknown intentions of the researcher, and 3) that they provide a measure of the *strength of evidence* that takes into account both the null and the alternative. Much of the discourse surrounding the switch to Bayesian statistics has focused particularly on

[9] In this section, we will use "Bayesian statistics" as a shorthand for a suite of approaches that include, but are not limited to, techniques for computing Bayes factors and approaches for estimating the values of unknown parameters. Bayesian statistics should not be taken to mean any procedure that makes use of Bayes Theorem. Bayes Theorem is simply derived from the rules of conditional probabilities. Bayesian statistics, however, is the approach to statistics that aims to produce outputs in the form of degrees of belief and/or degrees of support rather than supporting inferences by controlling certain kinds of errors.

the idea that Bayesian statistics may be the solution to problems caused by optional stopping, which have arguably contributed significantly to the replication crisis (e.g. Wagenmakers 2007; Rouder 2014). Others, however, have also focused on notions of *evidence* suggesting that the Bayesian conception of strength of evidence is more amenable to scientific reasoning or that it is closer to what researchers intuitively require (e.g., Morey et al. 2016b; Szűcs and Ioannidis 2017a). It is worth unpacking the claimed advantages of Bayesian statistics in more detail. We will examine the basis of these claims in the sections below.

3.1 Evidence Derived from Data Alone

In order to unpack the claim that Bayesian inferences are dependent only on the observed data and not data that might have been collected, but wasn't, it is necessary to understand how Frequentist statistics fall into this trap. This Bayesian critique of Frequentist statistics is based on the fact that Frequentist p values are calculated from the *sampling distribution*. As outlined earlier, the sampling distribution is a probability distribution of the values of the test statistic under a specified model, such as the null model. It includes all the values that the test statistic might take. And the p value is calculated from the tail end probabilities of this distribution—that is, the p value expresses: How often would I obtain a value this large *or larger* under this statistical model.

Given this, it is trivial to construct two statistical models (sampling distributions) where the probability of observing a *specific* value of the test statistic is the same, but the chance of observing other values (specifically, larger values) is different. Once a specific value is observed, and a p value is calculated, it will be different depending on the probability of obtaining larger values even though the two statistical models *say the same thing* about the observed data. As Jeffreys (1961) put it, the use of "p implies… that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred."

The second desirable property of Bayesian statistics is that, unlike p values, Bayesian statistics are not dependent on the unknown *intentions*[10] of the researcher. Consider again the case of Alice in the description of the uniformity assumption of the p value distribution. Alice collected data from 10 participants, did a significance test and found $p > .05$, added another 10 participants, re-running the test after every participant and then eventually found $p < .05$. Contrast this with Ashanti, who obtained a sample of 20 participants, ran a significance test and found $p < .05$. The Frequentist would say that Alice and Ashanti cannot draw the same inferences on the basis of their data, because the severity assessment of Alice and Ashanti's inferences would differ. As Wagenmakers (2007) states, examples like this "forcefully [demonstrate] that *within the context of NHST* [null hypothesis significance testing] it is crucial to take the sampling plan of the researcher into account" (p. 786). Furthermore, he goes on to state that within the context of Bayesian statistics the feeling people have that "optional

[10] The word *intentions* is often used in the literature. However, it is not the researcher's *intentions* that have an influence on the interpretations of p values. Rather, it is researchers' *behaviour* that influences the interpretation of p values.

stopping” amounts to “cheating” and that no statistical procedure is immune to this is “contradicted by a mathematical analysis”. The claim here is that Bayesian statistics are immune to optional stopping and that collecting more data until the patterns are clear is warranted if researchers are using Bayesian statistics.

3.2 Bayesian Statistics and a Measure of Strength of Evidence

These first two properties of Bayesian statistics, of the immunity to intentions, and of being dependent only on the collected data and not any other data, are derived from what is called the *Likelihood Principle*. The concept of the likelihood allows us to understand the third property of Bayesian statistics, namely that they give us a measure of the strength of evidence. To see this, it is important to know what is meant by *evidence*. As stated in a recent primer for psychological scientists, “The Likelihood Principle states that the likelihood function contains all of the information relevant to the evaluation of *statistical evidence*. Other facets of the data that do not factor into the likelihood function (e.g., the cost of collecting each observation or the stopping rule used when collecting the data) are irrelevant to the evaluation of *the strength of the statistical evidence*” (Etz 2017, our emphasis). The intuition here is obvious, if you want to know whether some data supports model A or model B, all you need to know is whether the data are more likely under model A or model B. On this view, the strength of evidence is just in the ratio of the likelihoods. If the observed data are three times more likely under model A than model B, then this can be read as a *measure* of the strength of evidence. Furthermore, if model A is the null model, then we can say something about the evidential support for this null.

A measure of the strength of evidence is meant to have an additional benefit for the Bayesian. We can weigh our evidence according to some background *pre-data* beliefs we have (e.g., that Model A is very unlikely to be true) and then use the data to update our beliefs. In Bayesian hypothesis testing, this updating factor is called a *Bayes factor*. Numerically, the Bayes factor can be interpreted as an odds ratio, and it is calculated as the ratio of two *marginal likelihoods* where the marginal likelihood is comprised of a model of the data and some predictions about likely parameter values (sometimes referred to as a *prior* (e.g., Rouder et al. 2009) or a *model of the hypothesis* (e.g., Dienes and Mclatchie 2017)). Rouder et al. (2009) give the marginal likelihood for hypothesis H as:

$$M_H = \int_{\theta \in \Theta_H} f_H(\theta; \mathbf{y}) p_H(\theta) d\theta,$$

where Θ_H represents the parameter space under the hypothesis H, f_H represents the probability density function of the data under the hypothesis H, and p_H represents the prior distribution of the parameter values expected by that hypothesis. The important point to note here is that calculating a Bayes factor requires the analyst to stipulate some prior probability function for the parameter that they wish to draw inferences about under each of the models they are comparing.

It is worth stepping through this in more detail to understand how this calculation works. To do so, we will consider the task of trying to determine whether a coin is fair (this example, makes use of discrete rather than continuous probability distributions and therefore the integral can be replaced by a sum). For this example, one might define the null hypothesis as $H_0 : \theta = 0.5$, or that the probability of obtaining heads is 50%. In order to calculate a *Bayes factor*, one needs another hypothesis. We might define this hypothesis as the probability of obtaining heads being some other fixed value—for example, $H_1 : \theta = 0.7$, or that the probability of obtaining heads is 70%. If we were to further consider H_0 and H_1 equally plausible, our Bayes factor value would simply be the likelihood ratio of these two hypotheses. For example, given a set of data such as the observation of 2 heads out of 10 flips we could conclude that this observation is 30.38 times more probable under the hypothesis that $\theta = 0.5$ than the hypothesis that $\theta = 0.7$.

However, we are ordinarily not interested in a single parameter value but are instead concerned with models in which the parameter may take one of several different values. In our coin flipping example, this might mean comparing $H_0 : \theta = 0.5$ and an alternative hypothesis H_1 composed of 11 point hypotheses ($H : \theta = 0$, $H : \theta = 0.1$, $H : \theta = 0.2$, … $H : \theta = 1$) spanning the entire range of values that θ might take. To calculate the *Bayes factor*, we first calculate the likelihood ratio of the data under H_0 to each of the 11 point hypotheses of H_1. The *Bayes factor* is then computed as the *weighted sum* of these 11 values, where the weights are determined by a *prior* assigned to each of the 11 point hypotheses that make up H_1. The prior makes predictions about what parameter values (bias values in our example) are expected under H_1.[11] If for example, we were to consider each possible value of θ to be equally likely under our biased coin model, then we would weigh each likelihood ratio equally. Because the prior is a probability distribution, the weights should sum to one, which means that each likelihood ratio would have a weight of 1/11. For our example of observing 2 heads in 10 flips this would correspond to a Bayes factor of 2.07 in favour of H_1.

This uniform prior is just one example of a prior one might choose. One might decide that the uniform prior is not very realistic and instead decide to employ a non-uniform prior. In our coin flipping example, we might use a prior that places more weight on values further from 0.5 than values closer to 0.5 if we believe trick coins are likely to be heavily biased (for example, a beta prior such as $\theta \sim \mathrm{Beta}(0.9{,}0.9)$). We might use a prior that represents our belief that trick coins will be heavily biased towards coming up heads (for example, a beta prior such as $\theta \sim \mathrm{Beta}(5, 1)$). Or we might believe that trick coins are unlikely to be heavily biased and instead use a prior that places most of its weight at values near 0.5 (for example, a beta prior such as $\theta \sim \mathrm{Beta}(10, 10)$). In each of these cases the Bayes factor will be different: We would obtain values of 0.5, 8.78, and 0.66 in favour of H_0 for each one of these three models or priors. In these examples, we have chosen to use the prior to quantify our beliefs about outcomes that are likely to occur when coins are unfair (that is, they are our models of what unfair coins are like). As Dienes and Mclatchie (2017) points out, this requires the analyst to specify the predictions of the models being compared and thus the Bayes factor can be interpreted as the relative predictive accuracy of the two models. That the

[11] In this context, *prior* refers to the weights we assign to each of the likelihood ratios for each of the possible parameter values. The term *prior* (sometimes *prior odds*) is also used to refer to our predata beliefs about how likely we think it is that H_0 or H_1 is true. This second type of *prior* doesn't factor into the calculation of the Bayes factor but, as noted above, can be used in conjunction with a Bayes factor to determine our post data beliefs. Consequently, if we think that biased coins are infinitesimally rare then even obtaining a large Bayes factor in favour of H_1 would not lead us to conclude that we have encountered a biased coin.

models have to make predictions about what data is likely to be observed has the added benefit that models that are overly vague are penalised. This can simply be illustrated by modifying the width of a prior so that a model predicts an increasingly wide range of implausible values. An example of this (using the default Bayesian *t*-test discussed below) is shown in Fig. 3.

There are two broad schools of thought about how one should go about specifying these model predictions. *Subjective* Bayes approaches seek to employ priors that reflect the analyst's prior beliefs about likely parameter values (Berger 2006; Rouder et al. 2009; Dienes 2014; Dienes and Mclatchie 2017; Gronau et al. 2018), as we have done with our coin flipping example. The *objective* Bayesian approach, on the other hand, seeks priors that are minimally informative.[12] Often priors are sought that are appropriate in as wide a range of scenarios as possible or priors that have good frequentist properties (Berger 2006). One such example is the JZS prior on the effect size parameter, which is found in the default Bayesian *t*-test (Rouder et al. 2009).

The fact that inference from Bayes factors depends on model specifications is not inherently problematic. As our coin flipping example shows, deciding whether a coin is fair or not is dependent on what we think it means for a coin to be unfair. That is, our inferences are specific to the models being compared. However, some difficulties can arise when it comes to specifying the models that are to be compared by the analysis. It is worth examining how disagreements about model specifications can give rise to different inferences by examining a few examples taken from Dienes and Mclatchie (2017). These examples will also be instructive because they were selected to highlight some of the putative benefits of the Bayesian approach over the Frequentist approach.

The first example reported by Dienes and Mclatchie (2017) is of an experiment where participants in two conditions were required to make judgements about the brightness of a light. Dienes and Mclatchie (2017) report the results from both the original finding and a subsequent replication attempt. In the original paper, the authors report a difference between the two conditions in brightness judgement of 13.3 W, and a corresponding statistically significant *t*-test ($t(72) = 2.7$, $p = .009$, cohen's $\mathrm{d} = 0.64$). For the replication attempt the sample size was increased such that if the true effect was of the same magnitude as the original finding the replication attempt would produce a statistically significant result approximately 9 times out of 10—that is, the statistical power would be 0.9. The replication attempt, however, failed to produce a statistically significant result($t(104) = 0.162$, $p = 0.872$, cohen's $\mathrm{d} = 0.03$), and a raw effect of approximately 5.47 W was observed. What is one to make of this failed replication attempt?

Dienes and Mclatchie (2017) state in the case of the second experiment that "[b]y the canons of classic hypothesis testing [that is, frequentist methods] one should accept the null hypothesis." As noted earlier in our discussion of Frequentist inference, a non-significant result does not warrant the inference *accept* H_0, at least not from a principled perspective. However, setting this aside,

[12] Minimally informative (or non-informative) is used here in the technical sense to refer to, for example, Jeffreys' prior, not in the colloquial sense of being vague. A *subjective prior* might be non-informative in the colloquial sense without being non-informative in the technical sense.

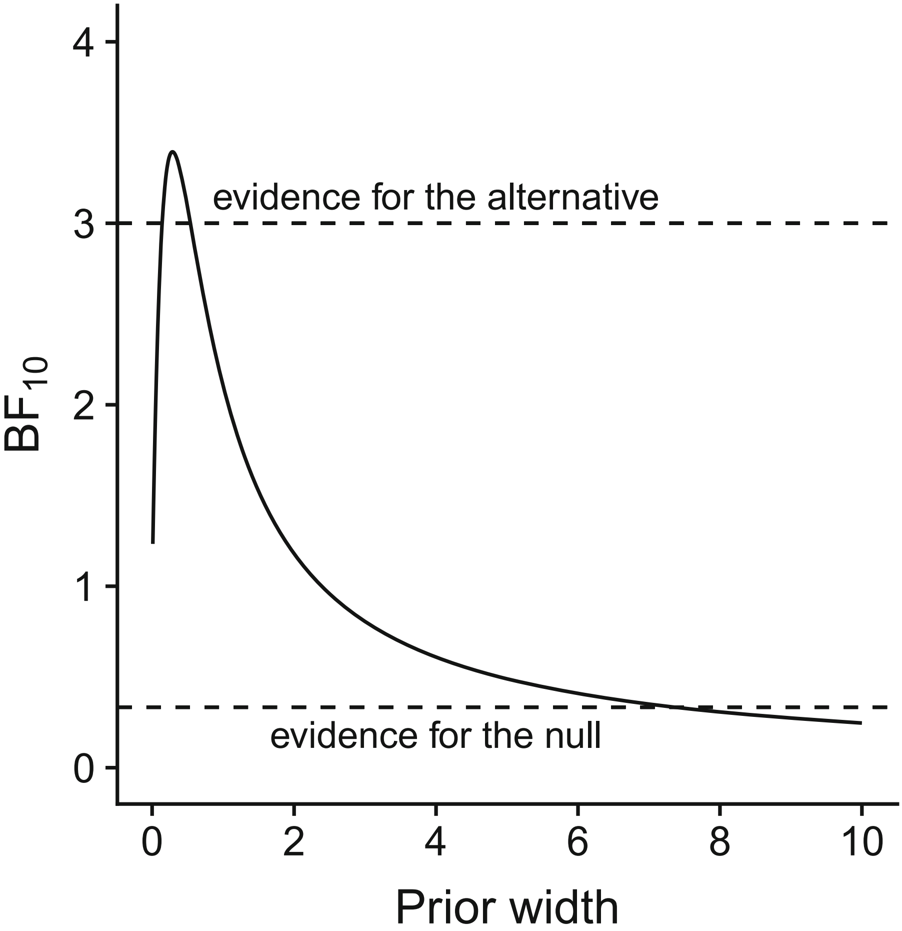

Fig. 3 Bayes factor values as a function of prior width

for now, we can ask what the Bayesian should conclude. According to the analysis presented by Dienes and Mclatchie (2017), the original finding, which reported a raw effect of 13.3 W, should inform the analyst's model of H_1. The resulting Bayes factor computed on the basis of this model after observing the new data (the raw difference of 5.47 W) is approximately 0.97. That is, the Bayes factor value indicates that the new data provide roughly equal support for the null and the alternative and the conclusion should be that the results are inconclusive. Dienes and Mclatchie (2017) may be justified in this specification of an informed prior; however, one might, either through a desire for "objectivity" or through a desire to compare one's inference to some reference, instead choose to use a non-informative prior. The JZS prior, employed in the default Bayesian *t*-test (Rouder et al. 2009), is one such example. Re-running the analysis employing this new model specification for the alternative hypothesis now instead results in a Bayes factor of 0.21—that is, the null is now preferred by a factor of nearly 5 to 1. Interestingly, this is just the same inference as the *heuristic* interpretation of the *p* value.

It is important to note, however, that the fact that the two Bayesian analyses give different results is not a problem, at least not from a Bayesian perspective. The analysis is simply providing a measure of the strength of evidence for one model relative to another model. A problem only arises when one seeks to interpret the Bayes factor as an indication of "*an effect*" *being present* versus "*an effect*" *being absent*. However, it is also worth noting that with default priors (that is, the JZS prior), the model being compared is not really a model of the theory in the same sense as Dienes and Mclatchie's (2017) model is, which somewhat breaks the connection between the statistical hypothesis and the scientific hypothesis. However, since any change in statistical practice is likely to depend on ease-of-use (both in terms of conceptual understanding and the availability of, for example, software tools) it seems likely that default priors may be the dominant type of model specification in use, at least in the short term. And therefore, it is necessary that the appropriate caveats are observed when drawing inferences on the basis of these procedures.

Just as Bayesian inference is relative to specific models, it is also important to reiterate that Frequentist inferences should be relative to specific alternative hypotheses that are assessed against actual observed results. This more sophisticated frequentist analysis would actually draw conclusions more similar to the inferences drawn by Dienes and Mclatchie (2017). For example, the Frequentist might want to use severity assessments to assess various hypotheses with respect to the observed result. If this was done, the inference, like the Bayesian inference would be similarly inconclusive. Inferences about only very small discrepancies being present are not tested with severity (that is, inferences that accord more with the null hypothesis would not be supported). The only inferences that would pass with severity are those that entertain the possibility of a wide range of discrepancies—from negligible to very large—being present (that is, an inconclusive result). Furthermore, a more sophisticated Frequentist might also choose to perform multiple statistical tests to test this one scientific hypothesis, and to build up evidence in a piecemeal manner. One way to do this would be to perform two one-sided tests against the twin null hypotheses of, for example, $H_0 : \mu > -10$ Watts and $H_0 : \mu < 10$ Watts. This would allow the analysts to draw inferences about practical equivalence within the range of, for example, −10 to +10 W. The results of such an equivalence test would be non-significant suggesting that the null hypotheses cannot be rejected and again suggesting that the result is inconclusive ($t(104) = -0.13$, $p = 0.45$).

It is an interesting exercise to apply severity reasoning to the other examples presented by Dienes and Mclatchie (2017). For instance, Dienes and Mclatchie (2017) shows that a Bayesian analysis can be used to show that a non-significant effect from an experiment with low a priori power need not be viewed as evidentially weak. However, severity assessments for non-significant results do not rely on *pre-experiment* power (that is, a power calculation performed before the data is known), as a naïve Frequentist might, but rather assess hypotheses with respect to the data *actually obtained*. For this example, it is possible to probe various hypotheses to see which pass with severity. Applying this reasoning to the same example as Dienes and Mclatchie (2017) would result in concluding that the data are consistent with the presence of a negligible to very small effect, but not consistent with a large effect. Or one might use multiple tests, taken together, such as in an equivalence test procedure, and find that one has good grounds to infer that any deviations from the null fall within the bounds of practical equivalence.[13] Furthermore, severity assessments of a *just* significant effects in a large study would lead one to conclude that there are not good grounds for inferring that anything but a negligible effect is present just as a significant (Frequentist) effect in a large study would lead to a Bayes factor that strongly favours the null model over the alternative model.

4 Two Approaches to Inference, Evidence, and Error

We have outlined a view of inference offered from the Frequentist, error-statistical, perspective in the form of the severity principle: One can only make claims about hypotheses to the extent that they have passed severe tests. And we have outlined a view of inference offered from the Bayesian perspective: One can make claims about hypotheses to the extent that the data support that hypothesis relative to alternatives. These two approaches are often pitched as rivals because it is argued that they can warrant different inferences when presented with the same data, as the examples presented by Dienes and Mclatchie (2017) are meant to show. However, as our discussion of Dienes and Mclatchie (2017) shows, this is not clearly the case. What these examples more clearly demonstrate is that the exact nature of the *question* being asked by Dienes and Mclatchie's (2017) Bayesian analysis and the naïve frequentist analyses they present are different. With different questions one need not be surprised by different answers. The same applies to asking two different Bayesian questions (one using a default prior and one using an informed prior)—a different question results in a different answer. Consequently, when Dienes and Mclatchie (2017) point out pitfalls of significance testing they are in fact pointing out pitfalls associated with a naïve approach. A more sophisticated use of Frequentist inference allows one to avoid many of the common pitfalls usually associated with significance testing and it is not necessary to adopt Bayesian methods if all one wants to do is avoid these misinterpretations.

There are, however, situations where Bayesian and Frequentist methods are said to warrant different inferences that are a consequence of the *process* that allows each type of inference to be justified. Consider, for example, the claim of Wagenmakers (2007) that the feeling that optional stopping is cheating is contradicted by a mathematical analysis. From an error statistical perspective any claims made as a result of optional stopping are not warranted (making those claims *is* cheating) because the claims have not been severely tested (the probability of detecting an error would be very low so not detecting an error is unimpressive). The same applies for data-dredging and a range of other behaviours. For the Bayesian, however, all that matters in assessing the *strength of evidence* is the ratio of the likelihoods. The Bayesian can be seen as regarding *data* as primary while the Frequentist can be seen as regarding the *process* as primary. As noted by Haig (2016), this is a difference between Frequentists (specifically, of the error-statistical variety) favouring local or context-dependent accounts of statistical inferences with Bayesians' favouring broad general or global accounts of statistical inference.

[13] In fact, running such an equivalence test on the data presented in their example does result in one rejecting the null hypothesis of an effect larger than practical equivalence (±1% difference between groups in the number of questions answered correctly) being present ($t(99) = 1.72$, $p = 0.04$).

The important question, however, is how does each approach fair as a system of *scientific inference*? The primary difference between the two can be seen as coming down to error control. Frequentists, like Mayo (Mayo 1996; Mayo and Spanos 2006; Mayo and Spanos 2011) insist that any system of inference must be so designed so that we are not lead astray by the data. Consider the case of collecting observations and then drawing inferences on the basis of these. It might be reasonable to ask whether those observations reflect some truth or whether they are possibly misleading. Bayesian statistics, however, does not care about error probabilities in assessing the strength of evidence. The strength of evidence (derived from the Likelihood Principle) is simply construed as the degree to which the data support one hypothesis over the other with no reference to how often the evidence might be *misleading*. This is in distinction to Frequentist approaches that fix at an upper-bound how often inferences will be in error. This highlights what Birnbaum (1964) called the "anomalous" nature of statistical evidence. Gandenberger (2015), similarly, cautions against using the Likelihood Principle to argue against Frequentist statistics, particularly the error statistical view. Whether the Likelihood Principle is true or not, is simply not relevant for this system of inference and, therefore, Frequentist violations of the likelihood principle are of no consequence (Gandenberger 2015). Similarly, ignoring error probabilities is of no consequence within the Bayesian system of inference (Haig 2016). Gandenberger (2015) states that the likelihood principle only applies if one wants to use methods that track "evidential meaning", but he goes on to state that while "tracking evidential meaning is intuitively desirable… [it may be] less important than securing one or more of [the] putative virtues" of Frequentist methods. These virtues, such as the ability to control error probabilities and the ability to *objectively* track truth (in, for example, the absence of priors), may be virtues that one wishes to retain.

The Bayesian view that the evidential import of the data is only reflected through the likelihoods is also more nuanced than is often recognised. Specifically, the adherence to the Likelihood Principle implies an immunity to stopping rules; however, this immunity must be qualified. There are many instances when the stopping rule may influence the inferences that the Bayesian wants to draw from the data obtained in an experiment. In these situations, the stopping rule is described as *informative*. Stopping rules are said to be *informative* if, for example, they provide information about a relevant unknown parameter that is not contained within the data itself. For example, when trying to estimate some parameter, θ, if the stopping rule is dependent on θ in some way other than through the data, such as by making some stopping rule more likely if $\theta = X$ and another stopping rule more likely if $\theta = Y$, then the stopping rule carries information about θ that is not in the data itself. To adapt an example from Edwards et al. (1963): If you are trying to count the number of lions at a watering hole, then the fact that you had to stop counting because you were chased away by all the lions should factor into any of your inferences about the number of lions. Roberts (1967) presents some more formal examples and suggests that in these cases it is right and proper to take this parameter dependence into account in the likelihood function.

Information about the stopping rule can also enter into a Bayesian inference through the prior more directly when objective priors are used. Consider the example of flipping a coin multiple times and after each flip recording whether it landed on *heads* or *tails*. Once the data is obtained, one might want to make an inference about the probability of obtaining heads. As pointed out by Wagenmakers (2007), for a Frequentist to draw inferences about the observed data they would need to have information about how the data was collected—that is, the stopping rule. Specifically, it would be necessary to know whether, for example, the data were collected until a fixed number of trials were completed or until a fixed number of heads were recorded. The two sampling rules can lead to identical observed data, but since the two sampling rules have something different to say about *possible* data that *could* occur under the null hypothesis, this information must enter into the Frequentist analysis. Etz (2017) also makes use of this example, not to show the flaw in Frequentist inference (which is what Wagenmakers (2007) deploys the example for), but to show how a Bayesian can make use of prior information when computing the posterior probability of obtaining heads. In his example, Etz (2017) shows how one can combine some prior beliefs (for example, the belief that the probability of obtaining heads is likely to be between 0.30 and 0.70) to obtain a posterior distribution of values for obtaining heads. In Etz's (2017) example, his prior quantifies his *pre-data beliefs*, and his posterior quantifies his *post-data beliefs* that have been updated in light of the data. However, how is one to perform the Bayesian analysis if one has no pre-data beliefs or no strong grounds for holding a particular set of pre-data beliefs?

As mentioned earlier, the use of objective priors is meant to circumvent the problems of specifying these subjective priors. The solution, therefore, is just to make use of one of the minimally informative objective priors. Box and Tia (1973) provide just such a set of non-informative priors derived from Jeffreys' rule; however, the exact prior that is appropriate turns out to be dependent on the sampling rule. That this, the "objective" Bayesian inference about the parameter from a set of data turns out to be different depending on how the data were collected. As noted by Hill (1974) and Berger (2006), this amounts to a violation of the Likelihood Principle. In Wagenmakers's (2007) terms, it would result in a Bayesian analysis that is dependent on the *unknown intentions* of the researcher. Box and Tia (1973 p 46) note that they find the observation that a difference in sampling rules leads to different inferences "much less surprising than the claim that they ought to agree." Indeed the requirement that one adheres to the Likelihood Principle in drawing inferences is not universally accepted even among Bayesian's. For example, Gelman and Shalizi (2013) encourage a kind of data-dependent model validation that might similarly violate the Likelihood Principle when the entire inference process is viewed as a whole. Furthermore, Gelman et al. (2014 p 198) state, "'the observed data' should include information on how the observed values arose". That is, good Bayesian inference should be based on all the available information that may be relevant to that inference. However, the assessment of evidence, once data and models are in hand can still be done in a manner that respects the Likelihood Principle.

In addition to cases where *informative* stopping rules are used, cases may also arise where stopping rules that are ostensibly *uninformative* from one perspective might be informative from another perspective. These kinds of situations are likely to arise more often than is often recognised. Gandenberger (2017) outlines such a situation. Consider two researchers, Beth employs the stopping rule: collect data until the likelihood ratio favours H_1 over H_0 by some amount. Brooke employs the stopping rule: collect data until reaching some fixed n. The stopping rule employed by Beth is technically *uninformative* because the stopping rule is only dependent on the data observed and is not dependent on other information about the parameter of interest not contained in the data. If it happens to be the case that Beth and Brooke obtain identical data then the Bayesian analysis states that Beth and Brooke are entitled to identical inferences.

However, consider a third party, Karen, who is going to make decisions on the basis of the data. For Karen, it might not be that easy to discount the stopping rule. For example, if she suspects that Beth might choose her stopping rule on the basis of a pilot experiment that showed evidence in favour of H_0 then the stopping rule contains information that is of some epistemic value to Karen. This situation, where there is a separation between inference-maker and data collector, is not uncommon in science. Other researchers who will make inferences on the basis of published research, journal editors, reviewers, or other end users of research may consider a stopping rule informative even when the researcher themselves does not.

Other instances might also exist where a Bayesian might want to consider stopping rules. One such example is suggested by Berger and Wolpert (1988). They suggest that if somebody is employing a stopping rule with the aim of making some parameter estimate exclude a certain value then an analyst might want to take account of this. For example, Berger and Wolpert (1988) suggest that if a Bayesian analyst thinks that a stopping rule is being used because the experimenter has some belief about the parameter (for example, that the estimate should exclude zero), then adjustments should be made so that the posterior reflects this. These adjustments, however, should not be made to the likelihood—that is, they should not affect the *strength of evidence*—but should instead be made to the prior so that some non-zero probability is placed on the value that the experimenter might be trying to exclude. This approach, however, has not been without criticism. Specifically, the practice of making adjustments to priors because an analyst might *think* that an experimenter *thinks* something about a parameter runs a severe risk of appearing ad hoc. This is especially the case given that much of the Bayesian criticism of Frequentist statistics is based on the claim that unknown *intentions* should not influence inferences. The Frequentist response is much more satisfactory. After all, the Frequentist can point to specific problematic *behaviour* that justifies their rule; however, Berger and Wolpert (1988) appear to suggest that the Bayesian really must care about the mental states of the data collector.

The upshot of examples like this is that far from immunity to stopping rules, the conditions under which stopping rules are informative can be poorly defined. Furthermore, the responses to these situations can be tricky to implement. The fact remains that many of the cases where Frequentists are worried about stopping rules may be the very same cases where stopping rules *should* worry a Bayesian too.

4.1 What Do we Really Want to Know?

What should we make of examples where stopping rules appear to influence the epistemic value of the data? One solution is to ask ourselves what we really need for scientific inference. For example, Gandenberger (2015) recognises that it is reasonable to care about error probabilities despite them having no influence on *evidence*. And Dienes (2011 p 286) suggests that "[u]ltimately, the issue is about what is more important to us: using a procedure with known long term error rates or knowing the degree of support for our theory." There are several legitimate reasons for deciding that *both* are important.

The reasons for wanting to know both is that the two kinds of inferences figure differently in scientific reasoning. Caring about error rates is important because one can learn from the absence of error, but only if there is a good chance of detecting an error if an error exists (e.g., Mayo 1996). When one collects observations it may be less important to know whether or not a particular observation is better predicted by theory A or theory B. Instead, it may be better to know whether inferences about the presence or absence of error are well justified, which is what can be gained from the severity principle. For instance, if we wish to conclude that an observation justifies a conclusion of some deviation from a particular model then whether we have good grounds for this inference can be determined with reference to the severity principle. Similarly, if we wish to conclude that we have good grounds for inferring that there is no deviation (within a particular range), then the severity principle can help here too. And all this can be done without needing to know whether and to what extent that deviation is predicted by two theories.

However, if one has good grounds for making one's models and good grounds for making predictions, then it seems reasonable to care about whether the evidence supports one model over its rival. With some observations in hand, along with some explanations or models, a Bayesian analysis allows us to judge which is the best explanation. Haig (2016) similarly echoes this view that both forms of inference are necessary by calling for *pragmatic pluralism*. However, for this to work it is important to understand the strengths and weaknesses of each approach, the inferences each approach warrants, and when each approach should be deployed. This, however, is a different kind of argument than that which is ordinarily made by those advocating statistical reform (Wagenmakers 2007; Dienes and Mclatchie 2017). The usual strategy here is to argue that Bayesian statistics should be adopted because they lead to more reasonable, more correct, or more intuitive inferences from data relative to Frequentist inference. As we have pointed out in Section 3.2, in our discussion of Dienes and Mclatchie (2017), the Frequentist inference and the Bayesian inference can often be similar on a gross level (the data are inconclusive, the data support an alternative hypothesis, the data do not support an alternative hypothesis) and, therefore, arguing that statistical reform is necessary because macro level inferences are different may not work as a strategy. A better strategy, we believe, is to argue that statistical reform is necessary because it is necessary to have the right tool for the right job in a complete system of *scientific inference*.

Tests of statistical significance find their strength where reasonable priors are difficult to obtain and when theories may not make any strong quantitative predictions. For example, when researchers simply want to know whether they can reliably measure some phenomenon (of a specific magnitude or range) then significance testing might play a role. (Significance tests play an analogous role in physics, see van Dyk 2014). In these contexts, however, it is important that researchers at least have some sense of the magnitude of the effects that they wish to observe so that analyses can be adequately powered. Furthermore, they might be useful in exploratory contexts. This kind of exploratory research is importantly different to data dredging—that is, rather than testing numerous statistical hypotheses, finding significance, and then claiming support for a substantive hypothesis, this kind of exploratory research involves the systematic collection of observations. Importantly, the systematic collection of observations will involve piecemeal accumulation of evidence, coupled with repeated tests and followups to ensure severity. In the psychological sciences, one such context might be neuroimaging[14] where a researcher simply wants to know whether some response can be reliably measured with the aim of later building a theory from these observations (see Colling and Roberts 2010). This is essentially a signal detection task and it does not require that one specify a model of what one expects to find. Instead, the minimal requirement is a model of the noise, and the presence of signals can be inferred from departures from noise. Importantly, theories developed in this way could then be tested by different means. If the theory takes the form of a quantitative model or, better yet, multiple competing plausible models then a switch to Bayesian statistics would be justified.

Bayesian statistics thrives in situations involving model comparison, parameter estimation, or when one actually wishes to assign credences, beliefs, or measure the degree of support for hypotheses. Significance testing has no formal framework for belief accumulation. However, to fully exploit these strengths psychological scientists would not only need to change the way they do statistics but also change the way they do theory. This would involve an increased emphasis on explanation by developing *quantitative* mechanisms (see Kaplan and Bechtel 2011; Colling and Williamson 2014). Unfortunately, the naïve application of significance tests does not encourage the development of mechanistic theories that make quantitative predictions. Rather, the focus on simple dichotomous reject/do not reject thinking can, and has, lead researchers to often be satisfied with detecting *any* effect rather than *specific* effects.

Importantly, the debates around statistical reform and the replication crisis highlight a deeper concern. Rather than *merely* a statistical issue, the replication crisis highlights the stark disconnect between those inferences that are *warranted* and *justified* and those inferences that scientists actually make, both with respect to their own work and with respect to the work of others. Haig (2016) and Szűcs and Ioannidis (2017a) raise similar concerns. Rather than offloading inferences onto sets of numbers produced by statistical procedures, researchers, and particularly students, need to have a greater understanding of how to construct appropriate explanatory theories and how to differentiate substantive and statistical hypotheses. Additionally, it is also important that researchers are able to identify contexts in which hypothesis tests (whether Bayesian or Frequentist) are appropriate and contexts in which parameter estimates are more appropriate—that is, when to *test* hypotheses and when to *measure* phenomena.

5 Conclusions

We do not think that the solution to the replication crisis lies in statistical reform per se. While there are undoubtedly problems with how people justify scientific inferences on the basis of statistical significance tests, these problems may lie less with the tests themselves than with the inferential systems people employ. And we have attempted to demonstrate how good inferences on the basis of statistical significance tests may be justified. We have also examined the Bayesian alternative to statistical significance tests

[14] This is just used as a hypothetical example. Whether this works in practice depends crucially on the ability to control error rates. While controlling error rates is *in theory* possible, in practise, this has proved more difficult (e.g., Eklund et al. 2016).

and explored some of the benefits of the Bayesian approach. The argument for Bayesian statistics is often framed in terms of the macro level inferences that they permit and in terms of the perceived shortcomings of Frequentist statistics. However, we have argued that well-justified Frequentist inferences can often lead to the same gross conclusions. Rather, the key differences lie in their view of evidence and the role error plays in learning about the world. That is, rather than furnishing different inferences, per se, each approach provides a different kind of information that is useful for different aspects of scientific practice. Rather than mere statistical reform, what is needed is for scientists to become better at inference (both Frequentist and Bayesian) and for a better understanding of how to use inferential strategies to justify knowledge.

Acknowledgements We would like to thank two anonymous reviewers for their comments on their manuscript. Their thoughtful comments greatly improved this manuscript. This work was supported by the James S. McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition (grant number 220020370).

References

Babbage C (1830) Reflections on the decline of science in England, and on some of its causes. B. Fellows.

Bem, D.J. 2009. Writing an empirical article. In *Guide to publishing in psychology journals*, ed. R.J. Sternberg, 3–16. Cambridge: Cambridge University Press.

Bem, D.J. 2011. Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology* 100: 407–425. https://doi.org/10.1037/a0021524.

Benjamin, D.J., J.O. Berger, M. Johannesson, B.A. Nosek, E.J. Wagenmakers, R. Berk, K.A. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C.D. Chambers, M. Clyde, T.D. Cook, P. De Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A.P. Field, M. Forster, E.I. George, R. Gonzalez, S. Goodman, E. Green, D.P. Green, A.G. Greenwald, J.D. Hadfield, L.V. Hedges, L. Held, T.H. Ho, H. Hoijtink, D.J. Hruschka, K. Imai, G. Imbens, J.P.A. Ioannidis, M. Jeon, J.H. Jones, M. Kirchler, D. Laibson, J. List, R. Little, A. Lupia, E. Machery, S.E. Maxwell, M. McCarthy, D.A. Moore, S.L. Morgan, M. Munafő, S. Nakagawa, B. Nyhan, T.H. Parker, L. Pericchi, M. Perugini, J. Rouder, J. Rousseau, V. Savalei, F.D. Schönbrodt, T. Sellke, B. Sinclair, D. Tingley, T. Van Zandt, S. Vazire, D.J. Watts, C. Winship, R.L. Wolpert, Y. Xie, C. Young, J. Zinman, and V.E. Johnson. 2017. Redefine statistical significance. *Nature Human Behaviour* 33 (1): 6–10. https://doi.org/10.1038/s41562-017-0189-z.

Berger, J.O. 2006. The case for objective bayesian analysis. *Bayesian Analysis* 1: 385–402. https://doi.org/10.1214/06-BA115.

Berger, J.O., and R.L. Wolpert. 1988. *The Likelihood Principle*. Hayward: Institute of Mathematical Statistics.

Birnbaum, A. 1964. *The anomalous concept of statistical evidence: Axioms, interpretations, and elementary exposition*. New York University.

Bishop, D.V.M. 2014. *Interpreting unexpected significant findings*. https://doi.org/10.6084/m9.figshare.1030406.v1.

Box, G.E.P., and G.C. Tia. 1973. *Bayesian inference in statistical analysis*. Weskey Publishing Company.

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

Cohen, J. 1992. Statistical power analysis. *Current Directions in Psychological Science* 1: 98–101. https://doi.org/10.1111/1467-8721.ep10768783.

Colling, L.J., and R.P. Roberts. 2010. Cognitive psychology does not reduce to neuroscience. In *9th conference of the australasian society for cognitive science*, 41–48. Sydney: Macquarie Centre for Cognitive Science.

Colling, L.J., and K. Williamson. 2014. Entrainment and motor emulation approaches to joint action: Alternatives or complementary approaches? *Frontiers in Human Neuroscience* 8: 67. https://doi.org/10.3389/fnhum.2014.00754.

Cramer, A.O.J., Ravenzwaaij D. van, D. Matzke, H. Steingroever, R. Wetzels, R.P.P.P. Grasman, L.J. Waldorp, and E.-J. Wagenmakers. 2015. Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review* 23: 640–647. https://doi.org/10.3758/s13423-015-0913-5.

de Winter, J.C., and D. Dodou. 2015. A surge of p-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too). *PeerJ* 3: e733. https://doi.org/10.7717/peerj.733.

Dienes, Z. 2011. Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science* 6: 274–290. https://doi.org/10.1177/1745691611406920.

Dienes, Z. 2014. Using bayes to get the most out of non-significant results. *Frontiers in Psychology* 5. https://doi.org/10.3389/fpsyg.2014.00781.

Dienes, Z., and N. Mclatchie. 2017. Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review* 100: 1–12. https://doi.org/10.3758/s13423-017-1266-z.

Edwards, W., H. Lindman, and L.J. Savage. 1963. Bayesian statistical inference for psychological research. *Psychological Review* 70: 193–242. https://doi.org/10.1037/h0044139.

Eklund, A., T.E. Nichols, and H. Knutsson. 2016. Cluster failure: Why fMRI inferences for spatial extent have inflated -positive rates. *Proceedings of the National Academy of Sciences of the United States of America* 113: 7900–7905. https://doi.org/10.1073/pnas.1602413113.

Etz A (2017) Introduction to the concept of likelihood and its applications. Advances in Methods and Practices in Psychological Science.

Fisher, R.A. 1925. Statistical methods for research workers. In *Oliver*. London: Boyd.

Gandenberger, G. 2015. A new proof of the likelihood principle. *The British Journal for the Philosophy of Science* 66: 475–503. https://doi.org/10.1093/bjps/axt039.

Gandenberger, G. 2017. Differences among noninformative stopping rules are often relevant to Bayesian decisions. arXiv:1707.00.214 [math.ST].

García-Pérez, M.A. 2016. Thou shalt not bear false witness against null hypothesis significance testing. *Educational and Psychological Measurement* 77: 631–662. https://doi.org/10.1177/0013164416668232.

Gelman, A., and C.R. Shalizi. 2013. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology* 66: 8–38. https://doi.org/10.1111/j.2044-8317.2011.02037.x.

Gelman, A., J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. 2014. *Bayesian Data Analysis*. 3rd ed. Boca Raton: CRC Press.

Gigerenzer, G. 1993. A handbook for data analysis in the Behaviorial sciences. In *The superego, the ego, and the id in statistical reasoning*, ed. G. Keren and C. Lewis, 311–340. New York.

Gronau, Q.F., A. Ly, and E.-J. Wagenmakers. 2018. Informed Bayesian t-tests. arXiv:1704.02479 [stat.ME].

Haig, B.D. 2016. Tests of statistical significance made sound. *Educational and Psychological Measurement* 77: 489–506. https://doi.org/10.1177/0013164416667981.

Hill, B.M. 1974. Review of bayesian inference in statistical analysis. *Technometrics* 16: 47800479. https://doi.org/10.1080/00401706.1974.10489222.

Ioannidis, J.P.A. 2012. Why science is not necessarily self-correcting. *Perspectives on Psychological Science* 7: 645–654. https://doi.org/10.1177/1745691612464056.

Jeffreys, H. 1961. *The theory of probability*. 3rd ed. Oxford: Claredon Press.

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

Kaplan, D.M., and W. Bechtel. 2011. Dynamical models: An alternative or complement to mechanistic explanations? *Topics in Cognitive Science* 3: 438–444. https://doi.org/10.1111/j.1756-8765.2011.01147.x.

Lakens, D. 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science* 8: 355–362. https://doi.org/10.1177/1948550617697177.

Lakens, D., F.G. Adolfi, C.J. Albers, F. Anvari, M.A. Apps, S.E. Argamon, T. Baguley, R.B. Becker, S.D. Benning, D.E. Bradford, E.M. Buchanan, A.R. Caldwell, B. Calster, R. Carlsson, S.-C. Chen, B. Chung, L.J. Colling, G.S. Collins, Z. Crook, E.S. Cross, S. Daniels, H. Danielsson, L. DeBruine, D.J. Dunleavy, B.D. Earp, M.I. Feist, J.D. Ferrell, J.G. Field, N.W. Fox, A. Friesen, C. Gomes, M. Gonzalez-Marquez, J.A. Grange, A.P. Grieve, R. Guggenberger, J. Grist, A.-L. Harmelen, F. Hasselman, K.D. Hochard, M.R. Hoffarth, N.P. Holmes, M. Ingre, P.M. Isager, H.K. Isotalus, C. Johansson, K. Juszczyk, D.A. Kenny, A.A. Khalil, B. Konat, J. Lao, E.G. Larsen, G.M. Lodder, J. Lukavský, C.R. Madan, D. Manheim, and S.R. Martin. 2018. Justify your alpha. *Nature Human Behaviour* 2: 168–171. https://doi.org/10.1038/s41562-018-0311-x.

Lindley, D.V. 2000. The philosophy of statistics. *Journal of the Royal Statistical Society: Series D (The Statistician)* 49: 293–337. https://doi.org/10.1111/1467-9884.00238.

Masicampo, E.J., and D.R. Lalande. 2012. A peculiar prevalence of pvalues just below. *05. The Quarterly Journal of Experimental Psychology* 65: 2271–2279. https://doi.org/10.1080/17470218.2012.711335.

Mayo, D.G. 1996. *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.

Mayo, D.G., and R.D. Morey. 2017. *A poor prognosis for the diagnostic screening critique of statistical tests*. https://doi.org/10.17605/OSF.IO/PS38B.

Mayo, D.G., and A. Spanos. 2006. Severe testing as a basic concept in a NeymanPearson philosophy of induction. *The British Journal for the Philosophy of Science* 57: 323–357. https://doi.org/10.1093/bjps/axl003.

Mayo, D.G., and A. Spanos. 2011. Error statistics. In *Philosophy of statistics*, ed. P.S. Bandyopadhyay and M.R. Forster. Oxford.

Morey, R.D., R. Hoekstra, J.N. Rouder, M.D. Lee, and E.J. Wagenmakers. 2016a. The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review* 23: 103–123. https://doi.org/10.3758/s13423-015-0947-8.

Morey, R.D., J.-W. Romeijn, and J.N. Rouder. 2016b. The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology* 72: 6–18. https://doi.org/10.1016/j.jmp.2015.11.001.

Neyman, J. 1976. Tests of statistical hypotheses and their use in studies of natural phenomena. *Communications in statistics—theory and methods* 5: 737–751. https://doi.org/10.1080/03610927608827392.

Neyman, J., and E.S. Pearson. 1933. On the problem of the Most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 231: 289–337. https://doi.org/10.1098/rsta.1933.0009.

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

Nuzzo, R. 2014. Scientific method: Statistical errors. *Nature* 506: 150–152. https://doi.org/10.1038/506150a.

Open Science Collaboration. 2012. An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science* 7: 657–660. https://doi.org/10.1177/1745691612462588.

Pashler, H., and E.-J. Wagenmakers. 2012. Editors' introduction to the special section on replicability in psychological science. *Perspectives on Psychological Science* 7: 528–530. https://doi.org/10.1177/1745691612465253.

Phillips, K.F. 1990. Power of the two one-sided tests procedure in bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics* 18: 137–144. https://doi.org/10.1007/BF01063556.

Roberts, H.V. 1967. Informative stopping rules and inferences about population size. *Journal of the American Statistical Association* 62: 763. https://doi.org/10.2307/2283670.

Rouder, J.N. 2014. Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review* 21: 301–308. https://doi.org/10.3758/s13423-014-0595-4.

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

Simmons, J.P., L.D. Nelson, and U. Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22: 1359–1366.

Simonsohn, U. 2015. Small Telescopes. *Psychological Science* 26: 559–569. https://doi.org/10.1177/0956797614567341.

Steegen, S., F. Tuerlinckx, A. Gelman, and W. Vanpaemel. 2016. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science* 11: 702–712. https://doi.org/10.1177/1745691616658637.

Stroebe, W., T. Postmes, and R. Spears. 2012. Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science* 7: 670–688. https://doi.org/10.1177/1745691612460687.

Szűcs, D. 2016. A tutorial on hunting statistical significance by chasing N. *Frontiers in Psychology* 7: 365. https://doi.org/10.3389/fpsyg.2016.01444.

Szűcs, D., and J.P.A. Ioannidis. 2017a. When null hypothesis significance testing is unsuitable for research: A reassessment. *Frontiers in Human Neuroscience* 11: 943. https://doi.org/10.3389/fnhum.2017.00390.

Szűcs, D., and J.P.A. Ioannidis. 2017b. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology* 15: e2000797. https://doi.org/10.1371/journal.pbio.2000797.

Trafimow, D., and M. Marks. 2014. Editorial. *Basic and Applied Social Psychology* 37: 1–2. https://doi.org/10.1080/01973533.2015.1012991.

van Dyk, D.A. 2014. The role of statistics in the discovery of a Higgs boson. *Annual Review of Statistics and Its Application* 1: 41–59. https://doi.org/10.1146/annurev-statistics-062713-085841.

Wagenmakers, E.-J. 2007. A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review* 14: 779–804. https://doi.org/10.3758/BF03194105.

Wagenmakers, E.-J., R. Wetzels, D. Borsboom, and Maas H.L.J. van der. 2011. Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology* 100: 426–432. https://doi.org/10.1037/a0022790.

Ware, J.J., and M.R. Munafò. 2015. Significance chasing in research practice: Causes, consequences and possible solutions. *Addiction* 110: 4–8. https://doi.org/10.1111/add.12673.

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

Yong, E. 2012. Replication studies: Bad copy. *Nature* 485: 298–300. https://doi.org/10.1038/485298a.

Yu, E.C., A.M. Sprenger, R.P. Thomas, and M.R. Dougherty. 2013. When decision heuristics and science collide. *Psychonomic Bulletin & Review* 21: 268–282. https://doi.org/10.3758/s13423-013-0495-z.