Journal of Mathematical Psychology 72 (2016) 78–89

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

How Bayes factors change scientific practice

Zoltan Dienes *

School of Psychology, University of Sussex, Brighton, BN1 9QH, UK
Sackler Centre for Consciousness Science, University of Sussex, UK

HIGHLIGHTS

- Bayes factors would help science deal with the credibility crisis.
- Bayes factors retain their meaning regardless of optional stopping.
- Bayes factors retain their meaning despite other tests being conducted.
- Bayes factors retain their meaning regardless of time of analysis.
- The logic of Bayes helps illuminate the benefits of pre-registration.

ARTICLE INFO

Article history:
Available online 7 January 2016

Keywords:
Bayes factor
Null hypothesis
Stopping rule
Planned vs post hoc
Multiple comparisons
Confidence interval

ABSTRACT

Bayes factors provide a symmetrical measure of evidence for one model versus another (e.g. H1 versus H0) in order to relate theory to data. These properties help solve some (but not all) of the problems underlying the credibility crisis in psychology. The symmetry of the measure of evidence means that there can be evidence for H0 just as much as for H1; or the Bayes factor may indicate insufficient evidence either way. *P*-values cannot make this three-way distinction. Thus, Bayes factors indicate when the data count against a theory (and when they count for nothing); and thus they indicate when replications actually support H0 or H1 (in ways that power cannot). There is every reason to publish evidence supporting the null as going against it, because the evidence can be measured to be just as strong either way (thus the published record can be more balanced). Bayes factors can be *B*-hacked but they mitigate the problem because a) they allow evidence in either direction so people will be less tempted to hack in just one direction; b) as a measure of evidence they are insensitive to the stopping rule; c) families of tests cannot be arbitrarily defined; and d) falsely implying a contrast is planned rather than post hoc becomes irrelevant (though the value of pre-registration is not mitigated).

1. Introduction

A Bayes factor is a form of statistical inference in which one model, say H1, is pitted against another, say H0. Both models need to be specified, even if in a default way. Significance testing (using only the *p*-value for inference, as per Fisher, 1935) involves setting up a model for H0 alone—and yet is typically still used to pit H0 against H1. I will argue that significance testing is in this way flawed, with harmful consequences for the practice of science (Wagenmakers, 2007). Bayes factors, by specifying two models, resolve several key problems (though not all problems). After defining a Bayes factor, the introduction first indicates the general consequences of having two models (namely, the ability to obtain evidence for the null hypothesis; and the fact the alternative has to be specified well enough to make predictions). Then the body of the paper explores four ways in which these consequences may change the practice of science for the better.

1.1. What is a Bayes factor?

In order to define a Bayes factor, the following equation can be derived with a few steps from the axioms of probability (e.g. Stone, 2013): Normative posterior belief in one theory versus another in the light of data = a Bayes factor, $B \times$ prior belief in one theory versus another. That is, whatever strength of belief one happened to have in different theories prior to data (which will be different for different people), that belief should be updated by the same

* Correspondence to: School of Psychology, University of Sussex, Brighton, BN1 9QH, UK.
E-mail address: dienes@sussex.ac.uk.

http://dx.doi.org/10.1016/j.jmp.2015.10.003
0022-2496/

amount, B, for everyone.[1] What this equation tells us is that if we measure strength of evidence of data as the amount by which anyone should change their strength of belief in the two theories in the light of the data, then the only relevant information is provided by the Bayes factor, B (cf Birnbaum, 1962). Conventional approximate guidelines for strength of evidence were provided by Jeffreys (1939, though Bayes factors stand on their own as continuous measures of degrees of evidence). If $B > 3$ then there is substantial evidence for H1 rather than H0; if $B < 1/3$ then there is substantial evidence for H0 rather than H1; and if B is in between 1/3 and 3 then the evidence is insensitive.

The term 'prior' has two meanings in the context of Bayes factors. $P(\text{H1})$ is a prior probability of H1, i.e. how much you believe in H1 before seeing the data. But the term 'prior' is also used to refer to setting up the model of H1, i.e. to state what the theory predicts, used for obtaining $P(\text{D}|\text{H1})$, the probability of obtaining the data given the theory. When measuring strength of evidence with Bayes factors, there is no need to specify priors in the first sense; but there is a need to specify a model (prior in the second sense). To know how much evidence supports a theory one must know what the theory predicts; but one does not have to know how much one believes in a theory a priori. In this paper, specifying what a theory predicts will be called a 'model'.

1.2. The consequences of having two models

The specification of two models in a Bayesian approach, rather than one in significance testing, has two direct consequences: One is that Bayes factors are symmetric in a way that p-values are asymmetric; and, second, Bayes factors relate theory to data in a direct way that is not possible with p-values. Here I clarify what these two properties mean; then the paper will consider in detail how these properties are important for how we do science.

First, a Bayes factor, unlike a p-value, is a continuous degree of evidence that can symmetrically favour one model or another (e.g. Rouder, Speckman, Sun, Morey, & Iverson, 2009). Let us call the models H1 and H0. By using conventional criteria, the Bayes factor can indicate whether evidence is weak or strong. Thus, the Bayes factor may indicate (i) strong evidence for H1 and against H0; or (ii) strong evidence for H0 and against H1; or (iii) not much evidence either way. That is a Bayes factor can make a three-way distinction. A p-value, by contrast, is asymmetric. A small p-value (often) indicates evidence against H0 and for the H1 of interest; but a large p-value does not distinguish evidence for H0 from not much evidence for anything. A p-value only tries to make a two-way distinction: evidence against H0 (i.e. (i)) versus anything else (i.e. (ii) or (iii), without distinguishing them) (and even this it does not do very well; Lindley, 1957). A large p-value is, therefore, never in itself evidence for H0. The asymmetry of p-values leads to many problems that are part of the 'credibility crisis' in science (Pashler & Wagenmakers, 2012). The reason why p-values are asymmetric is that they specify only one model: H0. This is their simplicity and hence their beguiling beauty. But their simplicity is simplistic. This paper will argue that using Bayes factors will therefore help solve some (but not all) of the problems leading to the credibility crisis, by changing scientific practice. The symmetry is particularly important in determining support for the null hypothesis, interpreting replications, and p-hacking by optional stopping, all practical issues discussed below.

The strict use of only one model is Fisherian; Neyman and Pearson (1967) argued that two models should be used, and introduced the concept of power, which helps introduce symmetry in inference, in that it provides grounds for asserting the null hypothesis. Unfortunately power is a flawed solution (Dienes, 2014) and that might explain why it is not always taken up. Power cannot be determined based on the actual data in order to assess their sensitivity; hence, a high powered non-significant result might not actually be evidence for the null hypothesis, as we shall see. Further, it involves (or should involve) specifying only the minimal interesting effect size, which is a rather incomplete specification of H1 (and it is the aspect of H1 most difficult to make in many cases). In practice, psychologists are happy to assert null hypotheses even when power has not been calculated, and inference is based on p-values alone (as we shall see).

The second consequence of having to specify H1 as well as H0 is that thought must be given to what one's theory actually predicts (Vanpaemel, 2010). In this way, Bayes factors allow a more intimate connection between theory and data than p-values allow. This issue is particularly important for dealing with issues of multiple testing and the timing of theorizing versus collecting data. I conjecture that a Bayesian view of these issues will lead to a more probing exploration of theory than significance testing encourages, a point taken up at the end.

The paper now considers in detail the specific changes to scientific practice the use of Bayes factors may bring about. Specifically it considers, in order, issues of obtaining support for the null hypothesis; of the effect of stopping rules on error rates; of dealing with multiple comparisons in theory evaluation; and, finally, of planned versus post hoc tests and the role of timing of theory and data in scientific inference. I will argue that Bayesian inference compared to significance testing leads to a re-evaluation of all these issues.

2. Changes to scientific practice

2.1. Supporting the null hypothesis

Here we consider in turn the problem of providing support for the null hypothesis; how Bayes factors help; and why the orthodox solution of using power does not solve the problem, as illustrated by high powered attempts to replicate studies.

The problem. The key problem created by the asymmetry of the p-value is that significance testing per se (i.e. inference by use of p-values) cannot provide evidence for the null hypothesis. Indeed, that is exactly how p-values are asymmetric. Despite that, a non-significant result is often in practice taken as evidence for a null hypothesis. For example, to take one of the most prestigious journals in psychology, in the 2014 April issue of the *Journal of Experimental Psychology: General*, in 32 out of the 34 articles, a non-significant result was taken as support for a null hypothesis (as shown by the authors claiming no effect), with no further grounds given for accepting the null other than that the p-value was greater than 0.05. That is, in the vast majority of the articles where there were no grounds for accepting the null hypothesis at all, the null hypothesis was nonetheless accepted, often in order to draw important theoretical conclusions. The effect of this practice can be disastrous. For example, the drug paroxetine was originally declared to have no risk of increased suicide in children because the increase of risk was non-significant (it was later shown to have such a risk, Goldacre, 2013). Human death aside, do we want

[1] In symbols:

$$P(\text{H1}|\text{D})/P(\text{H0}|\text{D}) = P(\text{D}|\text{H1})/P(\text{D}|\text{H0}) \times P(\text{H1})/P(\text{H0})$$

$P(\text{H1})/P(\text{H0})$ is the ratio of the probabilities (or strength of belief) in H1 versus H0, i.e. the prior odds of H1 versus H0. $P(\text{H1}/D)/P(\text{H0}|\text{D})$ is the ratio of the probabilities of the two theories in the light of the data; i.e. the posterior odds. The remaining term is the Bayes factor, B, which states that the data are B times more probable under H1 rather than H0. Briefly, posterior odds $= B \times$ prior odds.

to guide our theory development partly on conclusions that are groundless[2]?

Researchers may know that inferring the null hypothesis from a non-significant result is suspect. That obviously does not stop the practice from happening, it just makes sure it happens freely in papers where there also are also key significant results. But where the key result is non-significant, papers are less likely to be published (Rosenthal, 1979; Simonsohn, Nelson, & Simmons, 2014). The research record becomes a misleading representation of the evidence. Because the p-value is asymmetric, people seek to get the evidence in the only way it can appear to be strong—as against H0. Thus, apart from failure to publish relevant evidence concerning a theory, another outcome is p-hacking: Pushing the data in the one direction it can for it to be recognized as strong evidence, by use of analytic flexibility (John, Loewenstein, & Prelec, 2012; Masicampo & Lalande, 2012; Simmons, Nelson, & Simonsohn, 2011). No wonder there is a crisis in the credibility of our published results.

How Bayes factors help. Bayes factors partly solve the problem by allowing the evidence to go both ways. This means you can tell when there is evidence for the null hypothesis and against the alternative. You can tell when there is good evidence against there being a treatment side effect (and when the evidence is just weak); you can tell when the data count against a theory (and when they count for nothing). There is every reason to publish evidence supporting the null as going against it, because the evidence can be measured to be just as strong either way (thus the published record can be balanced). In fact, the Bayes factor is the only way for indicating the strength of evidence for a point null hypothesis (though for a Bayes factor H0 need not be a point value; Dienes, 2014; Morey & Rouder, 2011). People can still "B-hack" (i.e. massage data to get a Bayes factor just beyond some conventional threshold by the use of analytic flexibility), but we will explore how options are more limited than for p-hacking in important ways.

Power and replication. Replications are hard to evaluate by reference to p-values. If an original result was significant, and a direct replication non-significant, it might feel like a failure to replicate. But as p-values cannot indicate whether the null hypothesis is supported, a non-significant replication tells one nothing in itself. This is even true for high powered non-significant replications. The point can be illustrated conceptually by considering a high powered replication where both H0 and H1 specify point values. If the sample mean is exactly half way between H0 and H1, then no matter what the power, the data do not discriminate the theories in any way. In fact, if in a non-significant experiment, the sample mean were closer to H1 than H0, the data would support H1 more than H0 no matter how highly powered the experiment. Thus, it is rational to consider how the data actually come out to consider what they say, and power cannot do this.

Most theories allow more than just one point value; then Bayes factors can be used to specify the strength of evidence. For example, consider the Reproducibility Project (https://osf.io/ezcuj/) spearheaded by Brian Nosek (Open Science Collaboration, 2015). The aim was to establish how well 100 experiments published in 2008 in high impact journals in psychology replicate, when the exact methods specified are followed as closely as possible. In the replication of Correll (2008) by LeBel (https://osf.io/fejxb/wiki/home/), the original "PSD slope" reported in the Correll paper (Study 2) was 0.18, $SE = 0.077$, $F(1, 68) = 5.52$, $p < 0.02$. The attempted direct replication doubled sample size to achieve a power of 85%. The slope in the replication was 0.05, $SE = 0.056$, $F(1, 145) = 0.79$, $p = 0.37$. This looks like a "failure" to replicate. In fact, calculating a Bayes factor (see Dienes, 2014, 2015, for details of how to calculate), $B_{H(0, 0.18)} = 0.69$, indicating that the evidence is weak and does not substantially support either H0 or H1 (the value of B is between 1/3 and 3).[3]

In case it is thought that 85% power just is not good enough, consider the replication of Estes, Verges, and Barsalou (2008). These original authors found an incongruent priming condition caused more errors than a congruent condition, the difference being 4.8%, $SE = 1.6\%$, $F(1, 17) = 9.33$, $p = 0.007$. Renkewitz and Muller (https://osf.io/vwnit/) attempted an exact replication with a power of well over 95% for detecting this error difference. In the replication, they found a difference in errors of 1.4%, $SE = 1.1\%$, $F(1, 21) = 1.45$, $p = 0.24$. This is non-significant and hence a "failure" to replicate. However, $B_{H(0, 4.8)} = 0.79$, indicating the evidence was not discriminating between H0 and H1: there are no grounds for changing one's confidence in either H0 or H1 to any substantial degree based on the replication. On the other hand, it is quite possible to get evidence for the null using a Bayes factor in experiments with such numbers of participants; in the same replication, the effect on reaction times, which was significant in the original paper (a 37 ms effect, $SE = 6$ ms, $F(1, 17) = 40.19$, $p < 0.001$), was non-significant in the replication (0.2 ms, $SE = 6$ ms, $F(1, 21) = 0.001$, $p = 0.5$), and also $B_{H(0,37)} = 0.19$ (i.e. $B < 1/3$), with 22 subjects, indicating substantial support for the null. The point is that knowing power alone is not enough; once the data are in, the obtained evidence needs to be assessed for how sensitively H0 is distinguished from H1, and power cannot do this (Dienes, 2014). (Compare Etz, 2015, for a Bayesian analysis of the experiments in the Reproducibility Project.)

In sum, Bayes factors would enable a more informed evaluation of replications than p-values allow. The need for more direct replications is clear (Pashler & Harris, 2012); but replications are no good if one cannot properly evaluate the results.

Now we will consider some inferential paradoxes. The asymmetry of p-values leads to a sensitivity to stopping rules which is inferentially paradoxical, because the same data and theories can be evaluated differently depending on the intentions inside the head of the experimenter (e.g. Berger & Wolpert, 1988). We now consider this and other inferential paradoxes that allow p-hacking. The paradoxes mean that inferential outcome depends on more than the actual data obtained, and may depend on things which are in practice unknowable (the intentions and thoughts of experimenters; see Dienes, 2011 for explanation). The need to correct for multiple testing with significance testing is a paradox in that theories may pass or fail tests on data collected that was irrelevant to the theory, but corrected for anyway. Instead Bayesian approaches in which the model of H1 is informed by scientific context focus only on the relation between theory and the data that bear on specifically that theory. Similarly, the use of timing of theory versus data as inferentially relevant in itself disguises what is actually very important about pre-registration of studies, as we will discuss.

[2] The sample difference being small, zero, or in the wrong direction does not in itself provide sufficient grounds either; see Dienes (2014) for examples.

[3] Meta-analytic combined estimates should be analysed with Bayes factors too (Dienes, 2014; Rouder & Morey, 2011). In this case, the fixed effect combined mean estimate of Study 2 of Correll (2008) and the replication is 0.095, $SE = 0.0453$, $t(213) = 2.10$. In Study 1 of Correll the PDS slope was 0.18; Study 2 sought to manipulate this slope, and 0.18 remains a useful scale for predicting effects in Study 2 and the replication. On the combined data of Study 2 and the replication, $B_{H(0, 0.18)} = 3.78$, substantial support for H1 with all data combined. ($B_{H(0,0.18)}$ indicates that H1 was represented as a half-normal with a mode of zero and a standard deviation of 0.18; that is, the population difference is represented as being between 0 and roughly 2×0.18. See Dienes (2014), for explanation.) The Bayesian version of meta-analysis enjoys all the advantages of Bayesian inference in general; for example, it allows one to obtain support for a null hypothesis, not possible with a meta-analysis using significance testing.

2.2. The stopping rule

First we consider the problem, how stopping rules influence error rates, and thus allow cheating. Then we consider how this problem is side-stepped by Bayes factors. Finally, we consider how stopping rules can lead to biased estimates, and the Bayesian answer to this problem.

The problem. Imagine that after each addition of an observation to data, a p-value is calculated. If H0 is false, the p-value is driven towards small values. However, if H0 is true, the p-value does a random walk. That means sooner or later, if H0 is true, the p-value will randomly wander below 0.05 (Rouder et al., 2009). So if one uses significance testing, it is strictly forbidden to keep topping up participants, without a pre-planned correction. Yet John et al. (2012) estimate that virtually 100% of psychologists at major US universities have topped up participants after initially failing to get a significant result. If one decides to continue running until a significant result is obtained, significance is guaranteed even if H0 is true. Thus, one has to decide on the conditions one would stop in advance of collecting data—and then stop at that point. By contrast, a Bayes factor B is symmetric. If H0 is false, then, in the long run, B is driven upwards. If H0 is true, B is driven towards zero. Because B is driven in opposite directions dependent on which theory is true, when using a Bayes factor one can stop collecting data whenever one likes (Savage, 1962). Thus, use of Bayes factors respects the "stopping rule principle" according to which the only evidence about a parameter is contained in the data and not the stopping rule used to collect them (Berger & Berry, 1988a,b; Berger & Wolpert, 1988).

A useful rule would be to stop collecting data when either B is greater than 3 or less than 1/3; then one has guaranteed an informative conclusion with a minimum number of participants (cf. Schoenbrodt, Wagenmakers, Zehetleitner, & Perugini, in press). (Something which power cannot guarantee: A study can be high-powered but still the data do not discriminate between the models.) While significance testing allows p-hacking by optional stopping, one cannot B-hack by optional stopping.

The possibility that one can legitimately ignore the stopping rule would be such a dramatic and useful change to practice, that it might seem too good to be true. Consider the following argument for why the conclusion might be false. The value of B, as any statistic, is subject to noise, and surely one can capitalize on that noise by stopping for example when $B > 3$ (if it were to be), even when H0 is true? Indeed, Yu, Sprenger, Thomas, and Dougherty (2014) and Sanborn and Hills (2014) showed that one could indeed substantially raise the false alarm rate for B when H0 was true by using just such a stopping rule. The effect can be illustrated even with a symmetric stopping rule. Imagine an experiment where each participant provides a difference score, say their cognitive performance with and without a cognitive enhancer. We have prior information that implies that if there were to be an effect of a cognitive enhancer, it would be about one point for the dependent variable used. Following Dienes (2014), H1 is modelled as a half-normal with an SD of the expected size of effect (i.e. 1). For simplicity, assume the population standard deviation of scores is 1. When running for a fixed 100 trials, simulation of the experiment 1000 times (see Appendix A for details) showed that when H0 was true, B exceeded three 1% of the time, and B was less than a third 86% of the time. That is the false alarm rate was only 1%.

Table 1 indicates what happened when the stopping rule was as follows: After every participant, check to see if either $B > 3$ or else $B < 1/3$. If so, stop. Otherwise run another participant and continue until either the threshold is crossed or else 100 subjects are reached. In terms of researcher practice, this is a worst case scenario; researchers do not typically check after every participant, but maybe only two or three times when the initial result is non-significant; see Dienes (2011) for why the latter practice is wrong when uncorrected for orthodox statistics (and see Sagarin, Ambler, & Lee, 2014, for appropriate corrections). Each number in Table 1 is the outcome of 200 simulations. Appendix B gives the R code. Appendix A shows the results for different types of Bayes factors. Notice that when the same threshold for B (i.e. three/a third) is used as for our example with a fixed number of subjects in the last paragraph, the false alarm rate for when H0 was true increased from 1% to 14%. That is, the stopping rule affected the false alarm rate of the Bayes factor. Does this not contradict the claim that inference using B is immune to the stopping rule?

Why the stopping rule is a not a problem for Bayes factors. Rouder (2014) argued elegantly for why the sensitivity of the false alarm rate to the stopping rule is consistent with inference from B remaining immune to the stopping rule. Here the same argument will be put slightly differently. First notice that the equation 'posterior odds $=$ $B *$ prior odds' follows from the axioms of probability. That is, given that the axioms normatively specify how the strength of belief should be changed, B is normatively the amount by which the strength of belief should be changed regardless of the stopping rule. If strength of evidence is measured by how much in principle beliefs should normatively be changed, then B is normatively the measure of strength of evidence discriminating two theories. The stopping rule does not come into the equation, so the claim is true regardless of the stopping rule. But how does this fit with false alarm changing according to the stopping rule?

Notice that B is the measure of evidence regardless of the specific value of $P(\mathrm{D}|\mathrm{H0})$. That is, $P(\mathrm{D}|\mathrm{H0})$ can in principle vary as B stays the same. B will still be the measure of strength of evidence—because $P(\mathrm{D}|\mathrm{H1})$ will change by just the right amount. Experimental psychologists are used to such reasoning with signal detection theory. Discriminability in a perceptual decision task can remain the same as bias changes; we would never dream of measuring discriminability by measuring the false alarm rate in a signal detection experiment. Obviously the same point applies to H0 versus H1. That is, false alarm rate of a procedure can change when discriminating H0 versus H1 even when the ability of the procedure to discriminate remains invariant. The evidence provided by an observation remains the same even if the criterion is changed (and hence false alarm rate changes). B is the invariant measure of the strength of evidence for H1 versus H0, regardless of false alarm rate.

We as experimental psychologists have become fixed on false alarm rate for measuring the strength of evidence for a theory because we were taught to consider only one model (H0) for significance testing. It is like trying to perform signal detection theory with only one distribution, that for noise alone. But in signal detection theory terms, that is a nonsense; we need the signal distribution as well. Bayes considers two distributions: One for H0 and one for H1. False alarm rate is, by itself, uninformative about how well the theories are discriminated.

The Supplementary Materials[4] give R code for measuring the false alarm and hit rates for Bayes factors for optional stopping. One can vary, amongst other things, the threshold, population effect sizes, and the minimum or maximum number of participants before optional stopping can begin. Table 2 shows the same situation as Table 1, but with a minimum of 10 participants before optional stopping could start. The false alarm rate for a threshold of three is halved (see first column in Table 2 compared to Table 1). B will be most variable early on in testing, because B

[4] http://dx.doi.org/10.1016/j.jmp.2015.10.003.

Table 1
Per cent decision rates for accepting/rejecting H0 for $B_{H(0,1)}$ (i.e. a Bayes factor in which H1 has been represented as a half-normal, with mode = 0, and SD = 1). Each participant provides a single difference score, sampled from a normal distribution with a standard deviation of 1. Thus, the specified population effect sizes are dz's (Cohen, 1988). Maximum number of participants before stopping (MaxN) = 100; minimum number of participants before checking after every trial (MinN) = 1. H0 is rejected if B exceeds the stated threshold, and accepted if B goes below 1/threshold.

	Threshold:	3	4	5	6	7	8	9	10
Population effect:									
$dz = 0$	Reject H0	14	12	11	11	7	7	6	5
	Accept H0	86	87	86	86	85	79	74	69
$dz = 1$	Reject H0	97	100	100	100	100	100	100	100
	Accept H0	1	0	0	0	0	0	0	0

Table 2
Per cent decision rates for accepting/rejecting H0 for $B_{H(0,1)}$ as for Table 1, except that MinN = 10.

	Threshold:	3	4	5	6	7	8	9	10
Population effect:									
$dz = 0$	Reject H0	7	7	7	5	5	3	4	3
	Accept H0	93	91	88	81	83	82	73	66
$dz = 1$	Reject H0	100	100	100	100	100	100	100	100
	Accept H0	0	0	0	0	0	0	0	0

is driven in different directions according to which theory is true as data accumulates. Once B has picked up momentum in the right direction, it may never exceed a value in the opposite direction, even after an infinite number of participants (Savage, 1962). Thus, having a minimum number of participants, even a small amount, can reduce false alarm rate. Note that B is always and invariably the correct measure of strength of evidence for discriminating H0 versus H1, regardless of whether a minimum number of participants is used. Nonetheless if one wanted to control false alarm rate, in addition to discriminability, the Supplementary Materials (see Appendix B) would allow the reader to work out how to do so by, for example, changing the minimum number of participants, or raising the threshold of B. (There is another reason to run a minimum number of participants: the validity of the Bayes factor, as for any statistical test, depends on the assumptions of the statistical model of the data being approximately true. A minimum number of participants allows assumptions to be checked; Morey, Romeijn, & Rouder, 2013). The Supplementary Materials (see Appendix B) also provide results for different types of Bayes factors.

Appendix A illustrates how Bayes factors have better error properties as a function of the stopping rule not only than significance testing, but also than the use of confidence or credibility intervals.

In sum, Bayes factors can be used as a measure of evidence irrespective of the stopping rule, and hence optional stopping is not a form of B-hacking. In fact stopping when $B > 3$ or $< 1/3$ (or any other threshold) would enable stopping when the data are just as discriminating as needed. This guarantees the sensitivity of a study with a minimum of participants.

The issue of bias. It might be argued that, although Bayes factors are insensitive to the stopping rule as a measure of evidence, the estimates of population values can be biased by the stopping rule. Thus, we could be in the seemingly awkward position of having fine hypothesis testing but biased parameter estimation, depending on the stopping rule. To illustrate bias arising according to the stopping rule, if a researcher was interested in the effect of a drug on mood, she could decide to stop testing after she found three participants in a row who were happier on the drug than on placebo. The resulting estimate of how happy the drug made people would be biased upwards. Bias is a frequentist notion that therefore needs a reference class to define it; the reference class in this case is defined by the stopping rule. That is, let the researcher repeat the experiment an infinite number of times (and to allow the argument to be clear, assume the researcher can be taken as randomly sampling from the same population as before), each time stopping the experiment after three participants in a row were happier on the drug than on placebo. Even if the drug were ineffective, each estimate would have a tendency to indicate that people were happier on the drug; that is, the mean of all the estimates would show greater happiness on the drug than on the placebo. Is not this a problem for an experiment, even if analysed by Bayesian statistics?

The clue to the solution is that bias is inherently a frequentist notion, with need of a reference class (Howson & Urbach, 2006); yet it is the use of reference classes that leads to the inferential paradoxes in significance testing that do not apply to Bayesian analyses (Dienes, 2011; Lindley, 1993). Our researcher, as a Bayesian, would not simply average the results of the different experiments together (in an unweighted way). The experiments are all basic events in the reference class; but a Bayesian does not recognize the reference class as relevant to inference. Note that each experiment would have a different number of participants. The events in the reference class are just one arbitrary way of carving up the full set of data (as given by stringing together the infinite number of experiments the researcher runs). Different stopping rules (defining different reference classes) would partition the same full set of data into different events. The same data could be partitioned such that each experiment finished with three people in a row who were happier on placebo rather than drug (now the bias goes the other way). But all that matters is the complete data set, not the arbitrary partitionings of it. The experimenter should combine all her participants together, and then average such that each participant contributes equally. This procedure (of averaging over participants all the data that one has so far) converges in the limit to the correct value of the population mean (cf Rouder, 2014). The frequentist by contrast has to work within the reference class predefined by her, and so bias is a genuine worry: By frequentist methods, the average (over reference class events) converges to the correct value only if the stopping rule provides unbiased estimates.[5]

[5] It may seem that the Bayesian solution of weighting according to participant number is open to the frequentist; indeed, the frequentist may complain that the solution I provide above is just as frequentist as Bayesian. But the frequentist is conceptually obliged to respect reference classes even in meta-analyses. Consider Smith performing a study which obtained $p = 0.08$ and publishing. Jones, based

To recap: The stopping rule can introduce bias to estimates when the expected value of the estimate is taken over the events of a reference class. But such bias is irrelevant to Bayesian procedures, whether theory testing (Bayes factors) or estimating population parameters. Bayes factors would change scientific practice because hacking by optional stopping would be ruled out. Given the prevalence of optional stopping (John et al., 2012), this would produce a major change in the robustness of our science.

2.3. Corrections for multiple testing

First we consider the problem that multiple testing gives multiple opportunities for errors, yet correcting for this introduces inferential arbitrariness; then we consider the Bayesian solution, which removes arbitrariness.

The problem. One way people can cheat with inferential statistics is to make many comparisons and then focus on the one that was significant taken on its own. The frequentist solution is to correct for multiple testing. If with frequentist statistics one decides to correct for familywise error rate, the correction depends on an arbitrary specification of what the family is, allowing analytic flexibility (for evidence of wide spread prevalence of problems created by flexibility in defining relevant families, see Ioannidis, Munafò, Fusar-Poli, Nosek, & David, 2014; John et al., 2012). With Bayes the issue becomes one of specifying how different theories are affected by all the data relevant to them, which is not arbitrary. We consider an imaginary example to illustrate the issues and their solution.

An example is now presented in order to consider the issue of families of tests. Six studies are run testing the effect of referring to the general concept of "closing" on how quickly a sale is closed (i.e. how quickly the sale is agreed and completed). The maximum time allocated to the sale was 5 min in each study. A previous priming study using the same selling paradigm, but priming by seating the client in soft vs. hard chairs, obtained a priming effect of 15 s. Thus, based on the past study, in the current experiment one might expect a priming effect of on the order of magnitude of roughly 15 s if priming existed (so we can model H1 as a half normal with an *SD* of 15 s, following Dienes, 2014). In one study frequent verbal reference was made by the salesperson to closed doors compared to a control condition; in another condition the salesperson incidentally discussed Sunday closing rules; and, for example, in the final study, the salesperson made frequent hand gestures reminiscent of a closing door. Each condition had its own matched control. In one of the studies, the one with hand gestures of closing doors, reference to closure indeed resulted in faster closure of the sale as compared to its control condition (with opening hand gestures), mean effect = 10 s, $SE = 5$ s, $t(30) = 2.0$, $p < 0.05$. $B_{H(0,15)} = 3.72$, indicating substantial evidence for the effect of priming as opposed to the null hypothesis. None of the other studies were significant, nor had Bayes factors above 3.

A researcher might be tempted to report only the one study that worked. It did after all involve the most embodied references to closing (bodily hand gestures rather than word primes), and it might be presumed, that is why that particular study worked. The other studies, which were all different in a possibly relevant way, had therefore not found the right conditions for eliciting the effect[6]. This type of reasoning is very tempting and there must be a place for it in exploration. Often researchers explore the conditions for eliciting an effect before they find conditions that appear to work. Nonetheless, choosing one study from many is cherry picking. It is cherry picking because the other studies must have been designed in the first place because it was felt they did test the general theory that priming closure speeds closure. And when relevant data have not been reported because the results looked better without them, Bayes factors in themselves cannot make up for that systematic exclusion. So if only the one "successful" study were reported, both Bayesian and conventional statistics would inappropriately show the evidence to be stronger than it actually was for the general theory that priming "closing" speeds the closing of sales. (Pre-registration of studies, by itself neither Bayesian nor non-Bayesian, is a key solution to this problem; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012.) That is, bias introduced by the judicious dropping of conditions, discussed by Simmons et al. (2011), is not in itself solved by using Bayesian methods on the data that remains. Bayesian analyses will not solve all forms of bias (and then only the ones that are part of the formal statistical problem; Jaynes, 2003). A Bayesian analysis requires that all relevant data are included.

In fact, say that the authors report all studies, so cherry picking is avoided. What is the evidential value of the final study showing an effect? The orthodox approach corrects for multiple testing. Thus, if all studies are taken as a family, a threshold of $0.05/6 = 0.008$ may be used for *p*-values, now rendering the final study non-significant at the 5% level: The presence of an effect cannot be asserted for the embodied priming manipulation. However, from a Bayesian point of view, the evidence provided by the data from specifically the final study for the hypothesis that embodied primes are effective remains the same, no matter what other procedures are tested. The Bayes factor remains 3.72 for the evidential worth of the data from the final study, noteworthy evidence for H1 concerning this particular procedure. Is this not a problem for Bayes?

Before considering the Bayesian solution, first note the flexibility in the frequentist one. Families do not have to be defined by theoretical question in frequentist statistics, and indeed often are not (they may e.g. be defined by degrees of freedom in an omnibus test, e.g. Keppel & Zedeck, 1989). By contrast, the Bayesian solution is to consider the evidence for each theory. In frequentist terms there is no reason why families could not be made of various subsets of the studies. In frequentist terms if the sixth study was treated as planned it could be tested separately from the others, which are then each corrected at the 0.05/5 level as one family. We will consider planned vs. post hoc tests below. For now we consider how Bayes just depends on the relation of the data to theories.

Why multiple comparisons are not a problem for Bayes factors. The priming technique used in the final study is a variant of a number of different priming techniques addressing a common question. Let us say the mean priming effect for the other studies was 0. Now the overall priming effect across all studies is $(10 + 0)/6 = 1.7$. For simplicity, assume all studies had identical standard deviations and Ns. The standard error for the overall mean effect is $5/\sqrt{6} = 2.2$. Thus, $B_{H(0,15)} = 0.30$, support for the null hypothesis that priming closure does not lead to faster closures. In evaluating the general theory that priming closure speeds closure, all relevant data must be used. And when all data are used, the data sensitively support the null hypothesis (using for illustration

on Smith's *p*-value being tantalizing close to 0.05, runs 20 more participants, and combines the data together in a meta-analysis. The resulting meta-analytic $p = 0.04$ is not significant at the 5% level just because it was Jones who topped up and not Smith (see Section 2.1; so long as Jones topping up is conditional on the *p*-value obtained by Smith, the overall error rate of the Jones–Smith pair is above 5%). Frequentists may intuitively grasp for Bayesian solutions, but that does not make the frequentist version legitimate (for a similar argument for confidence vs. credibility intervals, see Morey, Hoekstra, Rouder, Lee, & Wagenmakers, in press).

[6] A test of the difference between embodied (mean = 10, $SE = 5$) and the others (mean = 0, $SE = 5/\sqrt{5} = 2.2$), $t(180) = 1.82$, ns; $B_{H(0,15)} = 2.83$, indicates only anecdotal evidence for a difference.

the guidelines of e.g. Jeffreys, 1939, for interpreting the size of Bayes factors, though note Bayes factors stand by themselves as continuous degrees of evidence).

The theory that embodied primes of closure speeds closure can be regarded as a specific or subordinate level theory of a general or superordinate level theory that priming closure speeds closure. Here the specific theory received substantial evidence considered on its own, while the general theory overall had substantial evidence against it. Naturally, in specific cases evidence pointed in somewhat different directions, as evidence will; but taken together, the evidence was clearly against the general theory. Note there was no need to correct for multiple comparisons; we just had to take account all evidence directly testing a given theory. (Scientific judgement will always determine the relevant specific and general theories.)

Bayes factors enable a properly nuanced extraction of information from data, unlike significance testing. That is, Bayesian inference reflects the fact there is evidence for a specific theory, while the data as a whole count against the more general theory. If one had independent reasons for thinking embodiment was especially important for making priming effective, more data could be collected, using the manipulation of the final study, until the evidence for both the subordinate and superordinate theories was convincing one way or the other. That is, publishing support for both the subordinate and superordinate theories (and showing statistically that embodiment was better than non-embodiment) would require more data than currently on the table. On the other hand, if previous research indicated that word and embodied primes were roughly similar in efficacy in other domains, one could simply go with the evidence for the superordinate theory, and regard it as substantially weakened by the data.

The same logic can be applied generally to cases where multiple tests are applied, all bearing on a common superordinate theory. Consider collecting data on a vast number of EEG and ERP measures on meditators and matched non-meditators, hoping to find indicators of superior attentional abilities in meditators than non-meditators. (So the scientific question has set the relevant superordinate hypothesis.) "Attentional ability" could manifest itself in a large number of different ways: More theta density? Larger P300 amplitudes? And on and on. The problem this design raises is that conclusions could rely on cherry picking a few EEG measures that came out as expected, and ignoring all those that gave non-significant results. Without corrections for multiple testing, how could Bayesian analyses protect against seeing patterns in noise? Here is one way to proceed. First specify which dependent variables measure attention, or the sort of attention we may be interested in. Then devise a way of meta-analytically combining all those measures into a single one. Finally test the strength of evidence that the overall measure provides for the superordinate theory that meditators have stronger attentional skills than non-meditators. Combining evidence is not peculiarly Bayesian (though combining evidence to obtain an overall strength of evidence is). But the procedure does show why Bayesian inference leads to sensible answers when it comes to multiple testing situations. When our real interest is in a general theory, we must assess all evidence for that theory. By realizing that non-significant results may carry evidential value, Bayesian inference encourages researches to use all available data. Significance testing can encourage ignoring non-significant results as non-evidential and hence cherry picking the significant ones.[7]

For another example, consider finding evidence for a difference in activation between conditions in one tiny voxel in an fMRI study. If that voxel is structurally and theoretically arbitrary, it means nothing for theory development. Results mean nothing except in so far as they inform interesting theory. The question is, when we combine activation across theoretically and structurally meaningful sets of voxels, what remains of the evidence? (And as soon as you construct a meaningful theory about what is going on, consider what other voxels are now implicated in testing the theory. Only when all evidence relevant to the superordinate theory has been taken into account can the superordinate theory be evaluated.)

The strategy suggested so far relies on using a Bayes factor to test a single degree-of-freedom hypothesis. This provides a simple broadly applicable strategy but the use of Bayes factors is not limited to this strategy. A superordinate theory that specifies a rank ordering of means in different conditions can also be tested with a Bayes factor using the methods of Hoijtink (2011). For example, a theory that specified that the mean for the first and second conditions should be the same but higher than those from a third, specifies a set of ordinal constraints which together are richer than a single degree-of-freedom comparison. An editor might be especially prepared to accept a paper in favour of or against a superordinate theory if the theory received substantial evidence as a whole (either for or against), regardless of the direction of specific cherry-picked comparisons. Of course, the single degree of freedom comparisons (first mean versus second mean; their average versus the third) would help pinpoint strength of evidence for specific claims made by the theory.

So far it might be thought that Bayes does little better than significance testing in dealing with multiple testing situations (after all, in orthodox statistics one could combine evidence across situations in theory relevant ways). Bear in mind that in Bayesian inference one is not at liberty to define families at will; one has to ask about the relation of data to each specific theory of interest, so "families" must be picked out as the tests relevant to a given theory. Bayesian inference can indicate the support for or against any specified theory. But Bayesian inference can do more, by taking into account the full Bayesian apparatus that lies beyond non-Bayesian approaches. A Bayes factor represents the strength of evidence data provides for one theory rather than another. That evidence informs the posterior probabilities for the different theories. The posterior probability that embodied priming of closure is effective may be affected by the evidence for priming using words; that is, if there is priming for words it increases the probability that there could be priming from gestures, and vice versa. The evidence from the other studies, using different priming procedures, may rationally affect the posterior probability of any one of the priming techniques working. This is because these specific theories fall under the same general theory. Gelman et al. (2013) and Kruschke (2010) describe how to set up hierarchical models whereby the posterior distributions of the means of different conditions is automatically influenced by the data from all conditions. This has the effect of making it harder to detect an effect of embodied priming if there were no priming in any other condition (cf correction for multiple testing); but easier if there were priming in other conditions. This rational adjustment cannot be done with non-Bayesian approaches. In essence the procedure provides a sort of correction for multiple testing—but not for the sake of correcting for multiple testing, but for the sake of making the most of all the relevant data.[8]

[7] When an informed model is tested, Bayesian inference requires one draw on all and only the relevant data. Strange as it is in hindsight, frequentist statistics just have not operated in that way. Frequentists could copy Bayesians in pooling data relevant to theories. But why not start from principles that directly lead to the right answer, rather than those that underspecify what to do?

[8] The procedure amounts to saying there is evidence relevant to the embodiment prime beyond that contained in the data for just the embodiment condition.

In sum, significance testing involves arbitrary corrections for multiple testing, where there is no need to define families by the theory the data are relevant to (indeed, people are often urged to define families by other criteria, like omnibus degrees of freedom in pre-packaged statistical routines such as ANOVA). Bayes factors (where H1 is motivated by theory) explicitly relate theories to data. It may be that specific theory receives support while a general theory is weakened (or vice versa). That is what the data say; what to do next is a matter for scientific not statistical judgement. Bayesian inference would change scientific practice because calculating a Bayes factor requires specifying two models, and thus encourages being clear about what theory the data bear on. Thus, families cannot be defined arbitrarily, but only by reference to theories of scientific interest.

2.4. Planned versus post hoc tests

First we consider the problem, that the timing of theory relative to data intuitively feels important, yet correcting for it introduces inferential arbitrariness; then we consider the Bayesian solution, which removes the arbitrariness.

The problem. One intuition is that it is desirable to predict the precise results one obtained in advance of obtaining them. Indeed, in an estimated 92% of papers in psychology and psychiatry, the results confirm the predictions (Fanelli, 2010). Yet when the predictions are made in advance of seeing the data, the confirmation rate is considerably less (Open Science Collaboration, 2015). Scientists feel a pressure to obtain confirmatory results. For significance testing it makes a difference whether one thought of one's theory before analysing the data or afterwards (planned versus post hoc comparisons). In Bayesian inference all that matters are the data and the theory, not their timing (because the Bayes factor depends just on the probability of the data given the theory).

At first, the Bayesian answer might seem strange. We have all read papers where when we got to the end of the introduction and read the "predictions", we thought "You are only saying that because that is what your results are". We feel cheated. A post hoc result is being falsely treated as a prediction. Is this not wrong? But wait a minute. You knew there was a problem just by reading what you had in front of you. That shows the real problem existed independently of the timing of events; the real problem was the relation of predictions to theory as evident in the paper itself. What really matters is how tightly and simply predictions follow from a simple and elegant theory. Those criteria are obviously not met by our example paper. The paper would be flawed just as much even if, in fact, the authors had thought of their predictions before looking at the data. The data are not actually likely given any stated general theory—that's the problem. Opposite or different predictions could just as well be generated from the stated general theory (if any theory were stated). Consider an opposite case: Einstein's finding that his theory of general theory, developed around 1915, explained the anomalous orbit of Mercury, known since 1859. It was a key result that helped win scientists over to his theory (Lanczos, 1974). First the result was known, then the theory was developed. But the theory had its own independent elegant motivation. What is important is the theory's simplicity and elegance both in itself and in application to the results, not which came first.

The role of timing in Bayesian inference. Timing is a proxy or correlate of what we are really interested in: Predictions genuinely made in advance are likely to be strongly motivated by a simple theory. Post hoc predictions are likely to be arbitrarily related to simple theory. A useful rule of thumb is that confirming novel rather than post hoc predictions is more likely to provide strong evidence for a simple theory. But that is not to do with some magic about when someone thought of a theory (someone's brilliance in mentally penetrating the structure of Mother Nature in advance may be relevant to their self-esteem but such personal brilliance does not transfer to the evidential support of the data for the theory: In science it does not matter who you are). The objective properties of theory and data as entities in their own right (Feynman, 1998; Popper, 1963) need to be separated from accidental facts concerning when certain brains thought of the theory. Gelman and Loken (2013) illustrate this beautifully by considering how, in a range of real examples, different results would have more simply confirmed a general theory than the results on offer. The metaphysics and the epistemology get put in their right place by Bayesian inference (getting a prediction right in advance has no metaphysical status as an indication of good theory; but it does help us know when we have one).

In considering what a general theory predicts in order to calculate the Bayes factor, one might be tempted to use the obtained data to refine the estimate of the magnitude of the prediction for those very same data. That is the Bayesian way of cheating. The data are thereby "double counted", once for connecting theory to predictions, then again for considering whether the predictions are confirmed, and so involve a violation of the axioms of probability (Jaynes, 2003; Jeffreys, 1939). Double counting has to be evaluated with respect to whether the axioms of probability are violated. For example, the general theory that 'priming occurs in this context' cannot be evaluated by using the obtained data to specify what the theory predicts (and then using the same data to test the predictions of the general theory). So what about if one found the Bayes factor not for the general theory but for a specific theory specifying the magnitude of the effect, which happened to be the magnitude shown in the data? That is now OK. All that matters is the probability of the data given the theory; where the theory came from does not matter, according to the principles of Bayesian inference. The issue, and hence the solution, is similar to that considered in the last section: If we are as scientists interested in the general theory, then an arbitrary version of it has no special interest to us beyond any other arbitrary version. While there may be evidence for the specific theory that priming occurs in this context with magnitude 12.63 s, there may be evidence against the general theory that priming occurs in this context (cf. Section 2.3). Further, if a mechanism of priming occurred to you after looking at the data, and for reasons independent of the data that mechanism would predict a likely priming effect of 12 ms, the data provide support for that theory.

The Bayesian answer helps show why pre-registered reports, such as used in Cortex and now at least 16 other journals (Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014; Wagenmakers et al., 2012; see the website Registered Reports, 2015, for regular updates) are valuable. It is not due to the magical power of guessing Nature in advance. Rather, pre-registration ensures the public availability of all results that are pre-registered, regardless of the pattern, which is important for all approaches to statistical inference, Bayesian or otherwise (Goldacre, 2013). This alone is sufficient to justify an extensive use of pre-registered reports. In addition, pre-registration may help us judge such things as simplicity and elegance of theory more objectively. How much judgements of the properties of theory and their relation to predictions are affected by knowing the results in a naturalistic scientific context needs to be investigated further, but it is likely

Thus, using his procedure amounts to a different assumption than if one just used the Bayes factor based on the embodiment data. What this evidence does is change the prior distribution for the embodiment prime; the posterior is thereby affected. Naturally, different scientific judgements concerning relevance can affect the Bayesian outcome.

to be a substantial factor, perhaps moderated by experience (Arkes, 2013; Slovic & Fischhoff, 1977). This is an extra-statistical consideration that does not undermine the direct conclusions that follow from a Bayesian analysis (how well a theory is supported by data relative to another theory), but does raise the issue of the context of scientific judgements within which those conclusions are embedded.

Finally, and very importantly, pre-registration helps deal with the problem of analytic flexibility (Chambers, 2015). There are generally various ways of analysing a given data set, each roughly equally justified. What should the cut off for outliers be—two or three SD, or something else? What transformation might be used, if the data look roughly equally normal with several? Should a covariate be added? Should the dependent variables be combined, or one of them dropped? And so on. Such considerations can affect Bayes factors as much as t-tests. It is possible to *B*-hack. Imagine that out of 10 roughly equally valid analysis methods, nine indicate support for H0 and one does not, as shown by a Bayes factor in each case. If one chose the final method because it fitted one's agenda better, the Bayes factor no longer reflects what the data on balance say. However, if one chose the full analysis method in advance, it will, with 90% probability, be one of the nine methods supporting H0. Thus, with pre-registration of the analysis protocol, the Bayesian analysis is more likely to reflect the overall message of the data. Note that in this case as well, the timing of predictions is just a proxy for the real thing; what actually matters are the objective properties of the data as they are. If by fluke (and it will sometimes happen) the pre-registered analysis method was the one method that did not obtain support for H0, the Bayesian analysis now fails to reflect the overall message of the data, even though the method was pre-registered. Thus, having all data transparently available must also be part of the solution. Then anyone with the time can check different ways of analysing the data for themselves. And in any argument that ensues, it may be worth bearing in mind that the pre-registered method is likely, but is not guaranteed, to reflect what the data say on balance.

The argument for pre-registration is particularly compelling for fMRI. Carp (2012a; see also 2012b) considers common analytic decisions made in the fMRI literature and shows these lead to 34,560 significance maps, which can be substantially different from each other. Considering all of these for each experiment is not feasible. Pre-registration would mean the Bayes factors calculated are likely to be reflective of the data.

In sum, using Bayes factors would change scientific practice by focusing attention on what matters—the relation of data to theory. No one would have pressure to pretend when they thought of the theory. People can focus on how simple and elegant the theory is and how tightly the predictions follow. Bayesian inference does not in itself solve all issues to do with the timing of events however; it should be combined with other solutions, such as pre-registration and full transparency. Even then, Bayesian inference sheds light on what the benefits of pre-registration actually are, and it should provide a conceptual framework to help focus discussions about the worth of different analyses (pre-registered versus after the data came in). For example, even if one boldly stated in advance that embodiment was more likely to produce priming than the use of words, that fact just in itself would not change the Bayes factors given in the last section (e.g. as weakening the general theory or supporting the specific one).

3. Discussion

Bayes factors provide a symmetrical measure of evidence for one model versus another (e.g. H1 versus H0) in order to relate theory to precisely the data relevant to it. These properties help solve some (but not all) of the problems underlying the credibility crisis in psychology. The symmetry of the measure of evidence means that there can be evidence for H0 just as much as for H1; or the Bayes factor may indicate insufficient evidence either way. *P*-values (even with power calculations) cannot make this three-way distinction, but making it is crucial to the integrity of science. Bayes factors can be *B*-hacked but they mitigate the problem because (a) they allow evidence in either direction so people will not be tempted to hack in just one direction; (b) as a measure of evidence they are insensitive to the stopping rule; (c) families of tests cannot be arbitrarily defined; and (d) falsely implying a contrast is planned rather than post hoc becomes irrelevant (though the value of pre-registration is not mitigated).

One advantage Bayesian inference can have is that it forces one to think about what one's theory really predicts. To calculate a Bayes factor, the theory (i.e. the psychological explanation) is represented as a distribution of possible population parameter values. Call this the model of H1 (i.e. a mathematical description of relations). Our psychological theories are rarely stated directly as probability distributions over parameter values; thus, there needs to be a translation from theory to model, a translation that can take into account other findings in the data or literature to refine predictions. The translation is not one to one; typically, the same theory could be translated to different models and different theories can be translated to the some of the same models. Strictly, the Bayes factor indicates the relative support for the model versus H0; it is an extra-statistical matter to decide what work the theory did and how much credit it should get. For example, the theory that caffeine improves concentration because it is a placebo predicts that a cup of coffee should enhance performance on a concentration task. The exact model for how much a cup of coffee enhances concentration could be informed by the effect sizes past studies using coffee. If the model is supported how much does that support bear on the theory? That is a matter of scientific judgement, not statistics per se, and will depend on the full context (cf Gelman & Rubin, 1995). The art of science is partly setting up experiments where interesting theories can be compared using simple models, so that the Bayes factor is informative in discriminating the theories. Thus, one should set up a test of a theory, that when translated into a model, makes a risky prediction, i.e. one contradicted by other background knowledge (Popper, 1963; Roberts & Pashler, 2000; Vanpaemel, 2014) so that the Bayes factor is likely to be discriminating if used to compare the contrasting theories.

One problem with using Bayes factors is precisely that the psychological theory could be translated to several models; yet the support indicated by any given Bayes factor strictly refers to the model not the theory. Thus, the distribution in the model needs to have those properties that capture relevant predictions of the theory in context, while the distribution's other properties should not alter the qualitative conclusion drawn from the resulting Bayes factor. If the outcome is robust to large distributional changes (while respecting the implementation of the same theory), the distributions are acceptable for use in Bayes factors, and the conclusion transfers to the theory (cf Good, 1983). This is referred to as robustness checking. For example if the application of a theory to an experiment indicates that the raw maximum difference should not be more than about m, then try simple distributions that satisfy this judgement yet change their shapes in other ways: Dienes (2014) suggests a uniform from 0 to m; a half-normal with mode 0 and standard deviation $m/2$; and a normal with mean $m/2$ and standard deviation $m/4$. In all cases the (at least rough) maximum is m yet in one case the distribution is flat, in another the probability is pushed up to one side, and in another peaked in the middle. If the qualitative conclusions remain unaltered, the conclusion carries from the models to the theory. A different approach may be to declare in advance which distribution will be

Table 3
Decision rates for accepting/rejecting H0 for NR[−0.1, 0.1] MaxN = 100 MinN = 1.

	Minimum width:	None	10 ∗ NRW	5 ∗ NRW	4 ∗ NWR	3 ∗ NRW	2 ∗ NRW	NRW	0.5 ∗ NRW
Actual effect:									
$dz = 0$	Reject	16	14	7	4	3	1	0	0
	Accept	0	0	0	0	0	0	0	0
$dz = 1$	Reject	100	100	100	100	100	100	0	0
	Accept	0	0	0	0	0	0	0	0

Table 4
Decision rates for accepting/rejecting H0 for NR[−0.1, 0.1] MaxN = 1000 MinN = 1.

	Minimum width:	None	10 ∗ NRW	5 ∗ NRW	4 ∗ NWR	3 ∗ NRW	2 ∗ NRW	NRW	0.5 ∗ NRW
Actual effect:									
$dz = 0$	Reject	19	12	6	5	2	1	0	0
	Accept	72	78	83	84	86	88	89	0
$dz = 1$	Reject	100	100	100	100	100	100	100	0
	Accept	0	0	0	0	0	0	0	0

used (with reasons) on the grounds such a distribution is likely to reflect the conclusion from most simple representations of the theory (see Section 2.4).

An example of Bayes factors motivating a closer consideration of theory is provided by Dienes (2015); see also the examples in Lee and Wagenmakers (2014): Sometimes Bayes requires that extra data are gathered on a different condition in order to interpret another condition, data not demanded by *p*-value calculations. For example, in order to claim that a measure of conscious knowledge shows chance performance, we need data to estimate what level of conscious performance could be expected if the priming or learning performance claimed to be unconscious had actually been based on conscious knowledge. Further, as soon as one thinks what level of raw effect size would be predicted in one's study, one has to carefully consider the literature with eyes one may not have had before, to estimate how well effect sizes in one paper might apply to one's own, given a change in context. Once effect sizes become relevant to the conclusions one draws, people may pay attention to them.

In conclusion, I argue that the use of Bayes factors is a crucial part of the solution to the crisis in which psychology (and other disciplines) find themselves. Now that the problems of what we have been doing up to now are evident (e.g. Ioannidis, 2005; John et al., 2012; Open Science Collaboration, 2015; Pashler & Harris, 2012), I hope Bayes is seriously considered as part of the solution—along with, for example, full transparency and online availability of materials, data and analysis (Nosek et al., 2015); greater emphasis on direct replications as well as multi-experiment theory building (Asendorpf et al., 2013); and increasing use of pre-registration (Chambers, Dienes, McIntosh, Rotshtein, & Willmes, 2015).

Appendix A. Comparing error properties of a Bayes factor with inference by intervals

One way of distinguishing H1 from H0 is by use of inference by intervals (Dienes, 2014). This requires specifying not a point null, but a null region, whose limits are the minimally interesting effect size. According to the rules of inference by intervals, if the interval (confidence, credibility, or likelihood[9]) is contained within the null region, then the null region hypothesis can be accepted. If the interval is entirely outside the null region, then H1 can be accepted. If the interval spans both the null region and regions outside, then the data do not discriminate H0 and H1 (Dienes, 2014). The stopping rule to guarantee a clear conclusion is therefore to stop when the null region is either entirely within or entirely without the null region.

In the limit as null region shrinks to [0, 0] the method becomes significance testing, with all its faults (including the inability to assert the null hypothesis). The properties of the method must depend on how big the interval is relative to the null region is. If the maximum number of participants before stopping (MaxN) is 100 and the null region is NR[0, 0] (i.e. a point null is used and so the method is significance testing), then H1 is accepted 36% of the time when H0 is true (using the same experimental set up as in Table 1 and Appendix A). Table 3 shows error rates (estimated by 1000 simulations for each value) for NR[−0.1, 0.1]. Even with no restriction on the width of the interval, false alarms fall from 36% to 16% (first column) compared to significance testing. This is similar to the Bayes factor for a threshold of 3. However, unlike for a Bayes factor, H0 is never accepted when it is true. One hundred participants are just insufficient to get the interval small enough to fit into the null region (even though a *dz* of 0.1 may be quite satisfactory for supporting H1 for many researchers: A *dz* of 0.1 corresponds to an *r* of 0.49; and for the 100 studies replicated by the Reproducibility Project to date, 70 had an original effect size smaller than this, Open Science Collaboration, 2015, and were used as evidence for H1).

Table 4 shows what happens when MaxN is increased to 1000. Now H0 is often accepted when true, especially if we do not allow stopping to happen until the interval is smaller than a certain width. However, increasing MaxN increases error rates. This is unlike the case for Bayes factors. For Bayes factors, if MaxN is increased to 1000 then for a threshold of 10, for the null hypothesis, the rejection rate is 5% (same as for MaxN = 100) and the acceptance rate is 93% (up from the 69% for MaxN = 100; see Table 1). Increasing MaxN only improves things for Bayes factors, but increases false alarms for inference by intervals. Better error rates can be achieved for MaxN = 100 for Bayes factors than MaxN = 1000 for inference by intervals.

As for Bayes factors, ensuring a minimum number of trials has occurred before stopping improves error rates for inference by intervals. Table 5 shows the improvement for requiring 10 participants to be run before optional stopping occurs as compared to Table 4. Still the error rates are higher than those for Bayes factors shown in Table 2.

[9] Though see Morey et al. (in press) for why it should be, or correspond to, a credibility interval.

Table 5
Decision rates for accepting/rejecting H0 for NR[−0.1, 0.1] MaxN = 1000 MinN = 10.

	Minimum width:	None	10 * NRW	5 * NRW	4 * NWR	3 * NRW	2 * NRW	NRW	0.5 * NRW
Actual effect:									
$dz = 0$	Reject	9	8	7	5	2	1	0	0
	Accept	79	82	82	82	87	88	88	0
$dz = 1$	Reject	100	100	100	100	100	100	100	0
	Accept	0	0	0	0	0	0	0	0

Appendix B. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jmp.2015.10.003.

References

Arkes, H. (2013). The consequences of the hindsight bias in medical decision making. *Current Directions in Psychological Science*, *22*, 356–360.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., … Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119.

Berger, J. O., & Berry, D. A. (1988a). The relevance of stopping rules in statistical inference. In S. S. Gupta, & J. O. Berger (Eds.), *Statistical decision theory and related topics. Vol. 4* (pp. 29–72). New York: Springer Verlag.

Berger, J. O., & Berry, D. A. (1988b). Statistical analysis and the illusion of objectivity. *American Scientist*, *76*, 159–165.

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle*. Hayward, CA: Institute of Mathematical Statistics.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, *57*(298), 269–326.

Carp, J. (2012a). On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments. *Frontiers in Neuroscience*, http://dx.doi.org/10.3389/fnins.2012.00149.

Carp, J. (2012b). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*, *63*, 289–300.

Chambers, C. D. (2015). Ten reasons why journals must review manuscripts before results are known. *Addiction*, *110*(1), 10–11.

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered reports: realigning incentives in scientific publishing. *Cortex*, *66*, A1–A2.

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead of "playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. *AIMS Neurosci.*, *1*(1), 4–17.

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

Correll, J. (2008). $1/f$ noise and effort on implicit measures of bias. *Journal of Personality and Social Psychology*, *94*, 48–59.

Dienes, Z. (2011). Bayesian versus Orthodox statistics: Which side are you on? *Perspectives on Psychological Sciences*, *6*, 274–290.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781. http://dx.doi.org/10.3389/fpsyg.2014.00781.

Dienes, Z. (2015). How Bayesian statistics are needed to determine whether mental states are unconscious. In M. Overgaard (Ed.), *Behavioural methods in consciousness research* (pp. 199–220). Oxford: Oxford University Press.

Estes, Z., Verges, M., & Barsalou, L. W. (2008). Head up, foot down: object words orient attention to the objects' typical location. *Psychological Science*, *19*, 93–97.

Etz, A. (2015). Retrieved 30 September 2015. http://alexanderetz.com/2015/08/30/the-Bayesian-reproducibility-project/.

Fanelli, D. (2010). Positive results increase down the hierarchy of the sciences. *PLoS One*, *5*, e10068.

Feynman, R. P. (1998). *The meaning of it all*. Penguin Books.

Fisher, R. A. (1935). *The design of experiments*. Oliver and Boyd.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman & Hall.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Unpublished paper, Retrieved 1 July 2014. Available from: http://www.stat.columbia.edu/_gelman/research/unpublished/p_hacking.pdf.

Gelman, A., & Rubin, D. B. (1995). Avoiding model selection in Bayesian social research. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 165–173). Oxford: Blackwell.

Goldacre, B. (2013). *Bad pharma: how medicine is broken, and how we can fix it*. Fourth Estate.

Good, I. J. (1983). *Good thinking: the foundations of probability and its applications*. University of Minnesota Press.

Hoijtink, H. (2011). *Informative hypotheses: theory and practice for behavioral and social scientists*. Chapman and Hall/CRC.

Howson, C., & Urbach, P. (2006). *Scientific reasoning: the Bayesian approach* (3rd ed.). Chicago: Open Court.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124.

Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*(5), 235–241.

Jaynes, E. T. (2003). *Probability theory: the logic of science*. Cambridge, England: Cambridge University Press.

Jeffreys, H. (1939). *The theory of probability*. Oxford, England: Oxford University Press.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, *23*, 524–532.

Keppel, G., & Zedeck, S. (1989). *Data analysis for research designs: analysis of variance and multiple regression/correlation approaches*. New York: Freeman.

Kruschke, J. K. (2010). *Doing Bayesian data analysis: a tutorial with R and BUGS*. London: Academic Press.

Lanczos, C. (1974). *The Einstein decade (1905–1915)*. London: Elek Science.

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: a practical course*. Cambridge: Cambridge University Press.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*(1–2), 187–192.

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, *15*, 22–25.

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below 0.05. *Quarterly Journal of Experimental Psychology*, *65*, 2271–2279.

Morey, R.D., Hoekstra, R., Rouder, J.N., Lee, M.D., & Wagenmakers, E.-J. The fallacy of placing confidence in confidence intervals, *Psychonomic Bulletin & Review*, in press.

Morey, R. D., Romeijn, J., & Rouder, J. N. (2013). The humble Bayesian: Model checking from a fully Bayesian perspective. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 68–75.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406–419.

Neyman, J., & Pearson, E. (1967). *Joint statistical papers*. Hodder Arnold.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*, 1422–1425.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 943–951. http://dx.doi.org/10.1126/science.aac4716.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*, 531–536.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence?. *Perspectives on Psychological Science*, *7*, 528–530.

Popper, K. (1963). *Conjectures and refutations*. London: Routledge.

Registered Reports, Retrieved 27 August 2015. https://osf.io/8mpji/wiki/home/.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–367.

Rosenthal, R. (1979). An introduction to the file drawer problem. *Psychological Bulletin*, *86*, 638–641.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*, 301–308.

Rouder, J. N., & Morey, R. D. (2011). A Bayes-factor meta analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, *18*, 682–689.

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

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science*, *9*, 293–304.

Sanborn, A. N., & Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review*, *21*, 283–300.

Savage, L. J. (1962). *The foundations of statistical inference: a discussion*. London: Methuen.

Schoenbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2015). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, in press.

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

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

Slovic, P., & Fischhoff, B. (1977). On the psychology of experimental surprises. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 544–551.

Stone, J. V. (2013). *Bayes' rule: a tutorial introduction to Bayesian analysis*. Sebtel Press.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.

Vanpaemel, W. (2014). Theory testing with the prior predictive. *Oral presentation at the 26th annual convention of the association for psychological science, 22–25 May, San Francisco.*

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779–804.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science*, *7*, 632–638.

Yu, E. C., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2014). When decision heuristics and science collide. *Psychonomic Bulletin & Review*, *21*, 268–282.