# How to use and report Bayesian hypothesis tests

Article (Accepted Version)

# How to use and report Bayesian hypothesis tests

Zoltan Dienes

University of Sussex

Correspondence:

Zoltan Dienes

dienes@sussex.ac.uk

School of Psychology

University of Sussex

Brighton

BN1 9QH

UK

## Abstract

This article provides guidance on interpreting and reporting Bayesian hypothesis tests, in order to aid their understanding. To use and report a Bayesian hypothesis test, predicted effect sizes must be specified. The paper will provide guidance in specifying effect sizes of interest (which also will be of relevance to those using frequentist statistics). First, if a minimally interesting effect size can be specified, a null interval is defined as the effects smaller in magnitude than the minimally interesting effect. Then the proportion of the posterior distribution that falls in the null interval indicates the plausibility of the null interval hypothesis. Second, if a rough scale of effect can be determined, a Bayes factor can indicate evidence for a model representing that scale of effect versus a model of H0. Both methods allow data to count against a theory that predicts a difference. By contrast, non-significance does not count against such a theory. Various examples are provided including the suitability of Bayesian analyses for demonstrating the absence of conscious perception under putative subliminal conditions, and its presence in supraliminal conditions.

Hypothesis testing is an integral part of many disciplines. The details of how we have been testing hypotheses has come under criticism and indeed has been targeted as an explanation for why various disciplines now face a credibility crisis (e.g. Halsey, Curran-Everett, Vowler, & Drummond, 2015). Using Bayesian hypothesis testing has been proposed as addressing some of the problems (e.g. Wagenmakers, Marsman, Jamil, Ly et al., 2018). Bayesian hypothesis testing will only address problems, of course, if understood by both the author of a paper and its readers. This article will provide brief guidance on interpreting and reporting Bayesian hypothesis tests, in order to aid their understanding. To use and report a Bayesian hypothesis test, predicted effect sizes must be specified. The paper will provide guidance in specifying effect sizes of interest (which also will be of relevance to those using frequentist statistics, because specifying effect sizes of interest is inferentially crucial for any statistical system).

Statistical inference can be divided into estimation and hypothesis testing (e.g. Jeffreys, 1939). Estimation involves saying how big something is; for example, finding a mean and its standard error. Hypothesis testing involves comparing two or more hypotheses, typically the claim that there is no effect (H0) versus the claim that there is an effect of interest (H1). Estimation and hypothesis testing complement each other. Given doubt about something's existence, it is useful to use Bayesian hypothesis testing to determine the strength of evidence that it exists. Given minimal credibility that something exists, it is reasonable to ask how big it is. Indeed, estimates are required for hypothesis testing to proceed. Thus, estimation for any effect deemed credible enough to research is always required. This paper will discuss Bayesian hypothesis testing to indicate what key problem it can solve, and then give guidelines on how to report the results. Although the point of the paper is to discuss Bayesian hypothesis testing, this is not to diminish estimation: It will be presumed that there will also always be accompanying estimates (e.g. means and standard errors). If estimates should always be provided, why test hypotheses? If one only used estimation, one would not be in a position to conclude that the data indicated that an effect did or did not exist (Haaf, Ly, & Wagenmakers, 2019). This paper is aimed at researchers for whom the existential claim of whether

or not something exists (i.e. a population numerical relationship exists) is relevant to their research. One has, for example, at least implicitly tested hypotheses when claiming that people performed at chance, that there was no interaction, or when leaving a previously considered variable out of a model. That is, scientists routinely wish to test hypotheses.

The typical non-Bayesian method of hypothesis testing is significance testing. Significance testing leads to the claim that an effect was or was not significant. Fisher (1935) pointed out that the null hypothesis is never established by a non-significant result. According to Fisher, we can disprove H0 by a significant result, but not establish it by a non-significant result. This asymmetry is unfortunate, because it means significance alone does not allow us to test theories that predict a difference: a non-significant result does not count against such theories. Confidence intervals illustrate why not. The confidence interval is the set of population values we still accept as possible in the light of data. An equivalent way of saying a result is non-significant (where H0 is the claim of no effect) is to say zero (no difference) is in the confidence interval. If zero is in the interval, zero is still possible - but note so are all the other differences in the confidence interval. Any confidence interval indicates that many population effects are possible. So, given a non-significant result, there is no reason to accept specifically zero as the population value. Yet many researchers still groundlessly use non-significance to count against a theory that predicts a difference (e.g. Abelson, 1995; Aczel, Palfi, Szollosi, Kovacs, et al. 2018; Dienes, 2016; Greenland, 2017). Drawing strong conclusions for baseless reasons must surely contribute to bad science.

Bayesian inference provides a solution (for other solutions see: Colling & Szűcs, 2018; Lakens, McLatchie, Isager, Scheel, et al.., 2018; for arguments in favour of the Bayesian solutions, see Dienes, 2016; Wagenmakers et al., 2018). There are two overall Bayesian approaches for testing hypotheses that we will consider in turn: The relation of a posterior distribution to a null interval; and Bayes factors.

## The relation of a posterior distribution to a null interval.

**Background**

In estimating an effect (e.g. a mean difference), one can start from prior information or constraints, represented by a prior distribution of the effect. For example, if it is known the population mean cannot be less than 0 or more than 1, although sample estimates can be (e.g. the meta-d'/d' of Maniscalco & Lau, 2012, when Type I and II decisions are made on the same information), a uniform distribution [0, 1] (i.e. the claim that values between 0 and 1 are equally plausible using the dependent variable, and values outside these limits are not possible) can serve as a prior distribution that represents this constraint in a simple way. Such a uniform would appropriately ensure that the credibility interval did not extend into an impossible range. Or when estimating the number of hits and false alarms based on a small number of observations in a signal detection study, the estimated population numbers can be improved by vague priors (which can be implemented by adding a count to each cell, Barrett, Dienes, & Seth, 2013, p. 545). Or when testing patients, and one has only a few observations, the variance of normal people may be used to improve the estimate of a patient's variance (consider when it is known that patient variance is generally large, but based on two observations a patient seems to have small variance)[1].

The prior distribution represents prior constraints. In Bayesian statistics, how the data are postulated to be distributed is represented by the "likelihood function". When the prior distribution is combined with the data (as represented by the likelihood function), a posterior distribution is obtained, giving the probability density of different effects taking into account both the data and prior constraints (for details on how this works see Etz & Vandekerckhove, 2018). If one wished to build in minimal prior constraints, the prior can be a uniform approximating [-∞, +∞] (cf. Gelman,

1 In conventional statistics one can either use the variance from the patient as it is; or, one could use a pooled variance from the patient and from normals, given the assumption that the variances were the same. In the Bayesian case one need not make such strong assumptions, either way. For example, one could use a prior worth one observation from the normal's data. In summing the squared differences for the patient add in the normals' variance as one squared difference, contributing one degree of freedom, to calculating the patient variance. This weak prior (simple to implement with no special programs) may make the estimation of the patient variance more accurate with a small number of observations, yet would be appropriately swamped by real patient data as more came in. A computationally more complex and subtle method would be Bayesian hierarchical modelling.

Simpson, & Betancourt, 2017, for caution against automatically using a uniform distribution). Then the posterior reflects the data (constrained only by how their distribution is modelled in the likelihood function).

A parameter is a population value we wish to make claims about, for example a population mean difference. If a broad uniform prior is used for a parameter (e.g. a mean difference with known variance), and the likelihood function is a normal distribution, the posterior will be a normal distribution the same as the (frequentist) sampling distribution of the mean difference. With an unknown variance and some standard assumptions about the prior on the variance, the posterior distribution for the mean difference will again be the same as the sampling distribution of the mean difference. Thus, under these conditions, the Bayesian 95% credibility interval will be numerically the same as the 95% confidence interval. For more details on the process of Bayesian estimation in more interesting ways (i.e. with results that can usefully differ from frequentist methods) see Gelman, Carlin, Stern, Dunson, et al. (2013), Greenland (2006, 2007), Kruschke (2014), McElreath (2016); for an accessible introduction to Bayesian estimation, see Wagenmakers et al. (2018). For brief guidance on the details of computing Bayesian estimates see Krushcke (2013) v,an Ravenzwaaij, Cassey, and Brown (2018). The only further comments this article will make about estimation is in interpreting the posterior distribution, in order to show how one may test hypotheses.

The posterior distribution (of, say, a mean difference) enables one to calculate the probability of the mean difference lying between any two possible values. The probability of the population mean difference lying between the limits -3 and +5 units is the area under the curve of the posterior distribution between -3 and + 5 (i.e. the black area in Figure 1a). As the two limits come closer together (Figure 1b) the area under the curve becomes smaller (compare black areas of 1a to 1b). As the lower limit becomes the same as the upper limit, the area goes to zero. That is, the probability of any specific population value is zero, by this representation. (The area under a single point on the curve is just a vertical line, so the area is zero.)

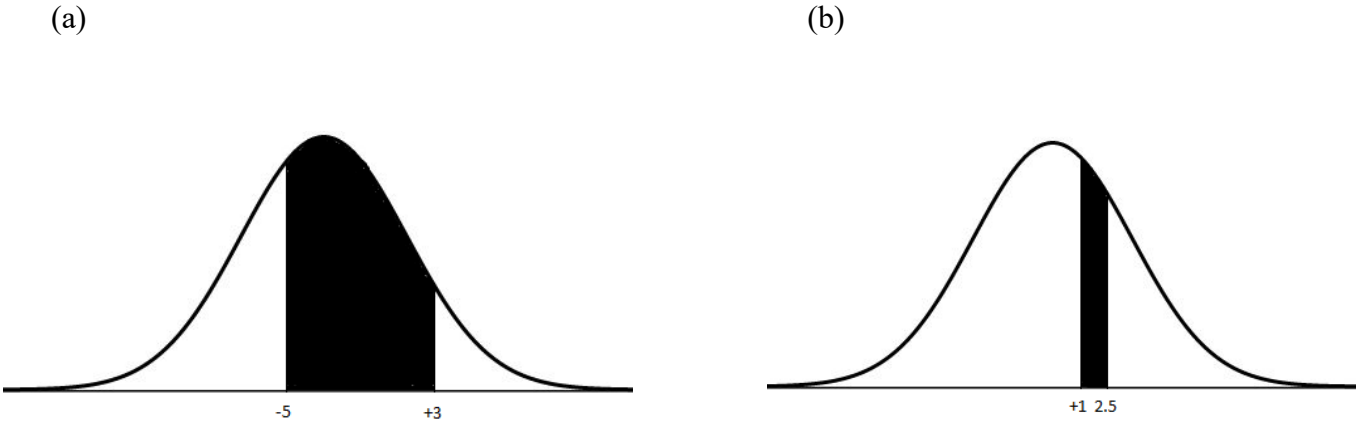



Figure 1

Area under the curve

The probability that the value is between two limits is the area under a probability density curve between those limits (the black area). Thus, as the limits come closer together (from (a) to (b)), the area shrinks, becoming zero for the probability of the value being precisely one value (e.g. precisely 2.1001).

Using the standard representations of a posterior distribution, for example a normal distribution or a t distribution, the probability of any particular population effect (such as zero, or no effect) is exactly zero; the area under that point is zero. Thus, whether or not the H0 value (no effect) is inside or outside a 95% credibility interval is irrelevant to that value's probability: There is always 100% certainty that there is some population effect (according to the posterior distribution,

no matter its shape or location). It is thus wrong to use the rule of whether or not zero is in a credibility interval to draw any conclusions about whether or not there is an effect.

How then can one draw inferences about whether or not there is an effect? There are two solutions. The first is to use a null interval hypothesis (a hypothesis that the population value lies between two limits, such as -0.1 and +0.1 units) rather than a point null hypothesis (a hypothesis that the population value is exactly one value, such as 0 units). If H0 defines an interval, then that interval can have a probability using standard posterior distributions; it is the proportion of the posterior distribution in that interval (that is, it is the area under the posterior curve between the limits defining the interval) (Greenwald, 1975; also see Kruschke, 2014). The second solution is to mathematically assign a probability to a point value, in order to represent the belief that the point could in fact be a good approximation to the truth - the option followed by Jeffreys (1939) in developing Bayes factors. We will first consider using a null interval.

**Crucial practical considerations**

In defining a null interval, the problem is to define a minimally interesting effect size, m, a magnitude below which the effect is too small to be relevant for theory or practice. The null hypothesis is then that the true population effect size is in the null interval [-m, +m]. No effect in that interval is privileged as more important than the others; if there is evidence for H0, there is evidence for the claim that the true effect is somewhere in that interval (i.e. it is not evidence that the effect is precisely zero). Kruschke (2014, 2018) suggests calculating a 95% CI (Credibility Interval, or Highest Density Interval, HDI, as it can also be called, given standard ways of calculating CIs) and using the rule: If the 95% CI is completely inside the null interval, accept H0; if completely outside the null interval, reject H0; otherwise suspend judgment. This would largely correspond in practice to Greenwald's (1975) subtler rule of determining if 95% of the posterior distribution is inside or outside the null interval. In following such a rule, two choices are made: One is the amount of evidence or plausibility one takes as good enough, which is determined by

whether one uses a 95% CI, a 90% CI etc. Kruschke recommends a 95% CI; this can be a convention for standard situations. In frequentist equivalence testing, in effect a 90% confidence interval is recommended (Lakens, Scheel, & Isager, 2018). Note that the choice of the x% used for the CI does not reflect any step change in the information a posterior distribution provides; the choice is just to allow decision making. The other choice is in specifying the predictions of one's theory: What counts as a minimally interesting effect size (Kruschke, 2018)?

Raw effect effect sizes are usually what are of theoretical or practical relevance. For example, if considering an intervention to lose weight, any effect less than 0.5 kg over one month might be regarded as too small to be of clinical interest. A raw effect size is an effect in a unit of measurement, as in this case, kg. A person interested in losing weight is not interested in how much noise is in the scales measuring them (the noise affects standardized effect sizes); just in the kg's lost over a period of time (Ziliak & McCloskey, 2008). A standardized effect size (such as Cohen's d , Pearson's r, or a standardized regression coefficient) depends on how noisy the measuring apparatus is: How many questions in a questionnaire, how many trials on a cognitive task, and so on. Having ten times as many trials changes Cohen's d (by participants) by about a factor of 3 (i.e. square root 10). Putting covariates and other factors in an analysis may change standardized effect sizes by reducing noise. So the first point is that one cannot pluck from the air an effect size of, for example, Cohen's d = 0.1, and say that is the minimally interesting effect size for all effects. It is a scientific matter what the minimally interesting effect is, and so there can be no general statistical solution. Any general statistical assumption about what a minimally interesting effect size is must be false in an indefinite number of scientific cases.

Because the inference that there is no effect depends crucially on the possible size of the effect, that size should be specified for objective reasons: That is, the numerical value for the effect should come from a public place one can point to (for example, data), so that other people can criticize the reason for choosing that value. The reasons should relate back to the theoretical or practical context. Back engineering a minimal interesting effect size from the number of

participants used in past studies is not informative, as it presumes the past studies determined their number of participants based on good reasons for detecting a certain a minimally interesting effect size. The problem is only pushed back. A committee deciding on a smallest effect of interest is useful in so far as the committee gives its reasons so they can be criticized and the estimate improved. Thus, black boxes like committees, or an expert's opinion, also push back the problem of what would constitute a good reason. Once a number has been chosen, ideally from a public place, judgment is needed at that point to indicate that the reasons are provisionally good enough (a judgment that must apply to every aspect of a statistical model, whether frequentist or Bayesian, or anything else). We now go through some heuristics for obtaining a minimally interesting effect size.

**i) The end user heuristic: The judgment of an end user.** In applied research what matters is whether the outcome is good enough for the end user. There have been a number of examples of the judgment of end users defining minimally interesting effect sizes. Taking the end users of a depression treatment to include the treating psychiatrists, Leucht et al. (2013) found a clinician's impression of 'no change' corresponded to a Hamilton rating scale change of 3 units, and an impression of 'minimally improved' to 7 units; this has motivated the use of 3 units and smaller on the Hamilton as defining an area of clinical irrelevance (Moncrief & Kirsch, 2015). Leucht et al. (2006) applied a similar methodology to schizophrenia. Taking the end users of a depression treatment to be the patients, Button et al. (2015) analysed studies that had asked depressed patients if they felt the 'same', 'better' or 'worse' after treatment. They estimated that the Beck Depression Inventory reduced by 17% as a criterion for patients moving from a "same" to a "better" response. Thus, 17% on the Beck Depression Inventory can also be taken as a minimal clinically relevant effect size for an intervention aimed at treating depression.

The end user can also judge if a change in pain is noticeable or meaningful. Dworkin et al. (2008) found that reductions in pain of approximately 10-20% were noticeable, and reductions of approximately 40% were judged as meaningful. Kelly (2001) found that the smallest change

associated with the judgment of feeling “a little better” or “a little worse” was 12 mm on a 100-mm visual analogue scale of pain intensity (this is very similar to the 10-20% difference obtained by Dworkin et al. as noticeable).

Anvari and Lakens (2019) applied the end user heuristic to affect as measured by the widely used Positive and Negative Affect Scale (PANAS; Watson, Clark, & Tellegen, 1988). Participants rated affect on both Wednesday and Friday (using a Likert scale going from 1 = “very slightly or not at all”, to 5 = “extremely”). On Friday they were also asked to indicate if their affect had changed a little, a lot, or not at all. When people indicated their affect had changed “a little”, the average change in Likert units was 0.3 scale points. Thus an intervention to lift mood might be regarded as effective only if it could lift mood by an amount regarded as noticeable, i.e. by more than 0.3 Likert units on the PANAS. This method could in principle be generalized to different judgments (for example, how mindful the person feels).

Sometimes a theory may indicate who a relevant end user is. Accordingly, Lakens, Scheel, and Isager (2018) suggest that a just noticeable difference may be an appropriate minimally interesting effect size in some contexts. For example, if women's faces become redder during the fertile phase, that result does not support an evolutionary signalling function of facial redness if the change in redness is below the just noticeable difference (jnd) for male observers. Thus, a one jnd may be a minimally interesting effect size. In this case, one may be able to refine the minimally interesting effect size by determining the degree of change in facial redness that was just enough to prompt men to act in an appropriate way (thereby using the calibration heuristic below), which may be more than one jnd.

**ii) Calibration.** Dienes (2014; supplemental data- Appendix 1, example 2) showed how one measure, for which we do not have a relevant interesting effect size, can be regressed against another, for which we do, in order to calibrate the former. For example, previous research has shown a cognitive task is related to depression. A study plans to investigate whether performance

on the task is changed by a treatment for depression, rTMS. Task performance may be a reaction time difference measured in milliseconds. A minimally interesting effect size may be obtained by a norming study in which the task is regressed against the Hamilton; the norming study may be past research or a new study. Then find what change in task performance is predicted by a 7 unit change in the Hamilton (taking 7 units as a minimally interesting effect). In this example, one also needs to consider the rTMS regime: Only if it were the same dose as a patient would receive in actual treatment would one take 7 units on the Hamilton as the calibrating amount. If the rTMS dose were half as strong (e.g. applied for half as many occasions), the simplest approach would be halve the expected effect, and thus use 3.5 units on the Hamilton as the minimally interesting effect size.

**iii) Checking the roughly smallest plausible value is still theoretically relevant**. For conceptual and direct replications using same dependent variable, where there have been past studies, one can look at the lower limit of 95% CI of the raw effect (cf. Perugini, Gallucci, & Costantini, 2014), and check if it is it still theoretically or practically interesting. If so, this value, for which the population effect is very probably greater (with probability 97.5%) , could be conventionally taken as a smallest interesting effect which is just plausible. For example, consider a study investigating whether a four-week mindfulness of walking intervention changes mindfulness on a 1-5 Likert scale. If past research using a four-week mindfulness of breathing intervention (i.e. a slightly different procedure) changes rated mindfulness on a 1-5 Likert scale by 0.5 units, with a lower limit of the 95% CI of 0.3 units, one can judge if 0.3 units is still meaningful. Admittedly this may be difficult, and the heuristic therefore unhelpful; but often the judgment that an effect is meaningful is easy if the effect is sufficiently large. For example, previous shorter two-week mindfulness interventions may have produced changes of 0.2 units in mindfulness, yet had other positive consequences. We just have to judge that 0.3 is interesting, not that it is only just interesting. If 0.3 units is interesting, it can be taken as the minimally interesting effect size.

The smaller the minimally interesting effect, m, (decided by any of the above methods) the easier to obtain a 95% CI outside the null interval [-m, +m]; but the harder to obtain a 95% CI inside the null interval. If the theory predicts a difference, a test of the theory is only severe if the probability of obtaining a 95% CI in the null interval is high. Thus, the null interval must be wide enough to allow severe testing (cf. Mayo, 2018; Popper, 1963). For example, in planning a study one may work out a sample size such that if the null region hypothesis were true, with the number of participants used, at least 80% of the time (or 90% etc) the 95% CI would fall within the null region. (Such planning has no effect on the conclusions reached when the data are in, but it is useful for working out what number of participants may be needed to test a theory properly.)

These methods cover many cases but far from all. Objective reasons for a minimally interesting effect can be hard to come by; but unless such reasons exist, the analysis does not connect to the theory. In the absence of objective reasons for a minimally interesting effect size, or if the minimally interesting effect size is unacceptably small for severe testing, one can consider if Bayes factors can be motivated for the analysis.

**Writing up**

For the APA guidelines on reporting Bayesian estimation, see Appelbaum, Cooper, Kline, Mayo-Wilson et al. (2018, p. 20); and for detailed advice on reporting Bayesian estimation see Depaoli and van de Schoot (2017). For the sake of argument, I will presume that a uniform prior was used, and the 95% credibility interval is numerically the same or very similar as the 95% confidence interval, in order to focus on hypothesis testing, and show how easily it may be done without complicated software, but that care is needed in considering the null region.

Always indicate what prior was used, and the form of the likelihood function. Give the mean and standard error of the estimate (i.e. mean and standard deviation of the posterior distribution, or other measure of central tendency and spread). It may be useful to give the 95% CI. Specify the minimal interesting effect size making clear what the reasons are for that value so the reader can

easily evaluate the reasons. Finally, indicate the relation between the posterior distribution and the null interval, for example whether the 95% CI falls within or without the null interval (Kruschke, 2014). A Robustness Region, RR, can also be provided. The Robustness Region is the set of possible minimally interesting effect sizes that lead to the same conclusion (see the next example).

Here is an example: "The minimally interesting effect size in affect was derived from Anvari and Lakens (2019), who found a difference of about 0.3 units was noticeable as a small change in affect on PANAS. Thus, the null interval was defined as [-0.3, +0.3 Likert units]. For all tests a Robustness Region is given, notated as RR [min, max], namely the set of possible minimally interesting effect sizes that lead to the same conclusion using the decision rules provided by Kruschke (2018), described earlier. First, the compassion meditation group was compared with the active control group (indifference meditation). For positive affect as measured by the PANAS, the mean for the compassion group was 3.4 (SE = 0.2) Likert units, and the mean for the control group was 2.6 (SE = 0.2) Likert units. A uniform prior was assumed on the difference score, with a t-distribution likelihood, so the 95% confidence interval was numerically the same as a Bayesian 95% credibility interval. The difference was 0.80 (SE = 0.35) Likert units, 95% CI [0.1, 1.5 Likert units]. The 95% CI spans both interesting values (i.e. > 0.3 units) and null interval values, so the null interval hypothesis cannot be rejected, RR [0.1, 1.5 Likert units]. No conclusion as yet follows about the clinical effectiveness of compassion meditation compared to the active control. Notice the Robustness Region includes values rather different than 0.3 in either direction, so the conclusion is robust. The robustness in this case is reassuring because Anvari and Lakens (2018) estimated noticeable changes in a rather different context, and of course with some noise (for positive affect in their study, a judgment of "little changed" had a mean change in PANAS of 0.27, SE = 0.04 Likert units)."

## Bayes factors

**Background**

A Bayes factor compares how probable the data are on one model (e.g. H1) compared to another model (e.g. H0), and thus provides the strength of evidence for the one model rather than the other (see Dienes 2014; Etz, Haaf, Rouder, & Vandekerckhove, 2018, for introductions to Bayes factors). The Bayes factor reflects the principle that data most supports the theory that best predicts it (Morey, Romeijn, & Rouder, 2016). The models specify how probable different population effects are. We are free, for example, to say H0 predicts just one population effect, no effect. Thus, no minimal interesting effect need be specified (though it could be; Morey & Rouder, 2011; Palfi & Dienes, 2019a). Because the evidence for one model rather than another depends on what the models are, the model of H1 should reflect the theory being tested (just as the minimally interesting effect size needed to reflect theory). For guidance on easily computing Bayes factors using an online calculator, see Dienes (2014). Conventions have also been suggested for the meaning of different values of the Bayes factor. Lee and Wagenmakers (2013) suggest treating greater than 3 as moderate evidence for H1 rather than H0, and, by symmetry, less than 1/3 as moderate evidence for H0 rather than H1; greater than 10 as strong evidence for H1 rather than H0, and less than 1/10 as strong evidence for H0 rather than H1. For journals that use a 5% significance level, a convention of 3 (and therefore 1/3) would be suitable for a Bayes factor to reflect about the amount of evidence shown by $p < .05$ when evidence favours an H1 over H0, though there is no monotonic relation between Bayes factors and p values (Lindley, 1957). Further, Bayes factors are continuous measures of evidence; thresholds are convenient when decisions are made but such thresholds do not reflect any step changes in the evidential value of a Bayes factor.

**Crucial practical considerations**

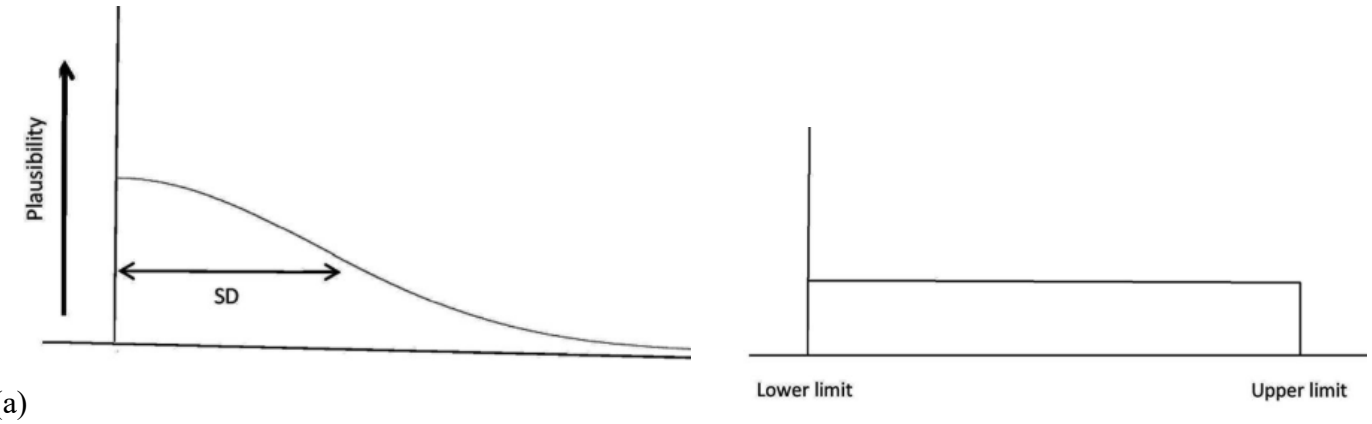
The model of H0 assumed by most Bayes factor calculators is that only one value (e.g. zero) is possible as the value of the population parameter (Dienes, 2008; Rouder, Speckman, Sun, Morey et al., 2009; van Doorn, van den Bergh, Bohm, Dablander, et al., 2019). Assuming a point H0 is

consistent with significance testing. What the typical Bayes factor requires that is new for people who are used to significance testing is a model of H1 (often called a prior), i.e. a representation of what the theory predicts. The model of H1 is a probability density function over different possible population effects (i.e. it represents how plausible different population effects are, given the theory). This may take any form the researcher can argue best represents theory in a simple way. This prior serves a different purpose from that used in estimation and therefore will rarely be the same. The prior distribution used in estimation has the purpose of allowing the most accurate estimate of parameters. The purpose of the model of H1 is to represent the predictions of a theory. If one used the model of H1 as a prior distribution for estimation, the estimates would be pulled towards the predictions of the theory, and it would become harder to have outcomes count against the theory. The test would not be severe (it would be hard to show the theory false when it was false). The function of the model of H1 is not to represent what the data say; it is represent what the theory says so that it may clash with the data if the theory were wrong. Thus, to avoid confusion, I do not call the model of H1 a prior[2].

A typical model of H1 is a normal (or Cauchy) distribution centred on zero, or such a distribution with the half below zero removed, as illustrated in Figure 2a. The half-normal distribution represents a theory making a directional prediction (where by convention the theory-predicted direction is represented as positive). It represents in a simple way that smaller effects are more likely than larger effects. One may justify this assumption by noting that published effects often over-estimate true effects (e.g. Open Science Collaboration, 2015), so if the predicted scale of effect is derived from published studies, smaller effects would be more plausible than larger ones.

2 A third referent of the term prior is the prior strength of belief in H1 vs H0. Not distinguishing these three referents (estimation prior, model of H1, prior odds of H1 vs H0) can lead to misplaced criticisms of Bayes factors,. For example, consider the argument that as how much one believes in subliminal perception is a subjective matter, Bayes factors are too subjective. But the model of H1 is a specification of what size effects are predicted by a hypothesis of subliminal perception, not an indication of how much one believes in that theory. (The prior belief in a theory can be relevant to scientific inference, for example when scientists hold very diverse judgments of plausibility, a study may plan for a more evidential Bayes factor than would otherwise be typical. But that is a different matter than how a Bayes factor is calculated.) Or consider the argument: Because the Bayes factor is sensitive to the prior (i.e. the model of H1, what predictions a theory makes), but the estimated mean is largely unchanged by variations in a vague prior, Bayes factors are a bit erratic, and we should just estimate. But of course, evidence for a theory should depend on what predictions it makes (Jeffreys, 1939).

Indeed, even when following up one's own unpublished work, one may be especially drawn to effects that have been over-estimated (Albers & Lakens, 2018; Dienes, 2017). The standard deviation of the half-normal (or normal) distribution represents the scale of effect postulated to exist. The task is to specify the scale in a way relevant to one's theory. One may do this in two ways. Either one specifies with reasons the sort of scale that characterizes the population effect, and the standard deviation is set to this number. Or one might have reasons for specifying the rough maximum population effect, then one sets the standard deviation to half the maximum (on the convention that a rough maximum for a normal distribution is two standard deviations out)[3]. If there were reasons for setting a maximum, and for values between zero and that maximum to all be roughly equally plausible, a uniform distribution can be used, as in Figure 2(b).




(a)

(b)

Figure 2

Possible models of H1

Two distributions that are often used as models of H1. The y axis is the plausibility of the effect given the theory; and x axis is the population effect size. (a) A half-normal distribution, which typically has a mode at zero (no effect), and thus requires setting only the standard deviation to the

[3] A Cauchy (or half-Cauchy distribution) can also be used; the Cauchy is a bell shaped distribution like the normal distribution but with thicker tails. It is useful when the rough maximum value expected is about 7-10 times the rough scale expected.

scale of effect size predicted. (b) A uniform distribution, which typically has the lower limit set to zero (no effect), and thus requires setting only the maximum.

The evidence for a theory depends on what sort of effect it predicts. One might think that the data itself gives the best estimate of the sort of effect that could be expected in the precise context used. However, if the effect size obtained is used to determine the effect size the theory predicts (by being used as the standard deviation of a half-normal distribution, for example), the obtained effect has been double counted: once to make a prediction; and, second, to test that prediction. Double counting is not legitimate in Bayes factors (Lindley, 1991). An indication of the problem of double counting is that such a test would not be severe (double counting makes it hard to find the theory wrong). Further, when double counting, the theory does no work in making the prediction. We now go through some heuristics for obtaining a relevant effect size for scaling predictions of a theory.

**(i) Direct replication.** In attempting to replicate a study, the effect found in the original study can be used as, for example, the standard deviation of a half-normal distribution for the model of H1 for the replication attempt (see e.g. Dienes & McLatchie, 2018, for worked examples). Ly, Etz, Marsman, and Wagenmakers (2019) present a related method. The theory tested is that implied by any empirical paper: That the methods of the original study (as given in the Methods section) describe a procedure for obtaining the sort of effect obtained (as reported in the Results section).

In the absence of a previous study, it might be thought that the same logic could be employed with a single study, by taking a random sample from it to act as an "original" estimate (e.g. O'Hagan, 1995). That is, use a proportion of the data (e.g. $\sqrt{N}$) to estimate the effect for modelling H1; then the rest of data to test H1 against H0. One could take many such samples, and average the (log) Bayes factors. This procedure may seem tempting as it removes all need to specify what a theory predicts. But that temptation shows the problem: There is no substantial theory being tested (Morey, Homer, & Proulx, 2018). And if there is no theory, and thus no independent

expectation of the effect size should an effect exist, there is no motivation to test a particular H1 versus H0. Without a clearly motivated model of H1, it may be appropriate to simply estimate parameter values. The difference from a replication of a previous study is that in choosing to replicate a study, the original study is thereby taken as scientifically interesting. Thus, the theory that the method reported produces the sort of effect reported is an interesting theory to test.

**(ii) Conceptual replication.** A study may be a conceptual replication of an original study. In this case, one may often take the effect from the original study as the scaling factor, just as in the case of the direct replication. The theory being tested is that the phenomena studied in both experiments belong to the same class.

For example, imagine that a study shows a subliminal priming effect using back masking on a reduced contrast image (e.g. Armstrong & Dienes, 2014), where participants chose a primed option by an amount 5% above a chance baseline, SE = 1.5%, N = 35, $p < .05$. The procedure is repeated but this time using gaze contingent crowding (Faivre, Berthet, & Kouider, 2012), in this case displaying the prime for about the same time as the original study. The priming effect is now 1% ,SE = 2%, N = 35 (imaginary data). Is there evidence for any subliminal perception using the new method? The outcome is non-significant ($t = 0.5$), but non-significance in itself does not mean anything. In the absence of a theory indicating why back masking should give a bigger or smaller effect, an approximate scale of effect is provided by the original study. Thus H1 could be modelled as a half-normal distribution with a standard deviation of 5%. Using the Dienes (2008) calculator, the Bayes factor is 0.56[4]. The value is quite close to 1; the non-significant result in this case is non-evidential. Thus, nothing follows about whether or not there was subliminal perception with gaze contingent crowding for these stimuli. More data are needed.[5]

---

4 In the Dienes (2008) calculator ( http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/Bayes.htm), enter the "sample mean" as 1, the "sample standard error" as 2, say "no" to Uniform, enter 0 for the mean for the normal, 5 for the SD, and 1 for the tails (i.e. making it half-normal).

5 A default Bayes factor on the gaze contingent crowding results (a "JZS" Bayes factor with scale factor = 0.7; http://pcl.missouri.edu/bayesfactor) is 0.20, i.e. moderate evidence for H0. However, this assumes a default scale factor of Cohen's dz of 0.70; in the original experiment, Cohen's dz was 0.08. While the default model of H1

**(iii) Basic effect heuristic.** The basic effect heuristic is to find basic effect and use that as estimate of scale of effect for intervention aimed at changing that effect. When investigating whether an intervention moderates an effect, the size of the basic effect itself may provide a scale appropriate for expecting how much the effect could be altered: A two-second effect may be modulated by a greater amount than a 50 ms-effect. Dienes, Baddeley and Jansari (2012) were interested in how much mood may change the amount of implicit learning. They used a new implicit learning paradigm, so there was no prior information relevant to that particular paradigm about expected effect sizes. Dienes et al. ran a pilot to measure only the amount of learning, without any mood manipulation. Then in the main experiment looking at the effect of mood on learning, the amount of learning found in the pilot was used to scale expectations of effects of mood (by being entered as the standard deviation in a half-normal distribution for the H1 modelling mood effects). See also Ziori and Dienes (2015) for use of the basic effect heuristic. In these cases, the main theory tested was supplemented by the auxiliary hypothesis that an intervention will have an effect somewhere between 0 and twice the basic effect, with smaller effects being more likely.

**(iv) Calibration with other data.** Dienes (2014; supplemental data- Appendix 1, example 2 looking at meta-cognition) showed how one measure, for which we do not have a relevant interesting effect size, can be regressed against another, for which we do, in order to calibrate the former. For example, Dienes (2015) describes how calibration can be used for testing whether knowledge is unconscious. A common schema for establishing subliminal perception by an objective threshold is: "Priming = X%, $p < .05$; identification = Y%, $p > .05$. Therefore there was subliminal perception." Such reasoning depends on asserting the null hypothesis for identification of the stimulus. But non-significance is not grounds for asserting H0. And obtaining a minimally interesting effect size for identification is difficult. But there is a crucial empirical question that can

produces an answer (0.20) not far from the informed model of H1 in the text (0.56), there is a reason for preferring the latter; namely, it is informed by the relevant scientific context.

be answered that allows Bayes factors to be used: What level of identification would be expected for the amount of priming obtained, if the knowledge were conscious?

Dienes (2015) recommended running a norming study in which the perception was conscious; for example, under conditions where the participants say they saw the stimulus (with some degree of clarity or confidence). If the level of priming was about that observed in the putatively subliminal case, then the level of identification in the conscious study indicates the rough level of identification expected in the putative subliminal study, if the perception were actually conscious. But it is unlikely the levels of priming would be the same. In that case, on a graph of identification performance against priming performance, plot the means of the conscious condition as a point. The theory being tested is that there is only one perceptual knowledge base underlying performance on identification and priming (e.g. conscious perception). So as identification goes to zero (or whatever is the level of chance performance), so does priming, on this theory. So draw a line from the plotted point to (0,0) (see Figure 3). Now read off from the line, given the level of priming in the putative subliminal study, what level of identification would be expected, if there were only conscious perception[6] (see Rouder, Morey, Speckman, & Pratte, 2007. for another Bayesian approach to subliminal perception)

6 Additional assumptions include (a) the relation is linear with the dependent variables used; so introducing a level of conscious perception that is not too high is useful for approximating this assumption; and (b) identification measures uniquely conscious perception, which on many theories of consciousness would be false (Dienes & Seth, 2010).

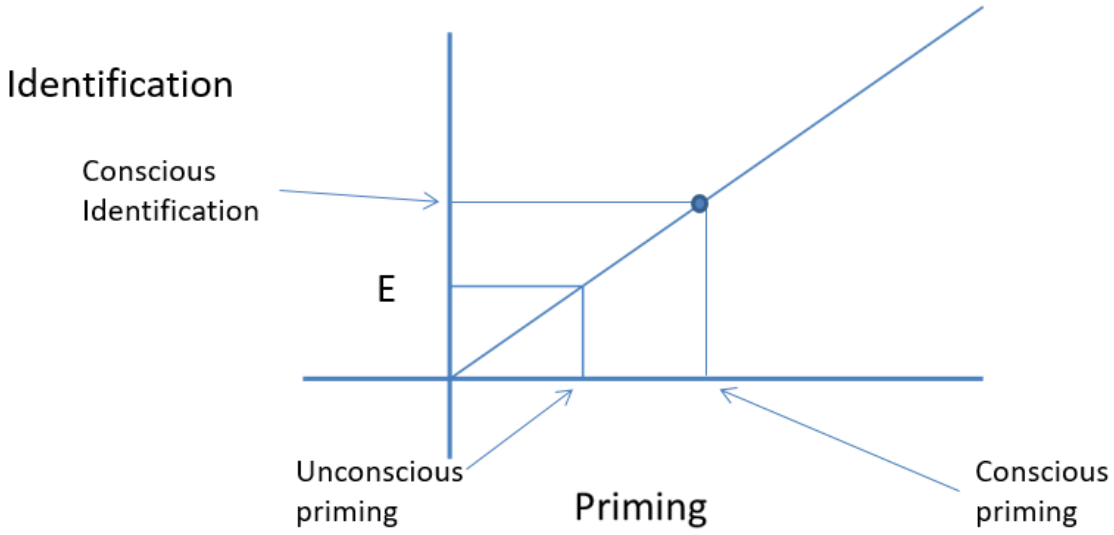



Figure 3

Calibrating the predicted level of identification against priming

In order to test whether or not perception is conscious in a putatively subliminal condition, one needs to know what level of conscious perception would explain the priming in the subliminal condition. Run a conscious condition and plot (conscious mean identification, conscious priming), the circle in the figure. Draw a line from there to (chance, chance). From the level of (putatively) unconscious priming, read off the expected level of identification, E. Use a Bayes factor to test the identification of people in the subliminal condition by modelling H1 as a half-normal distribution with standard deviation = E.

When using a past published study as the norming study, consider how performance is tested. For example, imagine a study by Lee finds evidence for priming, but not for conscious knowledge on a recognition test (e.g. Cohen's $dz = 0.16$, $t(19) = 0.56$). You find a study by Smyth using the same paradigm with about the same level of priming, and it did find evidence for conscious knowledge on a recognition test ($dz = 0.6$, $t(39) = 3.68$). Using the Rouder et al. (2009)

Bayes factor, scaled in Cohens *dz*[7], the Bayes factor is 0.30, moderate evidence for H0, i.e. for no conscious perception. That sounds as if implicit processing has been shown in the Lee study. But say Lee used only one trial for the recognition test achieving 58% accuracy, and Smyth used 48 trials, achieving 60% accuracy. Population Cohen's *dz* depends on the number of trials; thus, using Smyth's Cohen's *dz* as the expected effect size for the Lee study would be incorrect. One could adjust the expected Cohen's *dz* according to number of trials. But if one used raw units, such as percentage correct, the second study's effect size could be directly used as the predicted effect for the first (ceteris paribus), without assuming equivalent standard deviations between the studies; in this case, leading to a Bayes factor close to 1 as the obtained effect, 58%, is close to the scale of effect predicted, 60%, but with a large standard error (see Dienes, 2017). Using raw effect sizes as the basis for predictions motivates researchers to reduce noise in measurement; uncritical use of standardized effect sizes can motivate noisy measures when one wants to obtain evidence for H0 (especially if using a default Bayes factor, i.e. one with a fixed scale factor regardless of context).

**(v) Heuristics for when there are no relevant prior studies.** Dienes (2019) presents heuristics that can be used when there is not another study to inform expectations, different heuristics that can be used for mean differences, ANOVA, regression or mediation. An example for testing differences is the room-to-move heuristic: If a theory predicts that A will be smaller than B, then if B is above chance by an amount r, then A has to be between chance and r; r is the room to move. Dienes (2015) applies this heuristic in a couple of cases. For example, meta-d' cannot be larger than Type I d' (when Type II and Type I decisions are made on the same information); thus, Type I d' can be set as a maximum in modelling H1 for meta-d' (for an application in implicit cognition see Leganes-Fonteneau, Nikolaou, Scott, & Duka, 2019). Another example is using the Perceptual Awareness Scale (PAS) of Ramsøy and Overgaard (2004). The scale has four levels from PAS1 to PAS4, of increasing clarity of conscious visual experience. Given the theory that increasing conscious clarity

7 http://pcl.missouri.edu/bf-one-sample

is associated with increasing ability to discriminate (a theory corroborated by the numerous experiments that have since used PAS), the discrimination shown at one level of PAS can be set as the maximum that could be achieved at the level just below.

In planning a study one can work out how many participants would probably be needed to achieve given Bayes factor thresholds, assuming a model of H1 (with a scale factor determined by any of the above methods). Schönbrodt, Wagenmakers, Zehetleitner, and Perugini (2017) and Schönbrodt and Wagenmakers (2018) describe how to conduct simulations to achieve desired probabilities. Although these probabilities bear a similarity to power calculations in frequentist statistics, there are key differences. Unlike the role of power in significance testing, the probabilities play no role in final inference when it comes to Bayes factors, because the Bayes factor itself indicates the evidence for H1 over H0. Further, in setting up a model of H1, the scale factor is the scale of effect predicted, not the minimally interesting effect size. For frequentist power, one should use the minimal interesting effect size to calculate power. The role of working out a required likely number of participants is purely pragmatic, to consider if one has the resources to severely test a theory. A very quick way of calculating an approximate likely number of participants, which involves no simulations, is provided by Palfi and Dienes (2019b, version 3, Table 1). One can use as a stopping rule to stop when a certain threshold for the Bayes factor has been reached (e.g. "greater than 6 or less than 1/6"). When one specifies a model of H1, the Bayes factor is the evidence for that H1 over H0, no matter the stopping rule, or even when the model was specified (e.g. Dienes, 2016; Rouder & Haaf, 2020).

**Writing up**

For the APA guidelines on reporting Bayes factors gives three requirements, see Appelbaum et al. (2018, p. 20). The first requirement is to "specify the models being compared". Dienes (2014, 2019) introduces the following conventions. Use $B_{N(m, sd)}$ to indicate H1 was modelled using a

Normal distribution with mean m and standard deviation sd; $B_{HN(0, sd)}$ to indicate H1 was modelled with a half-normal distribution with mode 0 and standard deviation sd; $B_{C(m,s)}$ to indicate H1 was modelled with a Cauchy with mode m and scale factor s; $B_{HC(m,s)}$ to indicate H1 was modelled with a half-Cauchy with mode m and scale factor s; and $B_{U[min, max]}$ to indicate H1 was modelled with a uniform distribution with minimum min and maximum max.

One should also justify why one set the scale factors etc to the values one did. As the evidence depends on the tested predictions, the predictions must be relevant to one's theory (Vanpaemel, 2016). Not all theories make the same predictions, so there can be no default Bayes factor to be used in all cases. Conversely, a purely subjective Bayes factor based on each person subjectively determining the predictions of each theory is not a satisfying basis to provide a scientific test of a theory. Instead predictions should follow from theories using auxiliaries that are either well tested or simple, and publicly available to criticism.

The next APA requirement is to "report the Bayes factors and how they were interpreted." A decision may be made based on the Bayes factor, for example, a decision that a confound was controlled so the next study can move on from that issue; a decision that an intervention did not work so one theory will not be pursued further but another will be; a decision that an effect does exist so it is worth pursuing in a further study which tries to manipulate it. Decisions depend on not just the evidence for the alternatives (as represented in Bayes factors) but also the prior probabilities of the alternatives and the utilities of different outcomes. Given that the prior probabilities and utilities involved are vaguely defined, and assuming they do not vary much amongst alternatives (e.g. for different theories in cognitive psychology), it is often useful to keep them implicit and have a convention for what standard of evidence would motivate different decisions. In a context where significance has been set at 5% for decisions, the roughly equivalent amount of evidence would be a Bayes factor greater than 3 or less than 1/3. For example, one may conclude that two conditions produce the same outcome if the Bayes factor is less than a third, and thus decide that a confound has been dealt with. Such a standard of evidence has been criticized in some contexts (e.g.

Benjamin, Berger, Johannesson, Nosek, et al., 2018). Pragmatically, one can see what the journal one is submitting to requires (e.g. 6 and 1/6 for Registered Reports at *Cortex*). (Where there is a genuine concern about very different prior probabilities amongst researchers, for example when dealing with parapsychology, and very different utilities for different decisions, for example in considering plausible severe side effects of an intervention, these can be explicitly taken into account in interpreting Bayes factors, Lindley, 2014).

The final APA requirement is to "test the sensitivity of the Bayes factors to assumptions about prior distributions [i.e. models of H1]." A convenient way of determining the robustness of one's conclusions is with a Robustness Region (Dienes, 2019). Find all the scale factors that lead to the same conclusion. For example if one's preferred model of H1, using a half-normal distribution, led to the Bayes factor being above 3, the relevant threshold, then report all standard deviations for which the Bayes factor is above 3, as the interval [minimum, maximum]. This is illustrated below[8]. If the Robustness Region includes most standard deviations that are scientifically plausible, then the conclusion is robust.

For readers new to Bayes factors it is useful to explain how they work at the beginning of the Results section. Here is an example: "Bayes factors (*B*) were used to assess strength of evidence (Wagenmakers, Verhagen, Ly, Matzke, D. et al., 2017; see Dienes, 2015, and Sand & Nilsson, 2016, for the relevance of Bayes factors for implicit cognition). A *B* of above 3 indicates "substantial" (Jeffreys, 1939) or, better, "moderate" (Lee & Wagnemakers, 2013) evidence for the alternative hypothesis (H1) over the null hypothesis (H0); thus by symmetry a *B* below 1/3 indicates substantial (/moderate) evidence for H0 over H1 ("substantial" in the sense of just worth taking note of.) *B*s between 3 and 1/3 indicate the data collected do not sensitively distinguish H0 from H1. Thus we will report that there was no effect when $B < 1/3$. Here, $B_{HN(0, x)}$ refers to a Bayes factor in which the predictions of H1 were modelled as a half-normal distribution with an SD of x where x scales the size of effect that could be expected (see Dienes, 2014). Using the tonal inversion

[8] For relevant code see https://debruine.github.io/bfrr/

paradigm, when testing on the same length as training, Jiang, Zhu, Guo, Ma, et al. (2012) found an effect of 7% above a control; Li, Jiang, Guo, Yang et al. (2013) found 6%, and Qiao, Sun, Li, Ling et al. (2018) found 6%. Thus, the rough size of effect expected if there is learning of different lengths by a mechanism that learnt the inversions per se is 6% above baseline. We modelled H1 as a half-normal with an SD of 6%. With these assumptions for modelling H1, as it happened, where an effect yielded a *p* value above .05, the Bayes factor was above 3, and vice versa (cf. Jeffreys, 1939, p. 359, for this rough but not guaranteed correspondence between B and *p*; if the obtained effect is roughly the size expected on a half-normal model of H1 the correspondence typically obtains, Dienes, 2014, but there is in general no guarantee of a correspondence between *B* and *p* values, which are not monotonically related, Lindley, 1957). We will interpret all effects with respect to the Bayes factors, while also reporting *p* values.

To indicate the robustness of Bayesian conclusions, for each *B*, a Robustness Region is reported, giving the range of scales that qualitatively support the same conclusion (i.e. evidence as supporting H0, or as supporting H1, or there not being much evidence at all), notated as: $RR_{B>3}$ [x1, x2] where x1 is the smallest SD that gives the same conclusion and x2 is the largest; or $RR_{B<1/3}$ [x1, x2]; or $RR_{1/3<B<3}$ [x1, x2]."

Here is an example of reporting the results themselves: "The unequal length group performed at above chance levels, 55% (SE = 1.8%), t(30) = 2.78, p = .0094, $B_{H(0, 6\%)}$ = 19.72, $RR_{B>3}$[0.8, >50%], as did the equal length group, 57% (SE = 2.5%), *t*(30) = 2.80, *p* = .0088, $B_{H(0, 6\%)}$ = 21.61, $RR_{B>3}$[1.1, >50%]. There was no evidence one way or the other whether the two groups differed, Mdiff = 2%, SE = 3.1, *t*(60) = 0.65, *p* = .52, $B_{H(0, 6\%)}$ = 0.78, $RR_{B>3}$[0, 18%]. There was evidence that the control group performed at chance, 49% (SE = 1.5%), *t*(30) = 0.67, *p* = .51, $B_{H(0, 6\%)}$ = 0.16, $RR_{B>3}$[2.6, >50%].)" It reads much as a standard results section, and a significance tester could agree with the conclusions - yet distinctions are made unavailable to the significance tester: There can be evidence for no effect, or not much evidence either way for an effect. The distinction has consequences for the discussion. In the discussion the prediction that the different

length and same length groups would perform the same would not be counted as either confirmed or disconfirmed. Conversely, the claim that the control group performed at chance could be made, and the theoretical implications of this discussed.

## Discussion

Bayesian hypothesis testing is especially useful when answering the existential question of whether or not something exists, or determining the plausibility or degree of evidence for something existing. There are two methods for answering this question. Either the researcher determines what a minimally interesting effect is (and then considers how much of the posterior distribution falls in the null interval); or else one considers the full range of possible effects predicted by the theory (and calculates a Bayes factor). The choice of method depends on what aspect of the prediction of the theory is most salient and easy to motivate objectively. If what is most crucial is the minimal interesting effect size (m), and especially if the range of effect sizes predicted is hard to otherwise specify, then inference can be with respect to the null interval the minimal effect size defines ([-m, +m]). If the minimally interesting effect size is hard to determine and it is anyway very small compared to a readily predicted effect size, then inference is most convenient using Bayes factors testing against a point H0. What one cannot do is determine if a point "no effect" is outside or inside the credibility interval in order to make an existential claim.

Bayes factors and inference with respect to intervals are not guaranteed to come to apparently the same conclusion (they ask different questions). The Bayes factor uses the full range of predictions of the theory with respect to the effect tested; inference by intervals focuses on just one aspect of that prediction, the minimally interesting effect size. (Bayes factors can incorporate the latter, but they will give a very similar answer when testing against a null interval H0 as a point H0 so long as the null interval is small compared to the standard error.) One can for example, have evidence for no effect (with a Bayes factor) even though a range of possibly interesting effects remain plausible (the CI extends above the null interval) (See Palfi & Dienes, 2019b for an

example). This arises when the theory predicts that large effects – that is, an extensive range of effect sizes above the null interval - are plausible. A couple of examples will illustrate how predicting a range of possibilities can entail evidence against the theory even though not all theoretically relevant possibilities have been ruled out. Imagine you left your keys in either the bedroom or the kitchen. There are 10 places in the bedroom they could be. You have searched eight and they were not there. Whatever your prior confidence that the keys were in the bedroom rather than the kitchen, those odds should now be reduced – even though the keys still could be in the bedroom. Similarly, supersymmetry in physics predicts particles that could have a range of masses. That whole range has not been explored; but much of it has without finding supersymmetric particles. Because of this, some physicists have reported a dramatic reduction in confidence in the theory (Jha, 2013).

Making existential claims depends on having a theory of how big the thing is one is looking for. With no such theory, one cannot obtain evidence for something not existing. In that case, one could just estimate. Similarly one could just estimate if existence was certain and the only question of interest was how big it was. For example, in developing a new variant of an already reliable scale, finding the reliability of the new scale, or its correlation with the old one, may be a matter of how big the correlations are; establishing the mere existence of the correlations may be pointless. In that case, 95% Cis may be all that is needed (e.g. Palfi, Moga, Lush, Scott, et al., 2019). If one only estimates, be careful not to make existential claims (thereby acting as if such claims had been established by the reported statistics). Do not say the regression slope is different from zero, say it is probably between 0.5 and 0.8 ratings units/rating unit. If the posterior distribution is centred close to zero, do not say there is evidence for the slope being zero; say that the slope probably lies between the limits of the CI.

A continuous posterior distribution on a difference presumes that difference definitely exists. One may presume existence of the difference and wish to know if the difference is positive or negative. Thus, in estimating a difference one may conclude that the difference is probably positive

if, for example, 95% of the posterior lies above zero. In this case, no minimally interesting effect size nor scale of effect need be specified.

The examples given in the paper have been concerning whether a population parameter is zero, or one side of zero. However, it is one thing for a proportion of individuals to have a population effect one way, and a proportion of individuals a population effect the other way or a zero effect; and another thing for everyone to show an effect in the same direction (Haaf & Rouder, 2019). The same issue of determining relevant effect sizes arises in modelling this situation. What is the relevant effect size may then vary across individuals; for example, in determining the extent of metacognition according to meta-d', each individual's first order d' can be used to fix the predicted meta-d' for that individual (for an example see Leganes Fonteneau, Scott, & Duka, 2018).

Aczel, Hoekstra, Gelman, Wagenmakers, et al. (2020) considered the approaches of different Bayesian experts and concluded that the main point they had in common was that analyses should avoid ritual, and should engage with problems thoughtfully. An example of that principle in this article is advising thoughtful engagement with what size effect a theory predicts. Most of the experts surveyed in Aczel et al. did not like thresholds of evidence, on the grounds that Bayes factors are continuous indicators of evidence. Why do we need conventions for good enough evidence? One could avoid all arbitrary conventions, for evidence or anything else, and let people make their own minds up from data. The limit of this argument is for the results and discussion section to consist of the following sentence: "Here are the raw data: make of them what you will!" We should not be dichotophobic (the irrational phobia of making discriminations when discriminations are called for). At some point someone has to make something of the data. How shall they do it? When we decide to stop gathering evidence against one confound and move on to other issues, a decision is necessarily made about how much evidence is good enough. We just have to bear in mind that every statistical model, every statistical decision is provisional: At a later point, better reasons may emerge for a different minimally interesting effect size, or for a different model of H1, or for a confound in the design to seem more plausible, and the inferences motivated by the

data thereby change. Our statistical modelling and inferences are conjectural. But provisional conventions help us move on. (Compare Popper's 1963 argument for his measure of the degree of corroboration for a theory shown by some data being provisional.)

The approaches described in this article can be applied to the output of many different ways of modelling the data. Model the data as seems best. Once we have a parameter and its standard error, either inference by intervals or Bayes factors can be applied in order test hypotheses about that parameter. (And thereby test models and theories piece-meal, Mayo, 2018.) No new statistical packages need be learned. One need not be a "Bayesian" to benefit from the approaches described here. One need only care about whether one has good reasons for answering the existential question of whether something exists.

## References


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

Aczel, B., Hoekstra, R., Gelman, A., Wagenmakers, E. J,. ... van Ravenzwaaij, D (2020). Expert opinions on how to conduct and report Bayesian inference. *Nature Human Behaviour, https://doi.org/10.1038/s41562-019-0807*

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., ... Wagenmakers, E.-J. (2018). Quantifying support for the null hypothesis in psychology: An empirical investigation. *Advances in Methods and Practices in Psychological Science, 1,* 357-366.

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology, 74*, 187–195.

Anvari, F., & Lakens, D. (2019). Using Anchor-Based Methods to Determine the Smallest Effect Size of Interest. *PsyArXiv. 10.31234/osf.io/syp5a*

Appelbaum, M., Cooper, H., Kline, R. G., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal Article Reporting Standards for Quantitative Research in Psychology: The APA Publications and Communications Board Task Force Report. *American Psychologist, 73*, 3-25.

Armstrong, A. M., & Dienes, Z. (2014). Subliminal Understanding of Active vs. Passive Sentences. *Psychology of Consciousness: Theory, Research, and Practice, 1*, 32-50.

Barrett, A., Dienes, Z., & Seth, A. (2013). Measures of metacognition in signal detection theoretic models. *Psychological Methods, 18*, 535-552.

Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B. A...Valen, E. J. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*, 6–10.

Button, K. S., Kounali, D., Thomas, L. , Wiles, N. J., … Lewis, G. (2015). Minimal clinically important difference on the Beck Depression Inventory - II according to the patient's perspective. *Psychological Medicine, 45*, 3269-3279.

Colling, L.J., & Szűcs, D. (2018). Statistical Inference and the Replication Crisis. *Review of Philosophy and Psychology*, doi:10.1007/s13164-018-0421-4

Depaoli, S., & van de Schoot, R. (2017). Improving Transparency and Replication in Bayesian Statistics: The WAMBS-Checklist. *Psychological Methods, 22*, 240 –261.

Dienes, Z. (2008).*Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference*. Palgrave Macmillan.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psycholology, 5*: 781. doi: 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*. Oxford: Oxford University Press, pp 199-220.

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology, 72*, 78-89.

Dienes, Z. (2017). https://www.youtube.com/watch?v=9hFN0csyeO4&t=3355s Uploaded 19 April 2017; downloaded 29 Jan 2020.

Dienes, Z. (2019). How do I know what my theory predicts? *Advances in Methods and Practices in Psychological Science, 2*, 364-377.

Dienes, Z., & McLatchie, N. (2018). Four reasons to prefer Bayesian over significance testing. *Psychonomic Bulletin & Review, 25*, 207-218.

Dienes, Z., & Seth, A. (2010). The conscious and the unconscious. In G. F. Koob, M. Le Moal, & R. F. Thompson (Eds), *Encyclopedia of Behavioral Neuroscience, volume 1,* pp. 322–327. Oxford: Academic Press.

Dworkin, R. H., Turk, D. C., Wyrwich, K. W., Beaton, D., Cleeland, C. S., Farrar, J. T., ... & Brandenburg, N. (2008). Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *The Journal of Pain, 9*, 105-121.

Etz, A., Haaf, J. M., Rouder, J. N., & Vandekerckhove, J. (2018). Bayesian Inference and Testing Any Hypothesis You Can Specify. *Advances in Methods and Practices in Psychological Science, 1,* 281-295.

Etz, A., Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review, 25*, 5–34.

Faivre, N., Berthet, V., & Kouider, S. (2012). Nonconscious influences from emotional faces: a comparison of visual crowding, masking, and continuous flash suppression. *Frontiers in Psychology*, https://doi.org/10.3389/fpsyg.2012.00129.

Fisher, R. A. (1935). *The design of experiments.* Macmillan.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian Data Analysis*, 3rd Edition. Boca Raton: Chapman & Hall.

Gelman, A., Simpson, D., & Betancourt, M. (2017). The Prior Can Often Only Be Understood in the Context of the Likelihood Entropy 2017, 19, 555; doi:10.3390/e19100555

Greenland, S. (2006). Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *International Journal of Epidemiology, 35*, 765–775.

Greenland, S. (2007). Bayesian perspectives for epidemiological research. II. Regression analysis. *International Journal of Epidemiology, 36*, 195–202.

Greenland, S. (2017). A serious misinterpretation of a consistent inverse association of statin use with glioma across 3 case-control studies. *European Journal of Epidemiology, 32*, 87-88.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1–20.

Haaf, J. M., Ly, A., & Wagenmakers, E.-J. (2019). Retire significance, but still test hypotheses. *Nature, 567*, 461.

Haaf, J.M., Rouder, J.N. (2019). Some do and some don't? Accounting for variability of individual difference structures. *Psychonomic Bulletin & Review, 26*, 772–789.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature Methods, 12,* 179–185.

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

Jha, A. (2013). https://www.theguardian.com/science/2013/aug/06/higgs-boson-physics-hits-buffers-discovery Downloaded 30 Jan 2020.

Jiang, S., Zhu, L., Guo, X., Ma, W., Yang, Z., and Dienes, Z. (2012). Unconscious structural knowledge of tonal symmetry: tang poetry redefines limits of implicit learning. *Consciousness & Cognition, 21*, 476–486.

Kelly, A. (2001). The minimum clinically significant difference in visual analogue scale pain score does not differ with severity of pain. *Emergency Medicine Journal, 18*, 205–207.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6*, 299-312.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General, 142*, 573–603.

Kruschke, J. K. (2014). *Doing Bayesian data analysis: a tutorial with R and BUGS*, 2nd Edition. London: Academic Press.

Kruschke, J. K. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science, 1*, 270–280.

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2018). Improving Inferences about Null Effects with Bayes Factors and Equivalence Tests. *The Journals of Gerontology, Series B: Psychological Sciences*, *75*, 45–57

Lakens, D., Scheel, A. M., & Isager, P. M (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science, 1*, 259–269.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian modeling for cognitive science: A practical course.* Cambridge: Cambridge University Press.

Leganes-Fonteneau, M., Nikolaou, K., Scott, R., and Duka, T. (2019). Knowledge about the predictive value of reward conditioned stimuli modulates their interference with cognitive processes. *Learning & Memory, 26*, 460-464.

Leganes Fonteneau, M., Scott, R. and Duka, D. (2018). Attentional responses to stimuli associated with a reward can occur in the absence of knowledge of their predictive values. *Behavioural Brain Research, 341*, 26-36.

Leucht, S., Fennema, H., Engel, R., Kaspers-Janssen, M., Lepping, P., & Szegedi, A. (2013). What does the HAMD mean? *Journal of Affective Disorders, 148*, 243-248.

Leucht, S., Kane, J. M., Etschel, E., Kissling, W., Hamann, J., & Engel, R.R. (2006). Linking the PANSS, BPRS, and CGI: clinical implications. *Neuropsychopharmacology, 31*, 2318-2325.

Li, F., Jiang, S., Guo, X., Yang, Z., & Dienes, Z. (2013). The nature of the memory buffer in implicit learning: Learning Chinese tonal symmetries. *Consciousness & Cognition, 22*, 920-930.

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

Lindley, D. V. (1991). Comment on Aitkin "Posterior Bayes Factors." *Journal of the Royal Statistical Society. Series B (Methodological), 53*, 130-131.

Lindley, D. V. (2014). *Understanding uncertainty, revised edition*. New Jersey: Wiley.

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2019). Replication Bayes factors from evidence updating. *Behavior Research Methods, 51*, 2498-2508.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition, 21*, 422–430.

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond he statistics wars.* Cambridge: Cambridge University press.

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Boca Raton: Chapman & Hall.

Moncrieff, J., & Kirsch, I. (2015). Empirically derived criteria cast doubt on the clinical significance of antidepressant-placebo differences. *Contemporary Clinical Trials*, *43*, 60-62.

Morey, R., Homer, S. and Proulx, T. (2018). Beyond statistics: accepting the null hypothesis in mature sciences. *Advances in Methods and Practices in Psychological Science, 1*, 245-258.

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. J*ournal of Mathematical Psychology, 72*, 6–18.

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

O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparison. *Journal of the Royal Statistical Society: Series B (Methodological), 57*, 99-118.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349* (6251), 943–951.

Palfi, B., & Dienes, Z. (2019a). When and how to calculate the Bayes factor with an interval null hypothesis. *PsyArXiv*, https://doi.org/10.31234/osf.io/9chmw

Palfi, B., & Dienes, Z. (2019b). The role of Bayes factors in testing interactions. https://doi.org/10.31234/osf.io/qjrg4

Palfi, B., Moga, G., Lush, P., Scott, R. B., & Dienes, Z. (2019). Can Hypnotic Suggestibility Be Measured Online?. *Psychological Research*, https://doi.org/10.1007/s00426-019-01162-w

Perugini, M., Gallucci, M., Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science, 9*, 319–332.

Popper, K. R. (1963). *Conjectures and Refutations: The growth of scientific knowledge*. London: Routledge.

Qiao, F., Sun, F., Li, F., Ling, X., Zheng, L., Li, L., Guo, X., & Dienes, Z. (2018). Tonal symmetry induces fluency and sense of well-formedness. *Frontiers in Psychology, 9*, doi.org/10.3389/fpsyg.2018.00165

Ramsøy, T. Z. and Overgaard, M. (2004) Introspection and subliminal perception. *Phenomenology and the Cognitive Sciences, 3*, 1–23.

Rouder, J., & Haaf, J. M. (2020). Optional Stopping and the Interpretation of The Bayes Factor. https://doi.org/10.31234/osf.io/m6dhw

Rouder, J. N., Morey, R. D., Speckman, P. L., & Pratte, M. S. (2007). Detecting chance: a solution to the null sensitivity problem in subliminal priming. *Psychonomic Bulletin and Review, 14,* 597–605.

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.

Sand, A., & Nilsson, M. E. (2016). Subliminal or not? Comparing null-hypothesis and Bayesian methods for testing subliminal priming. *Consciousness & Cognition, 44*, 29–40.

Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review, 25,* 128–142.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods, 22*, 322–339.

van Doorn, J., van den Bergh, D., Bohm, U., Dablander, F., Derks, K., Draws, T., ... Wagenmakers, E. (2019). The JASP Guidelines for Conducting and Reporting a Bayesian Analysis. *PsyArXiv*, https://doi.org/10.31234/osf.io/yqxfr

van Ravenzwaaij, D., Cassey, P. & Brown, S.D. (2018). A simple introduction to Markov Chain Monte–Carlo sampling. Psychonomic Bulletin & Review, 25, 143–154. https://doi.org/10.3758/s13423-016-1015-8

Vanpaemel, W. (2016). Prototypes, exemplars and the response scaling parameter: A Bayes factor perspective. *Journal of Mathematical Psychology 72*, 183-190.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J.,... Morey, R. D. (2018). Bayesian inference for psychology. Part I: *Theoretical advantages and practical ramifications. Psychonomic Bulletin & Review, 25*, 35-57.

Wagenmakers, E.-J., Verhagen, A. J., Ly, A., Matzke, D., ... Morey, R. D. (2017). The need for Bayesian hypothesis testing in psychological science. In S. Lilienfeld & I. Waldman, (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 123-138). New York: John Wiley and Sons.

Watson, D., ,Clark, L. A., Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*, 1063-1070.

Ziliak, S. T., & McCloskey, D. N. (2008). *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Michigan: University of Michigan Press

Ziori, E., & Dienes, Z. (2015). Facial beauty affects implicit and explicit learning of men and women differently. *Frontiers in Psycholology, 6,* 1124. doi: 10.3389/fpsyg.2015.01124