What is a Bayes factor?

Schmalz, Xenia [1,2]*; Biurrun Manresa, José [3]; Zhang, Lei [4]

[1] Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, University Hospital, LMU Munich, Germany

[2] Department of Psychology and Cognitive Sciences, University of Trento, Italy

[3] Institute for Research and Development in Bioengineering and Bioinformatics (IBB-CONICET-UNER), National Scientific and Technical Research Council, Argentina

[4] Social, Cognitive and Affective Neuroscience Unit, Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna, Austria

* *Corresponding author. Email: xenia.schmalz@gmail.com, phone: +49 89 4400 56955, fax: +49 89 4400 55902, address: Pettenkoferstr. 8a, 80336 München, Germany.*

Note: This is a preprint of an article which is due to appear in *Psychological Methods.*

Abstract

The use of Bayes factors is becoming increasingly common in psychological sciences. Thus, it is important that researchers understand the logic behind the Bayes factor in order to correctly interpret it, and the strengths of weaknesses of the Bayesian approach. As education for psychological scientists focuses on frequentist statistics, resources are needed for researchers and students who want to learn more about this alternative approach. The aim of the current article is to provide such an overview to a psychological researcher. We cover the general logic behind Bayesian statistics, explain how the Bayes factor is calculated, how to set the priors in popular software packages, to reflect the prior beliefs of the researcher, and finally provide a set of recommendations and caveats for interpreting Bayes factors.

What is a Bayes factor?

A note on reading this article:
Reading an article about statistics is very different from reading an article about psychological sciences. When reading a psychology article, most experienced psychological scientists can read it once and understand its contents. Simply reading the paper is often not sufficient for understanding statistics articles, which can be frustrating. We suggest to first read this paper to get the general ideas behind it (if necessary, skipping the more maths-heavy sections - we flag the sections which can be skipped without compromising on the understanding of the general concepts). After a consolidation period, re-reading the paper more thoroughly may be helpful, and the reader may want to spend time on understanding the details: scrutinise the formulae, try to understand where each term comes from, and compute some additional examples (e.g., for Figure 1, calculating the posterior probabilities for different prior probabilities, power parameters, and alpha-levels).

The most common statistical tool for making inferences in psychological sciences is the *p*-value. The *p*-value has been criticised (e.g., Cumming, 2014; Wagenmakers, 2007), and its (mis)use has received some of the blame for the replication crisis in psychology (Dienes, 2016; Halsey et al., 2015; Open Science Collaboration, 2015; for a thorough discussion about *p*-values, see special issue on this topic in the American Statistician, e.g., Wasserstein & Lazar, 2016). As an alternative to *p*-values, some researchers have proposed to switch to Bayesian statistics for inference (Dienes, 2011; Rouder et al., 2009; Wagenmakers, 2007). Bayesian statistics follow a different philosophy compared to *p*-values, which are derived from a frequentist framework, and may be a valuable addition to the tools in a psychological scientist's repertoire. An often-cited advantage of Bayes factor analyses over significance testing is that the Bayes factor allows us to provide evidence for the null hypothesis, rather than only being able to reject it (but see also Harms & Lakens, 2018; Lakens, 2017). However, using Bayesian statistics without a thorough understanding is likely to result in the inheritance of one of the main issues with *p*-values: the application of a statistical ritual, which may lead to suboptimal decisions about how to analyse the data, and which conclusion one can draw from them (Gigerenzer, 2004). In the case of Bayesian statistics, the danger of the analysis turning into a ritual is exacerbated, compared to frequentist statistics, because

most researchers and current psychology students have covered the latter extensively throughout their studies.

Using Bayes factors requires the explicit specification of parameters that are not necessary for calculating a *p*-value (e.g., prior distributions); thus, arguably, greater familiarity with the logic underlying these analyses is required to make informed decisions about the parameter choices. The current paper therefore aims to bridge a gap in the literature on statistical tools, by explaining Bayesian statistics, and in particular the Bayes factor, to a researcher or student who has received only the basic statistics education that is standard as part of a psychology degree (see also Colling & Szűcs, 2018; Kruschke & Liddell, 2018; Tendeiro & Kiers, 2019, for texts with a similar aim but on a more advanced level, as well as Vandekerckhove, Rouder, & Kruschke, 2018, and the corresponding special issue). To be clear, the current paper does not aim to convert researchers to using Bayes factors for inference, but rather to provide them with a minimum amount of knowledge that is necessary to interpret a Bayes factor, should they decide to use it or come across it in another manuscript.

What do Bayesian statistics mean? An introduction to Bayes' Theorem

Bayesian statistics cover a range of procedures, though the most popular one in psychological science is the Bayes factor (Dienes, 2014; Mulder & Wagenmakers, 2016; Rouder et al., 2007). Their best-known property is likely to be the fact that they are susceptible to distribution *priors*: the prior belief that a researcher has about the models' predictions before collecting data or conducting the analysis, which can be explicitly included in the statistical model (see glossary and section "What is a Bayes factor" for a more detailed definition and explanation of different types of priors). This feature derives from the origin of the term "Bayesian statistics", namely Bayes' Theorem (for more detailed description of the theorem, see Rouder & Morey, 2019). A description of the theorem and its implication for making inferences follows; however, a reader who aims to get only a first-pass conceptual understanding may skip to the "Interim summary and discussion" subsection without compromising their understanding of the subsequent sections.

Bayes' Theorem describes the conditional probability of an event, P(A|B) (reads as "probability of A given B"). *Conditional probability* is an important concept not just for Bayesian statistics but also for properly understanding frequentist statistics: it is the probability of an event (*A*), under the condition of a different event (*B*), such as the

probability of a person having a disease (*A*) under the condition that they have scored positively on a test designed to diagnose it (*B*).

Bayes' Theorem has the following mathematical form:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$$

To illustrate the meaning of this equation with an example: We would like to get the conditional probability that a patient has a disease (*A*), given they tested positive on some diagnostic test (*B*). In order to calculate this, we need to know both the probability that this patient has the disease (*P(A)*) and the conditional probability that the test would detect this disease, under the assumption that the patient has it (*P(B|A)*, or the true positive rate). *P(B)*, in the denominator, refers to the overall probability that a person chosen at random will get a positive test, regardless of whether they are affected or not. *P(A)* is the prior probability; in this case one can substitute the prevalence of the disorder. If a disorder is extremely rare (e.g., the Black Plague), it is intuitive that, even if someone tested positive on a test that is supposed to detect it, the probability of the patient actually having it continues to be low. The term *P(A)* is multiplied by the ratio of positive test rates, which reflects the intuitive point that, all else kept equal, a low prior probability (or, as $P(A) \rightarrow 0$) will lead to a low posterior probability ($P(A|B) \rightarrow 0$).

Bayes' Theorem can be used to demonstrate the concept of applying prior knowledge in a formal calculation, though it does not directly translate to the way in which Bayesian statistics such as the Bayes factor are calculated. After introducing the logic behind Bayes' Theorem, we can apply it to a statistical null hypothesis test. Let's say we obtain an observation *B,* such as running an experiment and obtaining a *p*-value smaller than 0.05. For example, a psychological researcher may be keen to gain insights into whether a treatment is effective. We can write this as *A*: the probability that the null hypothesis of no treatment effect (H_0) is wrong ($A = P(H_1)$).[1] Applying these variables to Bayes' Theorem, we get the following:

[1] Note that we assume only two plausible hypotheses: H_0 and H_1. We can make this assumptions under two conditions: either when H_1 is a non-directional hypothesis, H_1: $\delta \neq 0$, or when we have a strong theoretical

$$P(H_1|p < 0.05) = \frac{P(H_1) \cdot P(p < 0.05|H_1)}{P(p < 0.05)}$$

$P(p<0.05|H_1)$ is the statistical power, because statistical power is defined as the conditional probability of a significant *p*-value, given that the alternative hypothesis is true (formally, a power calculation requires us to specify an effect size, which is often denoted as H_1). If power is 80%, then $P(p<0.05|H_1)$, also known as the true positive rate, is 0.80. So far, these are concepts that most readers will be familiar with from their use of frequentist statistics.

The novel concept in this equation is the probability of the hypothesis ($P(H_0)$) or rather its inverse, $P(H_1)$. We also need to calculate $P(p<0.05)$, the overall probability of a significant *p*-value. How this is calculated can be derived intuitively from Figure 1. For now, we focus on the $P(H_1)$-term in the numerator. We can plug in the values for the standardly assumed case of 80% power and alpha-level of 0.05, and assuming that we observed a significant *p*-value in an experiment:

$$P(H_1|p < 0.05) = \frac{P(H_1) \cdot 0.80}{0.80 \cdot P(H_1) + 0.05 \cdot P(H_0)}$$

For the probability of the null hypothesis being false ($P(H_1)$), the researcher needs to apply some informed judgement. Putting this into a concrete example, let's say, our null hypothesis is that precognition is not real: humans cannot foresee the future. To test this hypothesis, we recruit undergraduate students and ask them to predict, in an alternative-forced-choice experiment, which words they will see in a word list which will be presented to them afterwards (Bem, 2011). Most researchers will have a very strong prior that precognition is not real: we could put a number on this by saying that, in this case, our prior belief in the null hypothesis is 99%. In this case, the probability of the null hypothesis (H_0) is 0.99, and the probability of H_1 is 0.01. If we run the experiment and find that undergraduate students are significantly above chance level (with $p < 0.05$, i.e., the standard significance threshold), we can substitute the following:

rationale for assuming a more specific H_1 such that H_0 and H_1 are the only two plausible hypotheses (e.g., a directional hypothesis, where H1: $\delta > 0$.

$$P(H_1|p < 0.05) = \frac{0.01 \cdot 0.80}{0.80 \cdot 0.01 + 0.05 \cdot 0.99}$$

This calculation gives us a value of $P(H_1|p < 0.05) \approx 0.14$: Even after observing a significant *p*-value, given the parameters described above and a strong prior belief that precognition is not real, the probability of the null hypothesis remains approximately 86%. The obtained posterior probability may seem disappointing to a researcher searching for a yes-or-no answer. However, an alternative way to think about getting from a prior to a posterior probability is that we have updated our prior belief in light of incoming evidence: while the hypothesis had a very low probability of being true in the first place, we consider it approximately 14 times more likely after having observed the significant p-value (see https://www.youtube.com/watch?v=lG4VkPoG3ko for a video explaining this point). Figure 1 depicts Bayes' Theorem, and the steps for calculating the conditional probabilities, *P(H₁|p<0.05)*, using frequency counts rather than percentages, under two different assumptions: 50% confidence about the null hypothesis being true, or 99% confidence about the null hypothesis being true.

It is noteworthy that Figure 1 and the in-text description of the application of Bayes' Theorem to calculate *P(H₁|p<0.05)* rely on different definitions of probability (see Chapter 8 of Spiegelhalter, 2019, for a description and summary of different definitions of probability). In the text, we described the percentage, which we plugged in as the prior probability, as the "prior belief" associated with one particular hypothesis. This is in line with a Bayesian approach. Here, the outcome of a hypothesis test serves to shift the prior belief of a given hypothesis to a posterior belief, which incorporates incoming evidence. In contrast, in Figure 1, we take a large number of different possible hypotheses, and the prior is the percentage of possible hypotheses where H_0 is true. We do not assign a probability to a specific hypothesis, but rather calculate what would happen with the posterior probability *in the long run* (i.e., if we repeat the procedure frequently - giving rise to the term "frequentist statistics"), under different assumptions. Most types of data can be analysed with either Bayesian or frequentist methods: there is nothing inherent to a dataset that makes it "frequentist" or "Bayesian" or specifically suitable for these approaches.

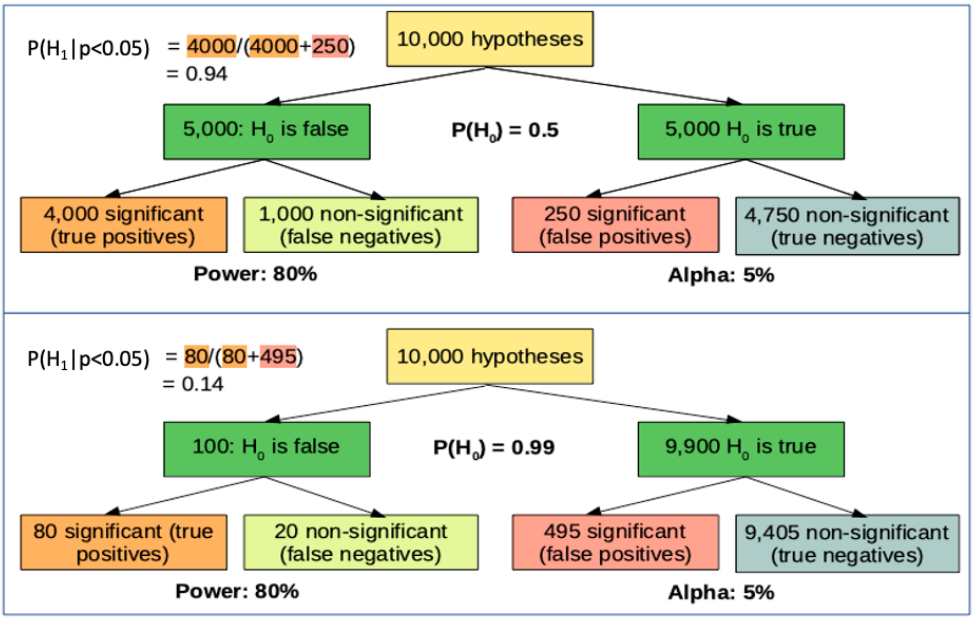


Figure 1. Calculation of the posterior probability of the null hypothesis not being true after observing a significant effect ($p < 0.05$), under the assumption of 80% power, alpha = 0.05, and (1) 50% of all experiments being conducted are on cases where the null hypothesis is true (upper panel), and (2) when 99% of all experiments being conducted are on cases where the null hypothesis is true (lower panel).

The reason why Figure 1 shows Bayes' Theorem in terms of frequencies rather than probabilities is to facilitate the understanding of the underlying concepts, as studies have shown that most people find natural frequencies more intuitive than probabilities (reviewed in Gigerenzer et al., 2007). The presentation of natural frequencies is in line with a frequentist approach: despite the use of Bayes' Theorem, the processes are described by making assumptions about what happens in the long run (e.g., about the percentage of hypotheses where H_0 is true). The probability of a single hypothesis being true or false cannot be assigned in the frequentist framework, because the long-run behaviour does not allow for inferences about a single event.

Interim summary and discussion

The section above aims to introduce the concepts of prior and posterior beliefs, and to explain the workings behind Bayes' Theorem. These are important concepts for Bayesian statistics, but it is important to stress that the calculations are not directly relevant to computing and interpreting a Bayes factor, which we will introduce in the next section. Bayes' Theorem illustrates the intuitive idea that, if a hypothesis is highly unlikely, it remains unlikely (though

slightly less so) even after we collect data that is consistent with it. Let's say we conduct two experiments: one on precognition (testing whether participants can foresee the future; Bem, 2011) and one on the Stroop effect (testing whether it takes longer to name the colour of the font when a colour word is written in a colour which is incongruent with it compared to when it is congruent; MacLeod, 1991). For both experiments, we happen to observe a *p*-value of $p = 0.026$. A naïve user of statistics may conclude that the identical *p*-values suggest that both effects are true. A Bayesian approach gives us a more intuitive justification for remaining more sceptical about precognition than about the Stroop effect (although theoretical and methodological justifications may be more pertinent).

An important point demonstrated in the calculations above is that we cannot calculate the posterior probability of a hypothesis without the prior, which can be conceptualised as the researcher's belief that a given hypothesis (e.g., the null hypothesis) is correct. The *p*-value gives us the conditional probability of *the observed data* (or more extreme observations), given the null hypothesis (*P(Data|Null Hypothesis)*): whenever we want to get from P(B|A) to P(A|B) (i.e., to calculate *P(Null Hypothesis|Data)*), we need to use Bayes' Theorem. Calculating the posterior probability of the hypothesis, given some data, is not possible without making some assumptions about its a priori probability.

Compared to using the *p*-value for frequentist inference, Bayesian statistics follow two principles which we have discussed so far: (1) They take into account the prior probability to calculate a posterior probability, and (2) they can describe the probability of a single hypothesis, as opposed to the long-term behaviour under a set of assumptions. Both of these principles are important to bear in mind when we aim to understand how Bayesian statistics such as the Bayes factor can be used for inference.

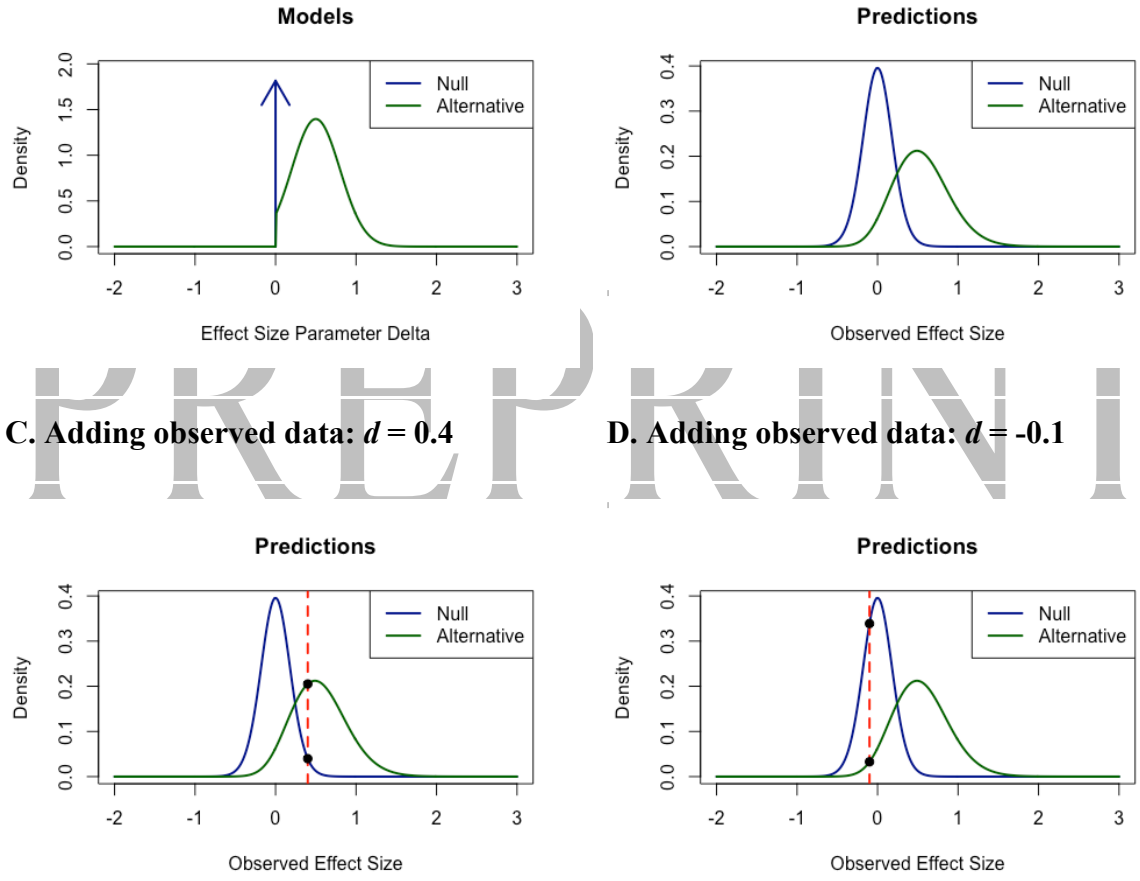
What is the Bayes factor?

In the following section, we provide a step-by-step instruction about how a researcher may construct a set of priors and use these to calculate a Bayes Fator (for a more advanced tutorial, see Wagenmakers et al., 2010). So far, we have focused on one hypothesis only: The null hypothesis. If we imagine any research question that we have been thinking about lately, and we describe the statistical hypothesis in words, it might go something like this: "According to my theory, there should be a difference between two conditions, where one group should have higher scores than the other. The difference probably won't be too big, but

neither too small. If my theory is incorrect, there will be no difference."[2] A sensible way to draw this theory-driven hypothesis is depicted in Figure 2, panel A (the plots in Figure 2 are generated from code provided by Rouder, 2016). The *y*-axis shows the density, which is proportional to the probability of a given effect size for the alternative hypothesis.

A. A theoretical model of sensible priors.

B. Prediction densities for the same model.



C. Adding observed data: d = 0.4

D. Adding observed data: d = -0.1

Figure 2. How to generate a set of sensible priors for the null and alternative (Panels A and B), and to incorporate incoming data (C and D).

Our theoretical null model is that there will be no effect: the difference between the conditions will be exactly zero. This gives us the point prediction, drawn as the blue arrow.

[2] This seems to be a typical way in which psychological scientists formulate a hypothesis, and it is a reasonable approach if the researcher has little information about the possible effect size, and if there is no practical or theoretical reason for specifying a smallest effect of interest.

An effect which is not too big, not too small, could be described as Cohen's $d = 0.5$ (Lakens, 2013). However, we do not want to confine the model by saying that any values that are slightly bigger or slightly smaller are incompatible with our alternative hypothesis: we consider a medium-sized effect to be most likely, but also values around it would also be compatible with our theory. Getting from the model descriptions in Panel A to a distribution describing the predictions for an experimental outcome requires a mathematical procedure (described by Rouder, 2016). In Panel B, the point prediction of the null hypothesis is represented by a distribution, as in reality, we predict that the observed effect size will vary around the zero point due to sampling variability. For the alternative hypothesis, we apply a transformation which makes the prior distribution broader: Again, due to sampling variability, our prediction about the observed effect size is less precise than our prior distribution about the population parameter, which is reflected by the broader distribution. With this, we have designed a set of sensible priors for the null and alternative models, and we are ready to collect data.

Once we have collected the data, we can calculate the size of our effect of interest. In Panel C, our observed effect size is $d = 0.4$ (depicted by the red dashed line). Any possible observed value is to some extent likely both under the null and under the alternative model, as the probability at this point is not zero under either model. However, if we follow the intersection of the red line and the two hypotheses (marked by black dots) and look at the corresponding density values on the *y* axis, we get the following: The intersection of the red line with the null hypothesis (blue line) gives us a predictive density of approximately 0.04. The intersection with the alternative hypothesis (green line), gives us a predictive density of approximately 0.21. The Bayes factor is the ratio between these two densities[3]. In this case, 0.21/0.04 gives us a Bayes factor of 5.25. Thus, the Bayes factor is 5.25 in favour of the alternative hypothesis: the data are approximately 5 times more likely under the alternative than the null.

Panel D shows what happens when our observed effect size is -0.1. Again, due to sampling error, we might observe a negative effect size even under the alternative hypothesis of a

[3] Density is not the same as probability. However, here, it is proportional to probability: hence, if we divide the two density values we get a value which is equal to the ratio of the two probabilities.

positive effect. However, the density under the null happens to be 0.34, and under the alternative it is 0.03, which gives us a Bayes factor of 11.3 in favour of the null hypothesis.

What we have shown so far is the formulation of the Bayes factor, in terms of the described models (M_1 and M_2) and the observed data (D) as:

$$BF = \frac{P(D|M_1)}{P(D|M_2)}$$

The attentive reader will have noticed that we are comparing the conditional probability of the data under the model. Getting from the conditional probabilities of the data to the conditional probabilities of the model can be achieved by applying Bayes' Theorem. Using Bayes' Theorem, we can express the *posterior odds* of the two models. This requires some algebraic manipulations, which a reader may skip for a first-pass reading. First, we express the Bayes factor with the *P(D|M)* term replaced by Bayes' Formula, which gives us:

$$BF = \frac{\frac{P(D|M_1) \cdot P(D)}{P(M_1)}}{\frac{P(D|M_2) \cdot P(D)}{P(M_2)}} = \frac{P(M_1|D) \cdot P(M_2)}{P(M_2|D) \cdot P(M_1)}$$

Instead of expressing the Bayes factor as the product of the posterior probability (*P(M|D)*) and a prior (*P(M)*), we can rearrange the information above to derive the *posterior odds* of the two models, or the ratio of the conditional probabilities of the models given the data. This can be obtained by multiplying the Bayes factor by the models' prior probabilities:

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(D|M_1) \cdot P(M_1)}{P(D|M_2) \cdot P(M_2)}$$

In words, we can paraphrase this as:

Posterior odds = Bayes factor x prior odds

Thus, to derive the posterior odds, we need to multiply the Bayes factor with our prior belief about how plausible each model is: for example, we might weigh the alternative hypothesis as more probable if we are conducting an experiment on the Stroop task than on precognition. The Bayes factor does not take into account this prior probability of each model, and is therefore distinct from the posterior odds (though the posterior odds are equivalent to the Bayes factor in the special case that the two models are equally likely). Instead, to calculate

the Bayes factor, the prior knowledge or beliefs are instead incorporated in the way in which the alternative hypothesis is constructed.

Interim summary and discussion

Getting from data to posterior odds requires two stages where a researcher needs to make a decision: First, they need to decide on the models that specify the two hypotheses, which is required for the calculation of a Bayes factor. Second, the Bayes factor can be multiplied by the researcher by the strength of the prior belief in a model, which gives the relative posterior probability of the models given the data. Unless a researcher depends on default priors (objective Bayesian analysis; Consonni et al., 2018), both stages require informed and occasionally subjective decisions from the part of the user, and are here collectively referred to as "priors". The first step, a consideration of the plausibility of the model specification, is of vital importance to the interpretation of a Bayes factor. To date, psychological researchers seem to rely solely on the Bayes factor for inference, without taking the second step of calculating the posterior odds. This second step provides a more direct answer to the question: How much more likely is Model 1 compared to Model 2? These posterior odds are not generally reported in articles; for any reader, it is simple to calculate their own posterior odds by multiplying the Bayes factor by their own prior beliefs (Tendeiro & Kiers, 2019). For formal inference, posterior odds should be reported with caution, and preferably only if the *a priori* probability of each model has been pre-specified before the Bayes factor has been calculated; otherwise, finding a value for the strength of *a priori* belief that would allow the researcher to "provide support" for a favourite hypothesis becomes a trivial algebraic exercise.

To calculate the Bayes factor, we first need to construct two hypotheses: Commonly, a null model and an alternative model. Theoretically, we could compare any two hypotheses against each other (Etz, Haaf et al., 2018), such as the hypothesis that an effect is negative versus the hypothesis that an effect is positive (if this makes sense in light of a research question). The models can often be derived directly from a verbal theory, though mathematical transformations are required for which psychologists may want to rely on statistical software packages.

Once we construct the models, we assess whether the data are more likely under one model than the other, by calculating the ratio between the two probabilities. The Bayes factor can be

interpreted as the degree of evidence for one model over another. The degree of evidence is always for the model in the numerator: if the probability relating to the alternative hypothesis is in the numerator and the Bayes factor is large, it provides evidence in favour of the alternative hypothesis. Which of the models should be put in the numerator is a matter of convention. Some researchers prefer to always have the more complex model in the numerator, such that a large Bayes factor corresponds to evidence for an alternative, and a small Bayes factor to evidence for the null model. Other researchers prefer to report the larger value (i.e., the value that is greater than 1), which translates more straightforwardly to a statement about how much more likely the data is under one model than the other (e.g., "the data is 3x more likely under Model 1 than Model 2" as opposed to "the data is 0.33 times as likely under Model 2 than Model 1"). The Bayes factor in favour of the alternative hypothesis is conventionally denoted B_{10}, and the Bayes factor in favour of the null hypothesis is denoted as B_{01} (Love et al., 2019). Thus, whenever a Bayes factor is reported, it should always be stated if large numbers provide evidence for the null or the alternative model.

A Bayes factor value of 1 means that the data is equally likely under either model. Bayes factors are continuous, as they can take any value between 0 and infinity; the interpretation of different magnitudes is a matter of convention. Bayes factors larger than 3 are generally taken as some evidence for the model in the numerator, values larger than 10 as strong evidence, and values larger than 30 as very strong evidence (e.g., Jeffreys, 1961; Rouder et al., 2009). The inverse of these cut-offs (values < ⅓, 1/10, and 1/30, respectively) provide evidence for the model in the denominator. The correct interpretation of the Bayes factor is as a continuous scale, and the extent to which the data is more likely under one model than the other (which corresponds to the extent to which one model is more likely than the other if both models are equally likely a priori). An intuitive way to interpret a given Bayes factor value is: *How much would I be willing to bet that this result would replicate?* Odds of 30:1 (BF = 30, very strong evidence), would elicit different decisions than odds of 3:1 (BF = 3, moderate evidence).

How do I decide on the priors and interpret the Bayes factor?

When submitting a paper with Bayes factors for publication, it is likely that the reviewers will ask the authors to justify their priors. The above section provides the reader with the tools to create their own priors for a simple design, especially in conjunction with the R script provided by Rouder (2016). However, most researchers calculate Bayes factors using

software, such as JASP (Love et al., 2019). This calls for a need to understand default priors, and what it means to change them. Here, we focus on *t*-tests, as the aim is to provide an understanding about how the priors work, rather than a manual for using JASP. For more complex designs, choosing a well-informed prior becomes increasingly more difficult, which is a reason for repeating the recommendation that an experimental design should be kept as simple as possible (Cohen, 1990). What follows is a technical explanation: a reader aiming to get a first-pass conceptual understanding may skip to the "Interim summary and discussion" section.

Software such as JASP and the R package *BayesFactor* (Morey et al., 2018) rely on the Cauchy distribution as a prior for many of the analyses. A Cauchy distribution looks similar to the well-known normal distribution: it is symmetrical and bell-shaped. The two parameters that define the shape of the Cauchy distribution are the location parameter and the scale (width) parameter: with these two values, the exact shape of this prior distribution can be deduced. The location parameter defines where the centre of the distribution lies. The width parameter (ω) defines how thick or slim this distribution is. If we increase the width parameter, we consider a wider range of effect sizes to be more highly plausible. If we decrease the width parameter, we indicate increased confidence that the effect size is close to what we predict. Figure 3 shows the Cauchy distribution, always with the same location parameter (0), but with varying width parameters: as the width parameter increases, values close to the centre become less likely, while the tails become thicker.

By default, the Cauchy distribution in JASP and in the *BayesFactor* package is centred around zero. At first, this may seem counter-intuitive, because this is the prior parameter for the alternative hypothesis. The reason for the centring is that it allows us to test a *non-directional* hypothesis. Given that the prior distribution is symmetrical and centred around zero, effect sizes of, say, greater than 0.5 or smaller than -0.5 are equally likely. The bell-shape of the Cauchy distribution further indicates a belief that smaller effect sizes are more likely than larger effect sizes. When we increase the width parameter (ω), we increase the a priori probability of large effect sizes (though small effect sizes will continue to be more likely).

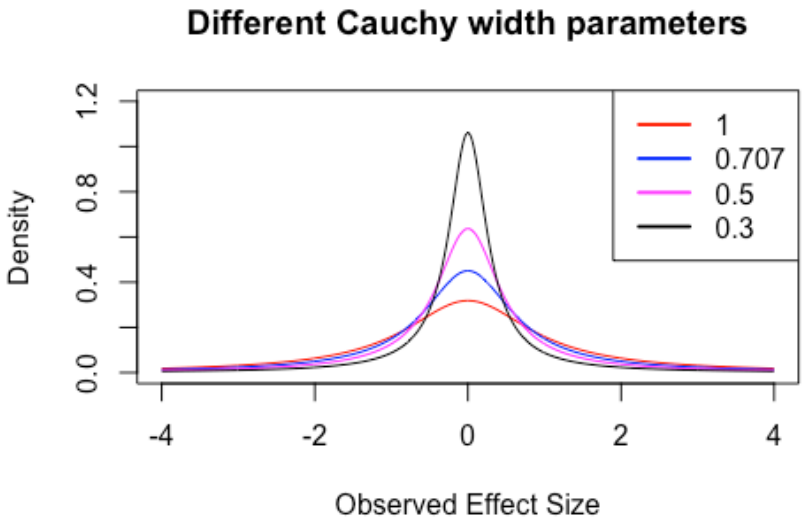


Figure 3. The Cauchy distribution, centred around zero, with different width parameters.

Interim summary and discussion

Software programmes such as JASP have a default prior. The default prior for the alternative hypothesis of a *t*-test is a Cauchy distribution (see Figure 3), which is centred around zero and has a width parameter of ω = 0.707. It is a common misconception that the width parameter is the expected effect size. Changing the width parameter does not shift the mode towards the most likely effect size values. Instead, it changes the relative probability with which larger effect sizes are proposed to occur: a wide distribution corresponds to a relatively higher probability of larger effect sizes (see Figure 3). Thus, while a large width parameter does correspond to the expectation of larger effect sizes, interpreting a width parameter as the expected effect size is incorrect.

The question is then how to translate the width parameter to the expected range of effect sizes. Here, it is important to bear in mind that the point of the Cauchy prior is not to predict a single effect size, but instead to specify a range of plausible effects. The width parameter of the Cauchy distribution (ω) corresponds to the bounds of the range of effect sizes which are proposed (by the researcher) to occur with a 50% probability. Thus, when we specify the width parameter as ω = 1, we do not put the heaviest weight on the effect size $d = 1$, but rather we make the statement: "*We are 50% confident that the effect size lies somewhere between d = -1 and d = 1*". The default prior, with the width parameter ω = 0.707, therefore,

makes the claim: "*We are 50% confident that the effect size lies somewhere between d = -0.707 and d = 0.707*". Thus, for the special case that we are 50% confident about an effect size range, the width parameter of the Cauchy distribution translates directly to the upper bound of the expected effect size range (see Table 1). If the researcher wants to choose a different level of confidence (e.g., we may be only 20% confident, or even 80% confident), the width parameter no longer corresponds to the bounds. In this case, some calculation is required. For a researcher who is 80% confident in their range of expected effect sizes, Table 1 provides a range of width parameters and effect size ranges that correspond to them.

The default width parameter, as explained above, is a non-directional prior. JASP allows us to test directional hypotheses in two different ways. The first is by specifying whether we expect Group 1 to have higher values than Group 2 ("Group 1 > Group 2"), or lower values ("Group 1 < Group 2"). The non-directional hypothesis is denoted as "Group 1 ≠ Group 2". For the directional hypotheses, the Cauchy is simply cut in half. The interpretation of a truncated prior is quite similar to that for the non-directional prior, as the width parameter translates to the upper bound of the range that we expect with 50% probability. Thus, the default parameter (ω = 0.707), with a truncated Cauchy, expecting a positive effect size, should be described as follows: "*We are 50% confident that the effect size lies somewhere between d = 0 and d = 0.707*".

An objective Bayes approach would require the consistent application of the default parameters on JASP, though it is still essential that the researcher interprets the results in light of the priors. A subjective approach requires expert judgement about the direction and magnitude of the effect, as well as the shape of the distribution that best describes the prior. In addition to changing the width parameter, newer versions of JASP also allow us to shift the location parameter. Here, the expected range (with 50% confidence) can be calculated by adding the amount of the shift to the lower and upper bounds. When we keep the default width parameter (ω = 0.707), but change the location parameter to $d = 0.5$, the interpretation of the prior is: "*We are 50% confident that the effect size lies somewhere between d = -0.207 and d = 1.207*". Furthermore, it is also possible to change the distribution from a Cauchy to a *t*- or a normal distribution (Gronau, Ly, & Wagenmakers, 2020).

Table 1: *Required JASP width (scale) parameters for different effect size ranges, when the researcher is either 50% or 80% confident that the effect size lies within this given range*

Range of effect sizes (non-directional)	Range of effect sizes (directional, halved Cauchy)	JASP Scale parameter for 50% confidence	JASP Scale parameter for 80% confidence
-2 to 2	-2 to 0 or 0 to 2	2	0.65
-1.5 to 1.5	-1.5 to 0 or 0 to 1.5	1.5	0.49
-1.3 to 1.3	-1.3 to 0 or 0 to 1.3	1.3	0.42
-1.1 to 1.1	-1.1 to 0 or 0 to 1.1	1.1	0.36
-0.9 to 0.9	-0.9 to 0 or 0 to 0.9	0.9	0.29
-0.7 to 0.7	-0.7 to 0 or 0 to 0.7	0.7	0.23
-0.5 to 0.5	-0.5 to 0 or 0 to 0.5	0.5	0.16
-0.3 to 0.3	-0.3 to 0 or 0 to 0.3	0.3	0.1

What misconceptions might affect the conclusions drawn from a Bayes factor analysis?

The above sections aim to explain to psychological scientists (both researchers and students) what Bayesian statistics is, how Bayes factors are computed, and how to interpret Bayes factors. We argued that such understanding is necessary to avoid Bayesian statistics inheriting the pitfalls of the *p*-value: usage without a thorough understanding of what these statistics mean. The next question is what kind of misunderstandings may occur among researchers, and what negative consequences might be expected when researchers use Bayes factors without a thorough understanding (see also Tendeiro & Kiers, 2019).

Misconception 1: When reporting the results, writing "Bayes factor > 3" is convincing

Figure 4 shows two plots which demonstrate how neither *p*-values nor Bayes factors should be reported. While both plots present valuable information about the observed mean, they lack information which is required for the reader to determine whether they should trust the conclusion that there is probably a group difference. Both for the *p*-value and for the Bayes

factor, a relevant point for interpreting the reported results is whether *p*- or *B*-hacking may have taken place. In addition, for the Bayes factor, more information about the priors would be needed. As we have seen, the Bayes factor requires us to specify a prior distribution. Even if we use the default prior in JASP, there are some underlying assumptions which may or may not be reasonable for our particular research question, and using different priors may change the conclusions we will draw. Therefore, it is important that a Bayes factor is always interpreted in light of the prior distributions that have been used.

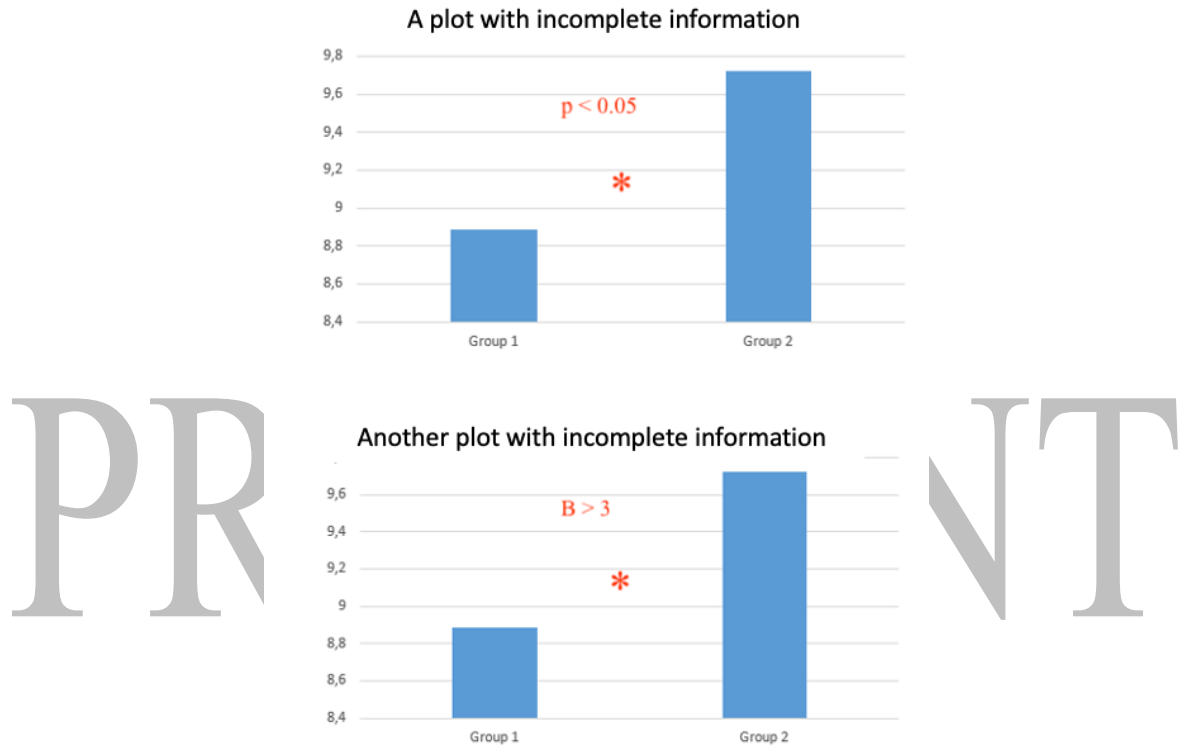


Figure 4. An example of how neither p-values nor Bayes factors should be presented.

Both graphs force the viewer to rely on dichotomous decisions. For the Bayes factor, a range of values greater than 3 encompasses a lot of possible values. The evidence for the hypothesis could range from barely above "anecdotal" ($B = 3.00001$) to "overwhelming" ($B =$ 3,000,000). A second reason why presenting the results of a Bayes factor as in Figure 4 is not recommended is that it loses this information. Treating the Bayes factor as a dichotomy or even trichotomy (H_0 supported, H_1 supported, inconclusive) would mean that it may inherit problems which are associated with the current use of *p*-values. Note that this information is also lost in the graph with the *p*-value: a viewer would be likely to interpret $p = 0.047$ differently than $p < 0.0001$. Critically, the use of a single threshold (be it $B > 3$ or $p < 0.05$) may encourage researchers to try out different analysis strategies to get above this threshold

(*p*- or *B*-hacking). If used as a continuous variable, this does not pose a substantial problem in terms of interpreting the degree of evidence, because *B*-hacking from, say, 2.7 to 3.2 by removing some inconvenient "outliers" would still provide a Bayes factor in the correct order of magnitude (though such practices would lead to systematic over-estimations of the effect sizes). Values which are substantially higher are more difficult to produce by such questionable post-hoc procedures, and also for this reason (in addition to the natural interpretation of the Bayes factor) should be taken to suggest stronger evidence. As Figure 4 would report a *B*-hacked value of 3.2 in an identical manner to a Bayes factor of 3,000, it loses information that is crucial to a correct interpretation. This, in turn, increases the overall impact of *B*-hacking on the conclusions that are drawn about the presence or absence of an effect.

In sum, recommendations about reporting Bayes factors are as follows: First, if accompanied by a figure, the figure should strive to show the distribution of the values (e.g., a violin plot) or an indication of the variability (e.g., labelled error bars), rather than only the central tendency. Note that this is true for any analysis method. Second, whenever a Bayes factor is presented, it should be clear which priors were used. The distribution (e.g., Cauchy or normal) should be stated, as well as its parameters (for a Cauchy, its location and width parameter; for a normal distribution, its mean and SD). At the very least, a statement that the default priors in JASP were used is essential. A justification for a prior would strongly increase the transparency. Third, rather than reporting the threshold which was exceeded by the Bayes factor ($B > 3$), reporting the actual obtained Bayes factor (e.g., 3.24) will encourage a continuous interpretation from the side of the reader. A more detailed set of recommendations for all stages of Bayesian data analysis in JASP can be found in van Doorn et al. (2020).

Misconception 2: Any Bayes factor should be taken at face value, regardless of how it was calculated, or: B-hacking is not a problem

Adjusting priors opens the possibility of changing the priors in the hopes that, eventually, the Bayes factor will provide evidence for the hypothesis that one really wants to be true. If done transparently, performing Bayes factor analyses on the same data but with different priors does not constitute misconduct. In frequentist analysis, such researcher degrees of freedom change the long-run probabilities of false and true positives and negatives, and thus render any results close to meaningless. As Bayesian inference is not based on these long-run

probabilities, it is sometimes argued that a Bayes factor value can be interpreted without a consideration of how many analyses were tried before the final model was decided on (Dienes, 2016). In practice, this argument should be treated with caution and if multiple models were tried, both this fact and the reasoning for deciding on the final model should be transparently reported.

Changing the parameters and assessing how this affects the results is a form of *sensitivity analysis* (or *robustness check*), where the robustness of results can be verified. For example, when we expect a large effect size, but find evidence for the null hypothesis, it could be that there is an effect, but it is smaller than expected and therefore more in line with the null model than with a large-effect-size prior. In a set of simulations, Schönbrodt et al. (2017) showed that, when the effect is small ($\delta = 0.2$), and using the default Cauchy prior ($\omega = 0.707$) and the conventional threshold of $BF < ⅓$ to draw conclusions about the absence of an effect, the false negative rate approaches 80% (using an optional stopping approach, where participants are recruited until the threshold is met): therefore, with a high probability, the researcher will erroneously conclude that the effect is not there when, in reality, there is a true but small effect. Verifying that the evidence for the absence of an effect persists when the prior is lowered would provide a stronger case. A sensitivity analysis can be conducted for a range of H_1 priors that reflect effect sizes that the researcher judges to be practically or theoretically meaningful.

Researchers are likely to overestimate the size of an effect when basing the effect size expectation on previous studies. This is because *p*-hacking and publication bias lead to overestimated effect sizes in the published literature (Gelman & Carlin, 2014). Thus, especially when we observe evidence for the null hypothesis with a relatively lax decision threshold, we might want to conduct a series of additional analyses, where we calculate the Bayes factor for smaller width parameter values. JASP has an inbuilt robustness check, which provides a plot of the evidence for H_1 or H_0 for width parameters ranging from $\omega = 0$ to $\omega = 1.5$. In addition to the plot, it also provides a value denoted "max BF_{10}": this gives the Bayes factor (for the alternative hypothesis) and the width parameter value (ω, denoted *r* in JASP) where the largest Bayes factor is found. The appropriate way to interpret the results of this analysis is *not* to report only this largest Bayes factor, but to provide the whole plot. In many cases, this might mean that we need to conclude that the hypothesis is supported only under a limited range of scenarios and that it may or may not exist.

Thus, conducting numerous analyses and reporting only the "best" one is a problem both for frequentist and Bayesian inference. Bayes factors do not provide a solution to this problem. However, a shift towards reporting robustness analyses might be an effective step towards reducing the crisis in psychological sciences. First, it will raise awareness of the fragility of many results. Second, it will encourage caution in interpreting results, when appropriate. Note that this is not specific to Bayesian statistics: when using *p*-values, especially when post-hoc decisions (e.g., about outlier removal) are involved, it is advisable to report the results from all analyses, and not only those that give significant results. In fact, in frequentist statistics, conducting many analyses until one observes a significant *p*-value, and then reporting only the significant *p*-value, is the definition of *p*-hacking. In an ideal case, to avoid *p*-hacking, the data processing and analysis steps should be defined in advance (e.g., via preregistration), and whatever result is observed using the predefined methods should be the basis of the conclusions. In reality, however, the data often behaves in unexpected ways, which may give legitimate reasons for the researchers to decide, a posteriori, to change the analysis plan. In this case, transparency is the key, which could be considered as exploratory analyses appended to the original analysis plan. The recommendation of assessing the robustness of analyses and reporting them in a paper has been made by Steegen et al., (2016), who refer to this approach as a *multiverse analysis*.

Reporting robustness or multiverse analyses would require a shift of mindset among authors, reviewers, and editors. The conclusion that the data might be equally likely under either model, because different analyses show different results, is often deemed unpublishable. However, making such data available is vitally important. First, even when individual studies yield inconclusive results, combining them in a meta-analysis should provide a more definite answer, in the long run. Second, selective reporting of studies with conclusive results - or, even worse, of analyses which yielded conclusive results even when other analyses of the same data did not show an effect - will distort the evidence for a given effect, by making it appear more robust than it actually is. Whether one is using a Bayesian or a frequentist framework, the observed data is subject to the same error term: by publishing papers and analyses which give the best results, we change the random error term to a systematic one, where studies are more likely to be published when the error term increases the observed effect size compared to the population (i.e., where $d > \delta$).

Misconception 3: The Bayes factor is never wrong

With the *p*-value, if the null hypothesis is true and we perform 10,000 experiments, we expect to get 500 significant results (5% of all experiments if our threshold is $p < 0.05$). This is a feature, not a bug: The side of significance on which the *p*-value falls is sometimes incongruent with whether an effect is actually present in the population, but the *p*-value is designed in such a way that it is wrong in a fixed percentage of experiments, under the assumption of the null hypothesis.[4] The Bayes factor follows a different logic, and as such, concepts such as the false positive rate are not central to its interpretation. However, same as the *p*-value, it can sometimes provide evidence for a model which is different from the population parameters. It is important that researchers are aware of this: a cause for the replication crisis is likely to be an overreliance on significant *p*-values as an indicator for the presence of the effect, while forgetting that they can sometimes occur under the null hypothesis (especially if the null hypothesis is highly likely a priori; see Figure 1).

The proportion of times a Bayes factor gives the wrong conclusion depends on (1) the effect size, (2) the sample size, (3) the prior parameters, and (4) the threshold at which one draws a conclusion (e.g., BF > 3 or BF > 10). Schönbrodt et al. (2017; see their Table 1) conducted a series of simulations to estimate the percentage of erroneous conclusions for a wide range of these three parameters, assuming an approach where sample size is determined by testing participants until a BF threshold is hit. For an effect size of $\delta = 0$, the default prior width parameter of $\omega = 0.707$, and the standard threshold of BF > 3, the false positive rate is 7.5%. Per se, this is not a fatal blow for the use of Bayes factor: as any statistic, it is susceptible to noise, and there will always be occasional data sets which seem to provide evidence for a wrong conclusion. However, it has implications for interpreting Bayes factors.

In particular, it is important to bear in mind that, if we conduct 10,000 studies where the null hypothesis is true and analyse them with the default Bayes factor in JASP, we will have some Bayes factors which are greater than 3. Note that, as sample size increases, such false positives will become rarer: this is an important difference compared to *p*-values, where the false positive rate is, by definition, fixed by the alpha-level and is independent of sample size. If we determine the sample size by sequential testing until we reach the desired threshold, we

[4] *p*-hacking, by definition, invalidates this feature, and may increase the false positive rate to up to 60% (Simmons, Nelson, & Simonsohn, 2011).

will obtain, on average, 750 false positives. In practice, false positives will become problematic when researchers conduct multiple comparisons. If the variables are not correlated (Bishop & Thompson, 2016), a researcher who has given 7 different tests to the participants and continues testing until at least one of them provides evidence for a group difference (BF > 3) already has a >50% chance, in the long run, to find support for at least one alternative hypothesis, even if the null hypothesis is true in all cases (7 * 0.075 = 0.525). Same as with *p*-values, this becomes a problem with selective reporting of variables or HARKing (hypothesising after results are known; Kerr, 1998).

Misconception 4: "Power" for Bayes factor analyses

When using frequentist statistics, a power analysis is mostly required by editors of Registered Reports, by ethics committees, or for grant proposals. The rationale is that we want to maximise the chance of being able to draw meaningful conclusions after the completion of a study: finding a non-significant *p*-value with a statistical power of 10% leaves us in the position where we learned very little about the posterior probability of the null hypothesis being true (the reader is encouraged to substitute a power of 0.1 to calculate a posterior probability of H_1, using Bayes' Theorem introduced in section "What does Bayesian statistics mean?"). Thus, considering power is important, from a frequentist perspective, to ensure an effective use of resources.

When we conduct a Bayes factor analysis, it is equally important that we maximise the chance to draw meaningful conclusions. However, the terminology "power" is defined from a frequentist perspective, and as such does not directly apply to the Bayesian framework. This is because the power is a long-run probability, and therefore an inherently frequentist concept. It is, however, possible to compute the long-run probabilities of false positives and false negatives. For Bayes factor analyses, these depend, as in the frequentist framework, on the population effect size δ (in the case of a false negative - for a false positive, by definition, δ = 0), but also on the prior parameters and the BF threshold which one takes as conclusive evidence. A table with false positive and false negative probabilities, for a wide range of scenarios, is presented by Schönbrodt et al. (2017).

One can also ask the question: How many participants do I need to test if I want to have a high long-run probability that I will obtain a Bayes factor exceeding a certain threshold? Again, this depends on the population effect size δ, on the prior parameters, and on the

decision threshold. Schönbrodt et al. (2017) provide the "average sample number": the average number of simulated participants that were required before the Bayes factor exceeded the threshold for a variety of effect sizes and prior parameters. These numbers are a very useful guideline to a researcher justifying their sample size and priors to an editor, funder, or ethics committee, and should be consulted while planning an experiment to maximise the probability of getting a meaningful result. However, given the continuous interpretation of a Bayes factor, even inconclusive data might be useful in the long term. When the Bayes factor provides only weak or anecdotal evidence, authors should refrain from drawing strong conclusions. However, researchers who have limited resources and cannot maximise their chance to draw strong conclusions can interpret a weak or anecdotal Bayes factor and contribute to the literature by providing a data set which can be included in meta-analyses.

Should we switch from *p*-values to Bayes factors?

An increasing number of publications recommend the use of Bayesian statistics instead of or in addition to frequentist inference (e.g., Dienes, 2011, 2014, 2016; Rouder et al., 2009; Wagenmakers, 2007). The desire for alternatives to the *p*-value is exacerbated by the crisis in psychological science, which resulted from a realisation that many results may not be true, even those that made it into undergraduate psychology textbooks (Open Science Collaboration, 2015). At the same time, the availability of easy-to-use software allows even researchers with little knowledge of Bayesian statistics to calculate Bayes factors. While learning about novel analysis methods is always advantageous, it is important to ensure that the psychological science community is aware of the problems that can and cannot be addressed by Bayesian statistics.

The premise of the current paper is that the researcher needs to understand the logic behind the test beyond knowing how to calculate it, no matter which statistical tool is used for inference (Gigerenzer, 2004). The current paper aims to provide an introduction to the logic and pitfalls behind Bayes factor to a psychology researcher or student with limited knowledge of probability and statistics. For practical reasons, the current introduction is limited in terms of its comprehensiveness. The interested reader is encouraged to read further articles on Bayesian statistics to get a fuller understanding: an annotated reading list is provided by Etz, Gronau et al. (2018).

What can Bayes factors do that the p-value cannot?

Both *p*-values and Bayes factors can be used correctly or incorrectly: an incorrect use of either statistic is problematic, but not in itself a valid argument to use one over the other. However, compared to *p*-values, the Bayes factor has several features that might allow for more appropriate inferences. First, the *p*-value is defined as a conditional probability for the case that the null hypothesis is true: it does not take the probability of the data under the alternative hypothesis into account. To calculate a Bayes factor, the researcher needs to take into account an alternative hypothesis: as most psychological researchers are technically interested in an alternative hypothesis (e.g., the presence of a treatment effect), this arguably allows for a closer mapping between the researcher's question and the statistical method. In fact, the posterior probability of H1 derived from the Bayes factor reflects the probability of H1 being true given the observed data (see section converting the Bayes factor to posterior odds).

Second, the Bayes factor encourages a continuous interpretation of the results. In the context of frequentist statistics, the *p*-value is often treated as a tool to make a binary decision: if we want to fix the long-term false-positive rate at 5%, $p < 0.05$ means that we can reject the null hypothesis, while $p \geq 0.05$ means that we cannot[5]. The interpretation of the *p*-value as the long-term error rate requires a dichotomous interpretation. In contrast, the Bayes factor allows for a continuous interpretation: values above 1 provide "anecdotal evidence"; values about 3 "some evidence", and values above 10 "strong evidence". These labels might encourage researchers to interpret the strength of evidence, and to be more tentative about their conclusions with a BF > 3 than when they observe a BF > 3,000. This also applies to interpreting the evidence for effects in the published literature: when building on effects which are supported with relatively small Bayes factor values the researcher may want to replicate the effect before building on it.

Third, the Bayes factor allows for inferences in favour of a null hypothesis, while a frequentist approach can only be used to reject a null hypothesis. By taking into account both an H_0 and H_1 model, the Bayes factor allows us to quantify the degree of evidence for the H_0

[5] A continuous interpretation of the *p*-value as strength of evidence has been proposed by Fisher. This interpretation is contentious, and incompatible with a frequentist interpretation. Some researchers will, however, rely on the intuitive notion that $p = 0.00001$ should be treated differently from $p = 0.048$, while simultaneously relying on the frequentist framework (Gigerenzer, 2004).

model, relative to the specified H_1. Psychology suffers from the selective publication of positive (i.e., significant) results (Rosenthal, 1979): this significance filter leads to the illusion of consistent evidence for an effect, even if the real effect is zero. One reason for this selective publication may be that a non-significant *p*-value cannot be interpreted as evidence of absence, thus yielding frequentist null-results difficult to interpret and thus difficult to use in theory building or for decision making (e.g., Schmalz & Mulatti, 2017). The ability to draw inferences about a null hypothesis may alleviate this problem. However, it is important to bear in mind that null results may be uninterpretable for non-statistical reasons: for example, poor measurement will yield both frequentist and Bayesian null-results uninterpretable, as it will be unclear whether a lack of a group difference or correlation reflects the absence of the effect or the poor measurement (Spearman, 1904). The ability of the Bayes factor to support the null hypothesis also comes with the caveat that the conclusions might not prove robust under different H_1 models, and thus, a sensitivity analysis with different priors is recommended to provide evidence for H_0.

What can Bayes factors not do?

p-values and Bayes factors share some of their pitfalls. Neither should be applied without some understanding of the underlying principles. In the case of *p*-values, research has shown that a majority of researchers have misconceptions about *p*-values which lead to inappropriate data analysis and inferences (John, Loewenstein, & Prelec, 2012; Simmons et al., 2011). This continuing misuse of *p*-values has contributed to a crisis of confidence in psychological science (Pashler & Wagenmakers, 2012). Bayes factors may prove more robust than *p*-values in the long term. However, this should not be taken for granted: simply switching from calculating *p*-values to calculating Bayes factors without understanding how true or robust these results are may result in misunderstandings that potentially have systematic detrimental effects on the research literature. For example, as explained above, trying out different priors and reporting the analysis that gives the biggest Bayes factor, in conjunction with publication bias, where inconclusive results are not published, will yield a literature with overestimated effect sizes for true effects, and an erroneous idea that there is convincing evidence for or against a given effect.

It has been argued that Bayes factors are immune to some of the questionable research practices that, in many cases, make the *p*-value uninterpretable. Examples of such *p*-hacking practices are selective reporting of variables, optional stopping (testing a number of

participants, calculating a *p*-value, and continuing to collect data if $p > 0.05$), or removing outliers conditional on $p > 0.05$. If Bayes factors are interpreted as continuous degrees of evidence, such practices are not problematic (Dienes, 2016; Rouder, 2014). This interpretation, however, might be counterintuitive to a researcher who is used to thinking of statistical test results in terms of making a binary decision. The bigger the Bayes factor, the smaller the probability of a false positive or false negative (Schönbrodt et al., 2017). However, if researchers treat both $BF_{10} = 3.1$ and $BF_{10} = 31$ as evidence for the alternative hypothesis, they might miss the point that a false positive is more likely when the former is observed. This is problematic when researchers analyse many variables and report only those that exceed a threshold. In an exploratory setting, a $BF > 30$ might be very rare unless there really is a true effect in the population, but one is likely to come across a $BF > 3$ even if the data consists of noise. Thus, selective reporting and multiple comparisons are likely to become problematic in cases when the Bayes factors are relatively small.

Both *p*-values and Bayes factors are tools for hypothesis testing: they provide a means for selecting between two models. This is orthogonal to effect size estimation (Kruschke & Liddell, 2018, but see also Keirs & Tendeiro, 2019; Rouder, Haaf, & Vandekerckhove, 2018). Effect size estimation can occur either via a frequentist or Bayesian framework: in the former case, this involves calculating the Maximum Likelihood Estimator or best linear unbiased estimator and plotting 95% confidence intervals around it (Cumming, 2014). In the latter case, a posterior distribution is calculated by combining the prior distribution with incoming data (Kruschke & Liddell, 2018). The peak of this distribution is the estimate of the effect size, and the bounds encompassing 95% of the area constitute the 95% credibility interval.

Hypothesis testing is likely to be more popular in psychological science than effect size estimation, because most hypotheses in psychology are only directional (Meehl, 1990): we predict that $\mu_1 < \mu_2$, but not by how much. Therefore, finding an effect size of $d = 0.7$ would not corroborate a theory any more or less than finding an effect size of $d = 0.2$. Note, however, that even if there is no theoretical prediction about the magnitude of an effect size, consideration of effect sizes is still required in practice for power calculations and required sample size estimations. Directional hypotheses make theories in psychological sciences difficult or even impossible to falsify (Meehl, 1990). Point hypotheses, in contrast, constrain the range of plausible values, such that showing that experimental results are inconsistent

with the plausible range provides evidence against the theory. With an accumulating amount of data and use of computational models, point hypotheses should be possible for some research questions. For example, one might have a theory that predicts that an effect X should be smaller than a well-established effect Y but greater than zero. When the mechanisms are well-understood and the measurement error is relatively small, computational models of these mechanisms might be able to provide such estimation.

Despite being suitable for testing directional hypotheses, both the *p*-value and the Bayes factor require the researcher to think about the effect size: in the case of frequentist statistics, for a power calculation, and in the case of the Bayes factor, for the specification of the prior (Rouder et al., 2016). The need to specify the prior might further push researchers to think about effect sizes, which would be a step towards thinking about point or range predictions rather than only testing the direction of an effect.

If effect size estimation (e.g., in a meta-analysis) is the ultimate goal (Schmidt, 1996), questionable research practices might be problematic both for *p*-values and for Bayes factors. With the example of optional stopping, we might test 20 participants; we observe a *p*-value of 0.07 or a Bayes factor of 2.6. We continue testing 5 participants at a time, after each batch of 5 participants we re-run the analyses and stop if the *p*-value or Bayes factor exceed the threshold. In both cases, the decision to terminate data collection is conditional on exceeding this threshold. This can happen for one of two reasons: It is possible that the observed effect size approaches the non-zero population effect size, in which case having collected more data would have helped to converge to the correct answer. The other possibility is that we happened to have tested a batch of participants with particularly high observed effects due to the error term. If the error term had resulted in a lower effect size, or cancelled out across the five participants, the test statistic would be less likely to reach the threshold. Thus, while the error term is generally considered to be random, conditioning on exceeding a threshold, which corresponds to larger observed effect sizes, leads to a slight but systematic overestimation of the effect size, in the long run. This may not be problematic to interpreting the Bayes factor; however, if there is a publication filter such that studies or analyses with Bayes factor values below 3 are not published, this will lead to a systematic inflation of effect sizes in the published literature.

Conclusion and final recommendations

Bayes factors, like *p*-values, come with a set of nuances and caveats. Using them effectively and convincingly requires an understanding of the basic concepts behind the Bayes factor, the way in which the prior is calculated, as well as a series of basic probability concepts that affect both the interpretation of *p*-values and of Bayes factors. The current paper aimed to provide an overview of such concepts to researchers or students with basic statistics knowledge.

A shift from frequentist to Bayesian statistics is unlikely to be sufficient to alleviate the symptoms of the replication crisis in psychology. The current article discusses the consideration of the a priori probability of a hypothesis, consideration of the expected effect sizes, and transparency. These are likely to be effective methods to increase the credibility of the literature in psychological sciences. Considering the a priori probability of a hypothesis involves building closely on theories and previous, credible work. This will reduce the *a posteriori* probability that a dataset, showing evidence for an effect, is not a false positive. The consideration of expected effect sizes will help researchers in planning an experiment, but it will also allow researchers to be sceptical about unrealistically big effect sizes reported in the literature (Gelman & Weaklim, 2009). One efficient way to encourage researchers to think about expected effect sizes is through the use of Registered Reports (Chambers et al., 2015). A Registered Report is a manuscript, consisting of an introduction, methods, and planned analysis section, that is submitted to a journal *before* data collection. As a Registered Reports requires the authors to outline the data processing and analysis plan, it will also increase transparency about the analyses that have been conducted. Changes to this plan may be required if the data behaves in unexpected ways, but these can be reported as exploratory analyses. To maximise the robustness of a paper, a sensitivity analysis is recommended.

We do not conclude that one way of analysing the data is inherently better than the other, but that both approaches can lead to either appropriate or inappropriate conclusions, depending on the way which they are used. At the same time, we encourage any researcher or student who is curious about Bayesian statistics to learn more about them. Three books which provide an introduction to Bayesian thinking, though not necessarily with a focus on Bayes factors, are Lambert, 2019; McElreath, 2020; Kurt, 2019; and van de Schoot et al., 2021. Etz, Gronau et al., 2018 also provide a list of articles varying in prerequisite knowledge. There are several reasons for learning more about Bayesian statistics in general, and Bayes factors

specifically: First, their use in psychological science is likely to increase due to the publicity and development of easy-to-use software, meaning that a researcher is likely to come across an article which used Bayes factors for inference. An understanding of the pitfalls will provide the researcher with the means to evaluate whether the conclusions are justified. Second, learning about Bayesian statistics is bound to also improve their understanding of frequentist statistics and their pitfalls. Third, with this additional statistical tool, the researcher will increase the flexibility in addressing research questions: one clear advantage of the Bayes factor is that it can be also used to provide evidence for the absence of an effect.

Glossary box

Terminology	Definition
B-hacking	Misuse of the Bayes factor, where different analyses are tried and the researcher selectively reports only the analyses with the biggest Bayes factor.
Bayes factor	A method of model selection which can be used for hypothesis testing and relies on quantifying the support for one model over another.
Bayes' Theorem	The formula that allows us to invert a *conditional probability*, i.e., to get from P(A|B) to P(B|A).
Bayesian statistics	A philosophy underlying the concept of probability, where probability can be seen as a degree of belief, which is obtained by combining prior knowledge with incoming evidence.
Cauchy distribution	A probability distribution, similar in shape to the normal (Gaussian) distribution, with two parameters defining its shape: a location parameter and width parameter (ω). This distribution is a popular choice for a prior for the computation of Bayes factors.
Cohen's *d*	A standardised measure of effect size, where a mean difference between two conditions is divided by an estimate of the standard deviation.
Conditional probability	The probability of an event (E) under the assumption that a given condition (C) is true: for example, the true positive is a conditional probability describing the probability of the disease (E) only in the subset of people who test positive on this test (C), written as P(E|C).
Frequentist statistics	An inferential framework, where probabilities are defined as the proportion of times an event would happen, under specified assumptions, with infinitely repeated sampling.

JASP	An open-source software that has an intuitive user interface which allows researchers to calculate Bayes factors as well as frequentist statistics for a wide range of experimental designs.
p-hacking	Misuse of the *p*-value, where different analyses are tried and the researcher selectively reports only the analyses with the significant *p*-values.
Model prior	Prior belief of hypothesis, before data is acquired.
Parameter prior	Prior belief of the model parameters (e.g., effect size), before data is acquired.
Posterior	Posterior belief of the hypothesis in light of data.

Author note:

A preprint of the article is available at https://osf.io/vgqbt/. A round of informal peer review was solicited before submission via Twitter and Facebook. We would like to thank Stephen Benning, Ondřej Kudláček, Borysław Paulewicz, Phillip Schäpers, Jorge Tendeiro, and Eric-Jan Wagenmakers for their helpful comments on an earlier version of this manuscript. At the time of writing, X.S. was supported by a grant from the Deutsche Forschungsgemeinschaft (SCHM3450/1-1) and by the Programme “Fellows Freies Wissen” from the Stifterverband, Wikimedia, and Volkswagenstiftung (H190 5909 5096 32135). Parts of the manuscript were prepared while XS was a visiting researcher at the University of Trento. L.Z. was partially supported by the Vienna Science and Technology Fund (WWTF VRG13-007).

References

Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407.

Bishop, D. V., & Thompson, P. A. (2016). Problems in using *p*-curve analysis and text-mining to detect rate of *p*-hacking and evidential value. *PeerJ*, *4*, e1715.

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

Cohen, J. (1992). Things I have learned (so far). *American Psychologist, 45(12),* 1304-1312.

Colling, L. J., & Szűcs, D. (2018). Statistical Inference and the Replication Crisis. *Review of Philosophy and Psychology*, 1-27.

Consonni, G., Fouskakis, D., Liseo, B., & Ntzoufras, I. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis, 13,* 627-679.

Cumming, G. (2014). The new statistics: why and how. *Psychological Science*. 25, 7–29. doi: 10.1177/0956797613504966

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*(3), 274-290.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781.

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

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2018). How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review*, *25*(1), 219-234.

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*(2), 281-295.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641-651.

Gelman, A., & Weakliem, D. (2009). Of beauty, sex and power: Too little attention has been paid to the statistical challenges in estimating small effects. *American Scientist*, *97*(4), 310-316.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587-606.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*(2), 53-96.

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2020). Informed Bayesian *t*-tests. *The American Statistician, 74*, 137-143.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle *p* value generates irreproducible results. *Nature Methods*, *12*(3), 179.

Harms, C., & Lakens, D. (2018). Making 'null effects' informative: statistical techniques and inferential frameworks. *Journal of Clinical and Translational Research*, *3*(2), 382.

Jeffreys, H. (1961). *Theory of probability* (3rd edition). New York, NY: 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*(5), 524-532.

Kiers, H., & Tendeiro, J. (2019, April 5). With Bayesian Estimation One Can Get All That Bayes factors Offer, and More. https://doi.org/10.31234/osf.io/zbpmy

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196-217.

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, *25*(1), 155-177.

Kurt, W. (2019). *Bayesian Statistics the Fun Way: Understanding Statistics and Probability with Star Wars, LEGO, and Rubber Ducks*. No Starch Press, San Francisco, USA.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology, 4*. 1-12. doi: https://doi.org/10.3389/fpsyg.2013.00863

Lakens, D. (2017). Equivalence tests: a practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355-362.

Lambert, B. (2018). *A student's guide to Bayesian statistics*. Sage, Los Angeles, USA.

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., ... & Matzke, D. (2019). JASP: graphical statistical software for common statistical designs. *Journal of Statistical Software*, *88*(2).

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological Bulletin*, *109*(2), 163.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press, Boca Raton, USA.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(1), 195-244.

Morey, R., Rouder J.N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2018). Package 'BayesFactor' 0.9.12-4.2. https://cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf.

Mulder, J., & Wagenmakers, E. J. (2016). Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments". *Journal of Mathematical Psychology*, *72*, 1-5.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

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*(6), 528-530.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, *86*(3), 638-641.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301-308.

Rouder, J.N. (2016). Roll your own: How to compute Bayes factors for your own priors [blog post]. http://jeffrouder.blogspot.com/2016/01/what-priors-should-i-use-part-i.html.

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, *25*(1), 102-113.

Rouder, Jeffrey N. and Morey, Richard D. 2019. Teaching Bayes' Theorem: strength of evidence as predictive accuracy. *The American Statistician 73 (2)*, 186-190. doi: 10.1080/00031305.2017.1341334

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E. J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science*, *8*(3), 520-547.

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*(2), 225-237.

Schmalz, X., & Mulatti, C. (2017). Busting a myth with the Bayes Factor: Effects of letter bigram frequency in visual lexical decision do not reflect reading processes. *The Mental Lexicon*, *12*(2), 263-282.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*(2), 115.

Schönbrodt, F. D., Wagenmakers, E. J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322.

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*(11), 1359-1366.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15(1)*, 72-101.

Spiegelhalter, D. (2019). *The art of statistics: Learning from data*. Penguin, UK.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702-712.

Tendeiro, J. N., & Kiers, H. A. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods, 24(6),* 774-795.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., ... & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*(1), 1-26.

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., ... & Wagenmakers, E. J. (2020). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*. doi: 10.3758/s13423-020-01798-5

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician, 70(2), 129-133,* doi:10.1080/00031305.2016.1154108

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158-189.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*(5), 779-804.