Psychological Methods

Workflow Techniques for the Robust Use of Bayes Factors

Daniel J. Schad, Bruno Nicenboim, Paul-Christian Bürkner, Michael Betancourt, and Shravan Vasishth

Online First Publication, March 10, 2022. http://dx.doi.org/10.1037/met0000472

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

ISSN: 1082-989X

https://doi.org/10.1037/met0000472

Workflow Techniques for the Robust Use of Bayes Factors

Daniel J. Schad[1, 2, 3], Bruno Nicenboim[2, 3], Paul-Christian Bürkner[4], Michael Betancourt[5], and Shravan Vasishth[3]

[1] Department of Psychology, Health and Medical University Potsdam
[2] Department of Cognitive Science and Artificial Intelligence, Tilburg University
[3] Department of Linguistics, University of Potsdam
[4] Cluster of Excellence SimTech, University of Stuttgart
[5] Symplectomorphic, New York, New York, United States

Abstract

Inferences about hypotheses are ubiquitous in the cognitive sciences. Bayes factors provide one general way to compare different hypotheses by their compatibility with the observed data. Those quantifications can then also be used to choose between hypotheses. While Bayes factors provide an immediate approach to hypothesis testing, they are highly sensitive to details of the data/model assumptions and it's unclear whether the details of the computational implementation (such as bridge sampling) are unbiased for complex analyses. Here, we study how Bayes factors misbehave under different conditions. This includes a study of errors in the estimation of Bayes factors; the first-ever use of simulation-based calibration to test the accuracy and bias of Bayes factor estimates using bridge sampling; a study of the stability of Bayes factors against different MCMC draws and sampling variation in the data; and a look at the variability of decisions based on Bayes factors using a utility function. We outline a Bayes factor workflow that researchers can use to study whether Bayes factors are robust for their individual analysis. Reproducible code is available from https://osf.io/y354c/.

Translational Abstract

In psychology and related areas, scientific hypotheses are commonly tested by asking questions like "is [some] effect present or absent." Such hypothesis testing is most often carried out using frequentist null hypothesis significance testing (NHST). The NHST procedure is very simple: It usually returns a *p*-value, which is then used to make binary decisions like "the effect is present/absent." For example, it is common to see studies in the media that draw simplistic conclusions like "coffee causes cancer," or "coffee reduces the chances of getting cancer." However, a powerful and more nuanced alternative approach exists: Bayes factors. Bayes factors have many advantages over NHST. However, for the complex statistical models that are commonly used for data analysis today, computing Bayes factors is not at all a simple matter. In this article, we discuss the main complexities associated with computing Bayes factors. This is the first article to provide a detailed workflow for understanding and computing Bayes factors in complex statistical models. The article provides a statistically more nuanced way to think about hypothesis testing than the overly simplistic tendency to declare effects as being "present" or "absent."

Keywords: Bayes factors, Bayesian model comparison, prior, posterior, simulation-based calibration

Supplemental materials: https://doi.org/10.1037/met0000472.supp

In the cognitive sciences and related areas, recent years have seen a rise in Bayesian approaches to data analysis. Many cognitive science journals have published special issues on Bayesian data analysis, including methodological journals such as the *Journal of Mathematical Psychology* (Lee, 2011; Mulder & Wagenmakers, 2016) and *Psychological Methods* (Chow & Hoijtink,

Daniel J. Schad https://orcid.org/0000-0003-2586-6823
Bruno Nicenboim https://orcid.org/0000-0002-5176-3943
Paul-Christian Bürkner https://orcid.org/0000-0001-5765-8995
Michael Betancourt https://orcid.org/0000-0002-2900-0931
Shravan Vasishth https://orcid.org/0000-0003-2027-1994

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project 317633480, SFB 1287 and by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 - 3907 40016.

The data are available at https://osf.io/y354c/

The experiment materials are available at https://osf.io/y354c/

Correspondence concerning this article should be addressed to Daniel J. Schad, Department of Psychology, Health and Medical University, Olympischer Weg 1, 14471 Potsdam, Germany. Email: danieljschad@gmail.com

2017; Hoijtink & Chow, 2017), but also the more experimental journal *Psychonomic Bulletin & Review* (Vandekerckhove et al., 2018). Other introductory articles have been contributed (see Doorn et al., 2021; Etz et al., 2018; Etz & Vandekerckhove, 2018; Nicenboim & Vasishth, 2016; Sorensen et al., 2016; Vasishth et al., 2018). That Bayesian analyses are so prominently discussed and used is an indication that Bayesian approaches are becoming increasingly mainstream (Gelman et al., 2013).

Bayesian data analysis plays an important role in cognitive science as it allows us to quantify the evidence in favor of one hypothesis over another, while taking the uncertainty of the parameters into account. Such Bayesian hypothesis testing can be implemented using Bayes factors (Gronau et al., 2017; Heck et al., in press; Jeffreys, 1939; Kass & Raftery, 1995; Rouder et al., 2018; Wagenmakers et al., 2010; for a critique of Bayes factors see Navarro, 2019; for Bayes factor design planning see Schönbrodt & Wagenmakers, 2018; Stefan et al., 2019).

Although Bayes factors are increasingly being used in the cognitive sciences and other fields of science (Heck et al., in press), there are several important issues with Bayes factors that researchers should take into account:

Issue 1: Bayes Factors in Complex Statistical Models Can Be Unstable

For most interesting problems and models in cognitive science, Bayes factors cannot be computed analytically. Instead, approximations are needed. One major approach is to estimate Bayes factors based on posterior MCMC draws via Savage–Dickey method (Dickey & Lientz, 1970) or bridge sampling (Bennett, 1976; Meng & Wong, 1996), implemented in the R package bridgesampling (Gronau et al., 2020). The approximate Bayes factor estimate may be unstable if insufficient MCMC draws are used, leading to different Bayes factors each time the analysis is performed (see Gronau et al., 2020). This sensitivity of the estimator to the particular Markov chain realization is also known as the variance of the estimator.

Issue 2: Bayes Factor Estimates Can Be Biased

Even if the estimation of Bayes factors via bridge sampling yields stable results, it is still unclear whether the computations are accurate or biased for complex problems, that is, whether the approximate Bayes factor estimate actually corresponds to the true Bayes factor. This stable error in the estimator is also known as the bias of the estimator. This potential bias is concerning, as—for realistic complex models—there is no guarantee that the Bayes factor estimate we obtain is correct. It is therefore crucial to calibrate Bayes factor estimates, which we do in the present work.

Issue 3: Bayes Factor Estimates Can Vary Dramatically Due to Sampling Variability

Any variability that is present in the data will also impact the results from Bayes factor analyses. Any inferences and decisions will always depend on the particular details of observed data and there's no way around that. Accordingly, computing Bayes factors does not mean that we can obtain some abstract and reliable "truth" from some observed data, which is still sampled with considerable noise. Bayes factors—just like frequentist *p*-values or any quantification of evidence—can vary considerably between replications of the same experiment. Excessive variation is a consequence of properties of the experimental design (the strength of the experimental manipulation, sample size, measurement error, etc.), which limits the conclusions that can be drawn from individual data sets (Oelrich et al., 2020).

Issue 4: Bayes Factors Can Be Highly Sensitive to Priors

The results of Bayes factor analyses can be highly sensitive to and crucially depend on prior assumptions about model parameters even in cases where posterior distributions remain virtually unchanged (Aitkin, 1991; Gelman et al., 2013; Grünwald, 2000; Liu & Aitkin, 2008; Myung & Pitt, 1997; Vanpaemel, 2010). That is, in Bayesian inference, researchers specify a priori assumptions about which parameter values they consider most likely before seeing the data. The same likelihood and data with different priors can lead to very different Bayes factors.

Issue 5: Discovery Claims Using Bayes Factors Require Decision Theory

We should not confuse inferences with decisions. Bayes factors provide inference on hypotheses. However, to obtain discrete decisions, such as to claim discovery, from continuous inferences in a principled way requires *utility functions*.

In order to quantify the performance of a decision-making process, we need to be able to quantify the consequences of any particular outcome. For example, what are the costs of choosing the wrong model and what are the benefits of choosing the right model? A utility function makes this quantification explicit, assigning to each outcome a total value for the total benefits less the costs. By studying the range of utilities that a decision-making process can generate, we can understand how useful we expect it to be in the context of any particular application.

In contrast to the utility-function based approach, common decision heuristics (e.g., using Bayes factor larger than 10 as a discovery threshold, as suggested by Jeffreys, 1939) do not provide a principled way to perform decisions; they are merely conventions and do not allow us to determine the consequences of making a discrete decision. Frequentist null hypothesis significance testing, for example, bases testing not on inferences but rather on false discovery rates and true discovery rates; these are examples of utility functions. To ensure that Bayes factors inform useful hypothesis tests, we need to define relevant utility functions and investigate the performance of Bayes factors in that context.

In the cognitive sciences, because it is often unclear how to define good utility functions, we argue that Bayesian decision making is premature: Inferences based on continuous Bayes factors should be reported instead of decisions.

In this article, we investigate these five different aspects of the performance of Bayes factors. A longer version of this article, with more details, including an example workflow analysis and R code to implement some of the analyses is available on arXiv (arXiv:2103.08744).

A Quick Review of Bayesian Methodology

Statistical analyses in the cognitive sciences often pursue two goals: to estimate parameters and to test hypotheses. Both of these goals can be achieved using Bayesian data analysis. Bayesian analyses focus on an "observational" model $\mathcal{M}$, which specifies the probability density of the data y given the vector of model parameters Θ and the model $\mathcal{M}$, i.e., $p(y \mid \Theta, \mathcal{M})$, or by dropping the model, $p(y \mid \Theta)$. It is possible to use the observational model to simulate data, by selecting some model parameters Θ and drawing random samples for the data $\tilde{y}$. When the data is given (fixed, e.g., observed or simulated), then the observational model turns into a likelihood function: $p(y \mid \Theta) = L_y(\Theta)$; this can be used to estimate model parameters or to compute evidence for the model relative to other models. To estimate parameters, in Bayesian data analyses the likelihood is complemented by the prior model, written $p(\Theta)$, that defines a probability distribution that encodes domain expertise. Bayes' rule specifies how to combine the likelihood and the prior to compute the posterior distribution of the model parameters $p(\Theta \mid y, \mathcal{M}_1)$:

$$p(\Theta \mid y, \mathcal{M}_1) = \frac{p(y \mid \Theta, \mathcal{M}_1) p(\Theta \mid \mathcal{M}_1)}{p(y \mid \mathcal{M}_1)} \tag{1}$$

Here, $p(y \mid \mathcal{M}_1)$ is a normalizing constant termed the "evidence" or "marginal likelihood," which is the likelihood of the data y based on the model $\mathcal{M}_1$ independent of the parameters Θ. The marginal likelihood can only be interpreted relative to other models. It is derived as $p(y \mid \mathcal{M}_1) = \int p(y \mid \Theta, \mathcal{M}_1) p(\Theta \mid \mathcal{M}_1) d\Theta$. Note that there is a key role of priors in this computation. Priors play an important role in Bayesian inference as they can regularize inferences when the data do not inform the likelihood functions strongly. However, they will influence marginal likelihoods even when the data are strongly informative, and are thus even more crucial for Bayes factors than for posterior distributions (Aitkin, 1991; Gelman et al., 2013; Grünwald, 2000; Liu & Aitkin, 2008; Myung & Pitt, 1997; Vanpaemel, 2010).

For very simple models, posterior density functions can be computed analytically. However, for most interesting models this is not possible and we have to rely on approximate methods. One such approach is sampling methods such as Markov chain Monte Carlo sampling, which is the method behind popular software implementing Bayesian analysis such as Stan (Carpenter et al., 2017), JAGS (Plummer, 2003), WinBUGS (Lunn et al., 2000), PYMC3 (Salvatier et al., 2016), Turing (Ge et al., 2018), and others.

Inference Over Hypotheses

Bayes factors provide a way to compare any two model hypotheses (i.e., arbitrary hypotheses) against each other by comparing their marginal likelihoods (Betancourt, 2018; Kass & Raftery, 1995; Ly et al., 2016). The Bayes factor tells us, given the data and the model priors, how much we need to update our relative belief between the two models.

To derive Bayes factors, we first compute the model posterior, that is, the posterior probability for a model $\mathcal{M}_i$ given the data: $p(\mathcal{M}_i \mid y) = \frac{p(y \mid \mathcal{M}_i) \times p(\mathcal{M}_i)}{p(y \mid \mathcal{M}_1) \times p(\mathcal{M}_1) + p(y \mid \mathcal{M}_2) \times p(\mathcal{M}_2)} = \frac{p(y \mid \mathcal{M}_i) \times p(\mathcal{M}_i)}{p(y)}$. Here, $p(\mathcal{M}_i)$ is the prior probability for each *model i*. Based on the posterior model probability $p(\mathcal{M}_i \mid y)$, we can compute the model odds for one model over another as:

$$\frac{p(\mathcal{M}_1 \mid y)}{p(\mathcal{M}_2 \mid y)} = \frac{[p(y \mid \mathcal{M}_1) \times p(\mathcal{M}_1)]/p(y)}{[p(y \mid \mathcal{M}_2) \times p(\mathcal{M}_2)]/p(y)} = \frac{p(y \mid \mathcal{M}_1)}{p(y \mid \mathcal{M}_2)} \times \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)} \tag{2}$$

$$\textit{Posterior ratio} = \textit{Bayes factor} \times \textit{prior odds} \tag{3}$$

The Bayes factor is thus a measure of relative evidence, the comparison of the predictive performance of one model ($\mathcal{M}_1$) against another one ($\mathcal{M}_2$). This comparison (BF_{12}) is a ratio of marginal likelihoods:

$$BF_{12} = \frac{P(y \mid \mathcal{M}_1)}{P(y \mid \mathcal{M}_2)} \tag{4}$$

BF_{12} indicates the evidence that the data provide for $\mathcal{M}_1$ over $\mathcal{M}_2$, or in other words, which of the two models is more likely to have generated the data, or the relative evidence that we have for $\mathcal{M}_1$ over $\mathcal{M}_2$. Under the assumption that all models are equally likely a priori, Bayes factor values larger than one indicate that $\mathcal{M}_1$ is more compatible with the data, values smaller than one indicate $\mathcal{M}_2$ is more compatible with the data, and values close to one indicate that both models are equally compatible with the data.

In the present work, we will consider the case of nested model comparison, where a null model hypothesizes that a model parameter is zero or absent (a point hypothesis: $p(\Theta_1 = 0 \mid y)$),[1] whereas an alternative model hypothesizes that the model parameter is present with some prior distribution and has some value different from zero that needs to be estimated from the data (a general hypothesis: $p(\Theta_1 \neq 0 \mid y)$).

In previous work, Tendeiro and Kiers (2019) have pointed out several issues with Bayes factors, and conclude that displaying posterior distributions may be the better way to go for Bayesian analyses. van Ravenzwaaij and Wagenmakers (2021) have responded to these issues and argue that "only Bayes factors can address the key question common to most empirical research in psychology: 'to what extent do the data support the hypothesis that there is an effect'" (van Ravenzwaaij & Wagenmakers, 2021, p. 38). Tendeiro and Kiers (2021) have responded to van Ravenzwaaij and Wagenmakers (2021) and discuss their points critically with the goal to contribute "toward a better understanding among psychologists of null hypothesis Bayesian testing" (Tendeiro & Kiers, 2021, p. 2). In the present work, we do not take any strong position either way; our goal is to show that, in cognitive science applications, several important issues need to be kept in mind if one decides to use Bayes factors.

Summary of the Proposed Bayes Factor Workflow

In the section entitled Misbehaving Bayes Factors, we discuss various potential problems associated with Bayes factor analyses.

[1] The fact that we investigate Bayes factors for point null hypotheses doesn't mean we are advocating for point null hypotheses. However, we use it here since point null hypotheses are widely spread and the norm in cognitive science.

Below, we outline a Bayes factor workflow to investigate these potential problems when carrying out an analysis. These problems can largely be investigated using one set of artificial data simulations in the context of simulation-based calibration (SBC; Betancourt, 2020b; Schad, Betancourt, et al., 2021; Talts et al., 2018). Consequently, we can integrate a set of analyses (that we illustrate below) into a coherent workflow for determining when Bayes factors are robust in any given application. For this workflow, we define the following steps:

1. Define the observational model.
2. Define the prior (prior model probabilities and prior parameter distributions), ideally verified with prior predictive checks.
3. Fit the model and estimate Bayes factors using bridge sampling on the same empirical data set multiple times to investigate whether the number of MCMC draws is sufficient to obtain stable Bayes factor estimates.
4. Carry out SBC to check whether Bayes factors are computed accurately.
5. Use simulations to investigate data variability of Bayesian inferences to support realistic expectations concerning their reliability.
6. If the simulations support accurate and reliable Bayes factor estimation, then one can use the Bayes factors obtained for the empirical data to support Bayesian inferences (possibly estimate Bayes factor repeatedly on empirical data and report median plus standard deviation); otherwise, improve experimental design or acknowledge limitations.

In many cases, having valid Bayes factor estimates will be sufficient because an important goal in cognitive science is to provide evidence in support of scientific hypotheses. This evidence is continuous in nature and thus reporting continuous evidence in scientific papers, without making discrete decisions, would be a natural approach. This is especially important given the large data variability inherent in evidence quantification (see below for illustrations), which often makes discrete decisions seem premature based on individual data sets. However, if discrete decisions are needed, for example, in order to make a discrete discovery claim, then the workflow can be expanded with the following steps:

1. If one wants to make a decision, e.g., on a discovery claim, then one can define utility functions based on principled grounds.
2. Use simulations to optimize the discovery threshold.
3. Use simulations to investigate data variability of decisions (false and true discovery rates).
4. Make a decision on discovery using an optimized discovery threshold.

We provide an example application of the workflow to a rather standard cognitive science data set; reproducible code is available on OSF (Schad, Nicenboim, et al., 2021).

We consider this workflow, in particular conducting SBC, to be the ideal way to approach Bayes factor analyses. However, we acknowledge that it takes a lot of time and computational resources to run this workflow for realistic research problems. It may therefore be difficult in research practice to implement this ideal workflow in every single analysis that one runs. We therefore suggest implementing this workflow once for a given research program, where different models and experimental designs may be similar to each other.

Based on this definition of the Bayes factor workflow, we now discuss in detail the problems and questions that motivate the workflow.

Misbehaving Bayes Factors

Issue 1: Bayes Factors in Complex Statistical Models Can Be Unstable

One question is how do we apply the Bayes factor method to models that we encounter frequently in cognitive science, such as for example, hierarchical models with many variance components. For these models (or any other complex model) it's not possible to calculate the marginal likelihood analytically. Bridge sampling (Bennett, 1976; Meng & Wong, 1996) has been proposed as a method for calculating the Bayes factor for complex models (for simpler models see Morey & Rouder, 2018).

The bridge sampling approach involves approximations of the marginal likelihoods. Bayes factor estimates based on bridge sampling can be unstable when based on models with too low effective sample size. Effective sample size is corrected for the autocorrelation of Markov chains and provides an estimate of how much information is in a chain relative to the number of independent samples (Vehtari et al., 2020). Estimates of effective sample size are quantity specific (Betancourt, 2020a) and may differ between an estimate for the posterior mean versus the marginal likelihood. Thus, even high effective sample size for the posterior density may not yield stable bridge sampling estimators, where effective sample size may still be low. Indeed, bridge sampling requires many more (effective) posterior samples than what is normally required for parameter estimation.

Running `brms` models with the default number of MCMC draws will induce instability in the Bayes factor estimates based on the bridge sampling, such that running the same analysis twice would yield different results for the Bayes factor. Moreover, bridge sampling in itself may be unstable and may return different results for different runs on the same posterior MCMC draws (just because of different starting values). This is very concerning, as the results reported in an article might not be stable if the number of posterior samples or effective sample size is not large enough. The `brms` defaults (2,000 iterations for each chain, with half of them used as "warm-up") were not set to support bridge sampling. Instead, they are intended for posterior inference on expectations (e.g., posterior means) for models that are not too complex.

To investigate the potential instabilities in Bayes factor estimation, we use the following experimental design:

We assume a design with a within-subject factor x with two levels and 15 subjects. Each condition (sum coded –1/+1, Schad et al., 2020) is measured twice per subject. We assume that our dependent variable is response times in milliseconds, and we assume that response times are log-normally distributed.

To explain response times in this experimental design, we aim to test two distinct hypotheses, which are implemented in two

different hierarchical (linear mixed-effects) models. The alternative hypothesis (H1) assumes that factor x influences the dependent variable, i.e., that the fixed effects estimate β_1 associated with factor x takes some value that has a prior distribution and is different from zero H_1: $\beta_1 \neq 0$. In R, the corresponding model formula can be written as: `log(rt) ~ 1 + x + (1 + x | subj)`. The null hypothesis (H0) assumes that factor x does not influence the dependent variable response times, that is, H_0: $\beta_1 = 0$. In R, the corresponding model formula can be written as: `log(rt) ~ 1 + (1 + x | subj)`. We will use Bayes factors to compare the general hypothesis H1 to the point hypothesis H0.

For this simulation, we set known true values for the parameters: Because the response times are assumed to be log-normally distributed, the parameters are defined in this log-normal distribution model. They can be interpreted as the parameters for a linear mixed-effects model on log-transformed response times. We use values for the fixed effects for the intercept of 6 (i.e., median response times of `exp(6) = 403 ms`) and for the effect of x of –1. (This effect of –1 constitutes a very strong experimental effect of nearly 1,000 ms.) For the random effects we assume standard deviations of .5 and a correlation of .3. The residual noise is set to .5. We simulate the data using the R function `simLMM()` (in the `designr` package; Rabe et al., 2021) and make use of the functionality `empirical=TRUE`, which makes sure that the fixed effects in the data correspond precisely to the indicated values (i.e., the intercept is exactly 6 and the effect of x exactly –1).

For this fixed simulated data set, we estimate the same model 100 times, that is, each time performing new MCMC sampling using the same data, model and priors. We compute Bayes factors for each of the 100 models, and then investigate whether the 100 Bayes factors are the same in each run. We run this analysis using bridge sampling using the default number of 2, 000 samples and also using the larger number of 10, 000 samples.

The results, displayed in Figure 1, show that Bayes factor estimates become quite unstable when using a smaller number of MCMC samples.

These results demonstrate that bridge sampling with a large number of samples is required to obtain stable Bayes factor estimates. Of course, if Bayes factors are not estimated in a stable way, but depend on random noise in the MCMC chain, then these Bayes factor estimates cannot accurately represent information that is contained in the data.

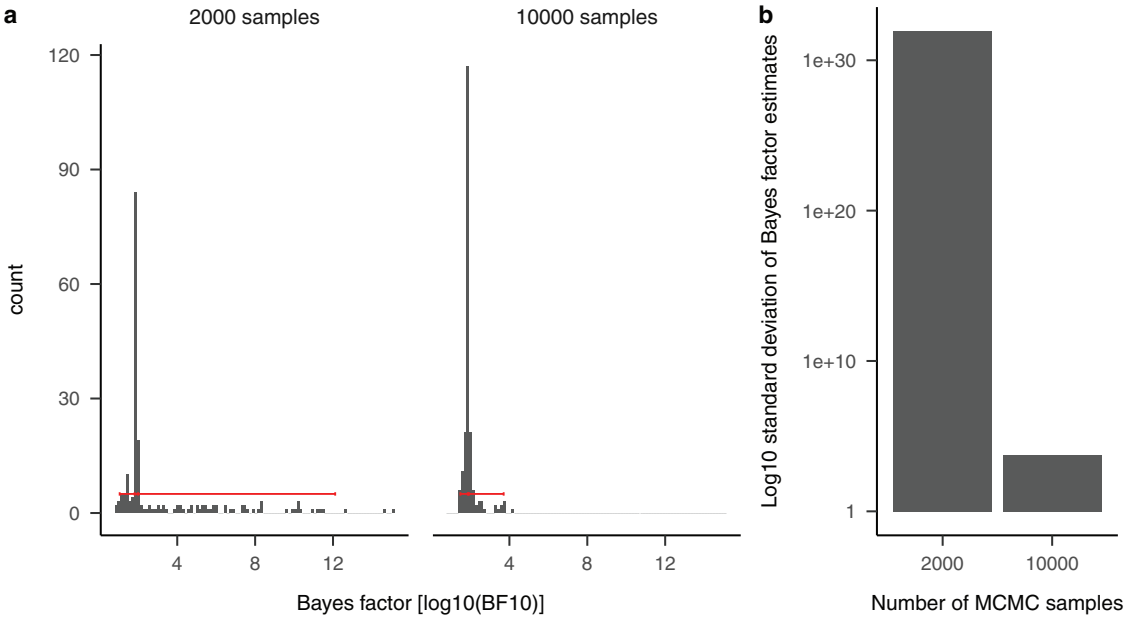
The stability of bridge sampling needed in any given application depends on how small a difference between marginal likelihood one wants to be able to resolve. If the two models compared are very different, then even large variation might be acceptable. In this case, it is important to report that variation in Bayes factors. However, if the two models are very close to each other, then small variation might be problematic.

Next, we studied the stability of Bayes factors when in the true data simulation process, H0 was actually true. We used the same simulated data set as in the previous analysis, with the only difference that we set the critical fixed effect estimate to zero.

The results show (see Figure 2) that larger number of samples yielded more stable Bayes factors. In general the instability of Bayes factor estimates against the MCMC draws (and starting values of the bridge sampler) demonstrates that it is necessary to use a large number of iterations when computing Bayes factors using

Figure 1

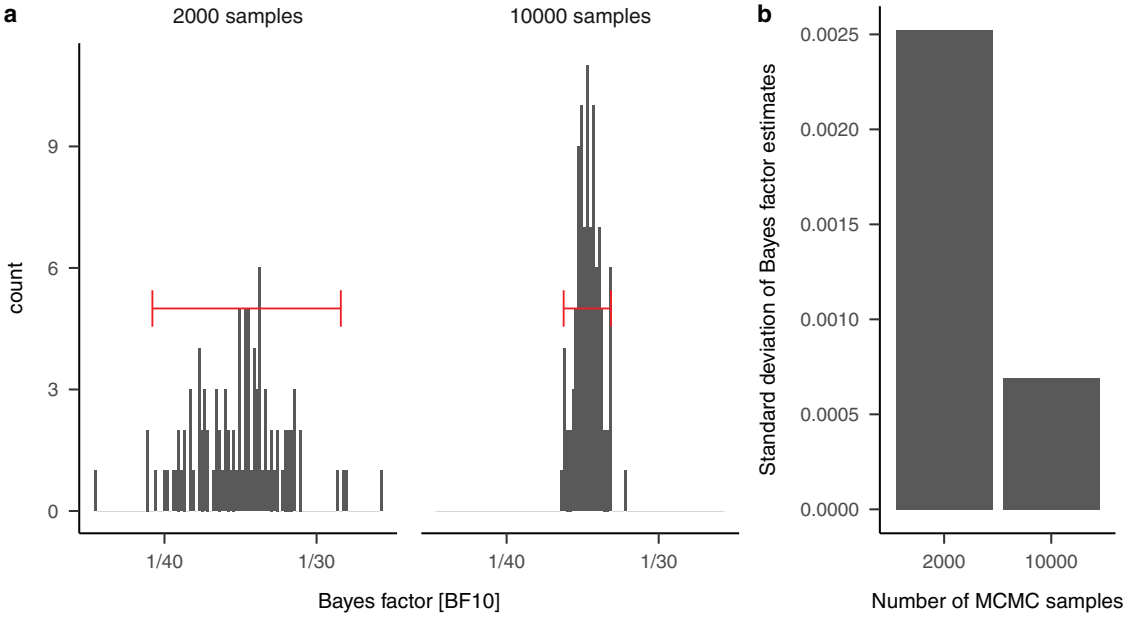
Stability of Bayes Factor Estimates Against Different MCMC Chains



Note. (a) Histograms of 100 Bayes factor estimates obtained from the same data and model, where only the MCMC chains differ between runs. The histograms are shown for bridge sampling for the default number of 2,000 samples (left panel) and for a larger number of 10,000 samples (right panel). Red horizontal error bars indicate 95 percent quantiles. (b) Bars show the log10 standard deviation across 100 Bayes factor estimates displayed in (a) for each number of samples separately. See the online article for the color version of this figure.

Figure 2

Stability of Bayes Factor Estimates Against Different MCMC Chains in a Situation Where H0 is the True Model



Note. (a) Histograms of 100 Bayes factor estimates obtained from one same data and model, where only the MCMC chains differ between runs. The histograms are shown for bridge sampling for the default number of 2,000 samples (left panel) and for a larger number of 10,000 samples (right panel). Red horizontal error bars indicate 95 percent quantiles. (b) Bars show the standard deviation across 100 Bayes factor estimates displayed in (a) for each number of samples separately. See the online article for the color version of this figure.

`brms` and `bridge_sampler()`. Moreover, it shows that we have to check for each data set that we analyze that our Bayes factor estimate is stable. It is possible to do this by running the analysis a few times (at least five to six times) to test whether the obtained Bayes factor estimates are stable.

That said, it's important to note that stability doesn't mean accuracy. Bridge sampling with a large number of samples returns estimates with low variability. However, it is not clear from the stability analysis how those estimates relate to the true Bayes factors! We turn to this issue in the next section.

Issue 2: Bayes Factor Estimates Can Be Biased

Importantly, for approximate computations of Bayes factor estimates using bridge sampling, there are no strong guarantees for their accuracy. In principle, it could very well be that a stably estimated Bayes factor based on bridge sampling is in fact biased. The technique of simulation-based calibration (SBC; Betancourt, 2020b; de Heide & Grünwald, 2021; Hendriksen et al., 2021; Rouder, 2014; Schad, Betancourt, et al., 2021; Talts et al., 2018; Tendeiro et al., 2021) can be used to investigate this question.

Even when Bayes factor estimates based on bridge sampling are computed in a stable way (i.e., stability over different sets of MCMC draws), it is unclear whether the estimates are unbiased for the kinds of (multilevel) models that are standardly used in psychology and psycholinguistics. To understand these problems, it is useful to discuss the typical set, which is the "set containing the bulk of the posterior probability mass" (Gabry et al., 2019, pp. 394–395). MCMC explores the typical set and uses that exploration to estimate expectation values of functions of the parameters. When the algorithm enjoys a *central limit theorem*, that exploration is effective and the error in an estimator is determined by how much the variation of the corresponding function is contained within the typical set. Bayes factors, however, are given by the posterior expectation of the reciprocal likelihood function which usually varies most at extreme values far away from the typical set. Even under ideal conditions the MCMC estimators for these expectations can suffer from large errors. Therefore, calibrations (Betancourt, 2019) are needed to test whether Bayes factor estimates correspond to the true Bayes factor in a given application. We will discuss this issue and perform such calibrations below.

The estimation error of Bayes factors can vary with the observed data. Therefore, we need to quantify that data variation if we want to use this method responsibly. We can implement this calibration by observing how the Bayes factor outcomes vary across prior predictive simulations (Betancourt, 2019). We do this using SBC.

We can formulate SBC for model inference, where $\mathcal{M}$ is a true model used to simulate artificial data y, and $\mathcal{M}'$ is a model inferred from the simulated data.

$$p(\mathcal{M}') = \int \mathrm{d}y \int \mathrm{d}\mathcal{M} p(\mathcal{M}' \mid y) p(y \mid \mathcal{M}) p(\mathcal{M}) \tag{5}$$

We can read this equation sequentially: first, we sample a model from the model prior, $p(\mathcal{M})$. Next, we simulate data based on this model, $p(y \mid \mathcal{M})$. This in fact involves two steps: simulating parameters from the parameter prior, $p(\Theta \mid \mathcal{M})$, and then simulating artificial data from the parameters and model, $p(y \mid \Theta, \mathcal{M})$, i.e.,

$p(y \mid \mathcal{M}) = \int p(y \mid \Theta, \mathcal{M}) \times p(\Theta \mid \mathcal{M}) \mathrm{d}\Theta$. Next, we estimate the posterior model probabilities from the simulated data, $p(\mathcal{M}' \mid y)$. This again involves several steps: we estimate the model parameters based on the simulated data, $p(\Theta' \mid y)$, for each of the two models, use this to compute marginal likelihoods, $p(y \mid \mathcal{M}')$, and Bayes factors for each of the two models using bridge sampling, and then compute the posterior probability for each model given the data, $p(\mathcal{M}' \mid y)$, by adding the model prior. If we now integrate out the simulated data y and the model $\mathcal{M}$, then we can see that this procedure results in $p(\mathcal{M}')$. That is, the average posterior should be exactly the same as the prior $p(\mathcal{M})$. The key idea is that if the computation of Bayes factors and posterior model probabilities is performed correctly (and of course the data simulation is implemented correctly), then the average posterior probability for a model should be the same as its prior probability. Interestingly, this can be used to test the accuracy of Bayes factor estimation.

Critically if SBC does not show a difference between the average posterior (i.e., the left-hand side of Equation 5) and the prior, then this doesn't guarantee that the computation for every posterior will necessarily be good; it is a necessary condition but not a sufficient one.

A note of caution here: This consistency condition, that the average posterior is equal to the prior, holds only for the average posterior over prior predictive simulations; we have no guarantees on how any individual posterior distribution will behave in these simulated data, let alone for observed data. Thus, the statement that the average posterior is equal to the prior does not apply to Bayesian inference on a single data set, where a prior is used to infer a posterior distribution, but is specific to SBC.

Applied to our current example of null-hypothesis testing, we define a prior on the model space, for example, we can define the prior probabilities for a null and an alternative model, specifying how likely each model is a priori. From these priors, we can randomly draw one hypothesis (model), for example, $n_{sim} = 500$ times. Thus, in each of 500 draws we randomly choose one model (either H0 or H1), with the probabilities given by the model priors. For each draw, we first sample model parameters from their prior distributions, and then use these sampled model parameters to simulate artificial data. For each simulated artificial data set, we can then compute marginal likelihoods and Bayes factors (between the models H1 and H0) using bridge sampling, and we can then compute the posterior probabilities for each hypothesis using the true prior model probabilities (i.e., how likely each model is a posteriori). As the last, and critical step in SBC, we can then compare the posterior model probabilities (averaged across all 500 simulations) to the prior model probabilities. A key result in SBC is that if the computation of marginal likelihoods and model posteriors is performed accurately by the bridge sampling procedure and does not exhibit bias, that is, if the approximate Bayes factor estimate corresponds to the true Bayes factor, then the data-averaged posterior model probabilities should be the same as the prior model probabilities.

An Example With an Unbiased Estimate of the Bayes Factor. We investigate whether the approximate Bayes factor estimates (based on bridge sampling) are unbiased with the same experimental design presented before.

In order to perform SBC, we need to define the prior model probabilities. For simplicity, we assume that both hypotheses (H0 and H1) are both equally likely a priori, which also has the advantage that both hypotheses are equally frequently sampled in the SBC. (However, see Schad & Vasishth, 2019; for a different prior with higher probability for the null.)

Next, we define priors for the model parameters. For the intercept we assume a normal distribution with mean 6 and standard deviation .5. For the fixed effect estimate for factor x, we assume a normal distribution with mean 0 and standard deviation 1. For the random effects standard deviations, we assume a half normal distribution with mean 0 and standard deviation 1.5, which is truncated to take only positive values. For the residual noise term, we assume a normal distribution with mean 0 and standard deviation .5, which is again truncated to take only positive values. For the random effects correlation between the intercept and the estimate for x, we assume an LKJ prior (Lewandowski et al., 2009) with parameter value 2. This LKJ(2) prior regularizes the prior distribution on the correlation such that extreme values like –1 or +1 are heavily downweighted.

Based on these priors, it is now possible to simulate a priori data for the artificial experimental design. First, we use the prior probabilities for the hypotheses to sample a hypothesis from the prior. We do so 500 times (i.e., 500 runs of SBC).

We see that H0 and H1 are each sampled approximately 250 times (see Table 1). We will perform a formal SBC analysis below. Next, we sample model parameters from the priors based on the model that was sampled in each run, and we can then simulate data based on the sampled hypothesis and the sampled parameters. The next step is to estimate the Bayesian models on the simulated data. For each simulated data set, we estimate the posterior parameter distributions for each model (H0 and H1) using `brms`, then we perform bridge sampling, and then we use this to compute a Bayes factor for each of the 500 simulated data sets.[2]

For each model (H0 and H1), we compute marginal likelihoods, and we compute the Bayes factor by comparing marginal likelihoods. Note that in the null model, we do keep the random effects of factor x varying across subjects, that is, `fakert ~ 1 + (1+x | subj)`. That is, we do assume that effects of factor x could be present for individual subjects, but importantly, by removing the fixed effect of x we assume a priori that the overall mean effect across all subjects is zero. The model comparison therefore targets only this fixed effect of factor x, but not the random effects.

While this is not required in and part of SBC, we show here the distributions of Bayes factors given the true hypotheses (see Figure 3). The results show that the Bayes factor estimates exhibit wide distributions when either H0 or H1 are true. It is clear that when H1 is the true hypothesis in the data simulation, then the Bayes factors provide more evidence for H1 on average. By contrast, when H0 is the true hypothesis in the data simulation, then the distribution of Bayes factors is clearly shifted toward evidence for H0. Interestingly, these distributions are quite asymmetric such that Bayes factors accrue evidence in favor of true H1 at a faster rate than they do in favor of H0 (Tendeiro & Kiers, 2019).

[2] We specify a large number of sampling iterations for each of four chains (iterations: 10,000, warmup samples: 2,000) to obtain stable Bayes factor estimates as we discussed in the previous section. The control parameter `adapt_delta` is set to 0.9, and `max_treedepth` is set to 15. These ensure that the posterior sampler works correctly (Betancourt, 2016, 2017; Gabry et al., 2019).

Table 1
Number of Sampled Hypotheses for the H0 and for the H1

Quantity	H0	H1
n samples	245	255

Note that there was one outlier data point for H0 with a $BF_{10} = -3 \times 10^{-86}$. This resulted from an unstable marginal likelihood because the bridge sampling did not converge. Thus, even in the simple example case we use where models were fit with a large number of iterations (10,000), there can occasionally be problems with bridge sampling.

Last, we can compute the posterior probability of the hypothesis H0 being true given the data. The posterior odds, $p(H1 \mid y)/p(H0 \mid y)$, can be obtained by multiplying the Bayes factor with the prior odds (which is $p(H1)/p(H0) = .5/.5 = 1$ in our example):

$$p(H1 \mid y)/p(H0 \mid y) = BF_{10} \times p(H1)/p(H0) \tag{6}$$

If the probability $p(H0 \mid y)$ is p, then the above equation can be written as follows:

$$(1 - p)/p = BF_{10} \times p(H1)/p(H0) \tag{7}$$

We solve for p and obtain:

$$\begin{aligned} p(H0 \mid y) = p &= \frac{1}{BF_{10} \times p(H1)/p(H0) + 1} \\ &= \frac{1}{p(H1 \mid y)/p(H0 \mid y) \times p(H1)/p(H0) + 1} \end{aligned} \tag{8}$$

Because $p(H1 \mid y) = 1 - p = q$, we can use the fact that $p = \frac{1}{BF_{10} \times p(H1)/p(H0) + 1}$ to work out $1 - p = q$, which gives us:

$$\begin{aligned} p(H1 \mid y) = q &= \frac{BF_{10} \times p(H1)/p(H0)}{BF_{10} \times p(H1)/p(H0) + 1} \\ &= \frac{p(H1 \mid y)/p(H0 \mid y) \times p(H1)/p(H0)}{p(H1 \mid y)/p(H0 \mid y) \times p(H1)/p(H0) + 1} \end{aligned} \tag{9}$$

As the last step, across the 500 simulation runs, we average the posterior probabilities for each model, i.e., by computing the mean posterior probability across all 500 runs: $\mu_{\mathcal{M}_x}^{post} = \frac{1}{n_{sim}} \Sigma_{i=1}^{n_{sim}} p(\mathcal{M}_x \mid y_i^{sim})$, where each y_i^{sim} is one out of $n_{sim} = 500$ simulated data sets, $\mathcal{M}_x$ is one selected model, and $\mu_{\mathcal{M}_x}^{post}$ is the average posterior probability for model x.

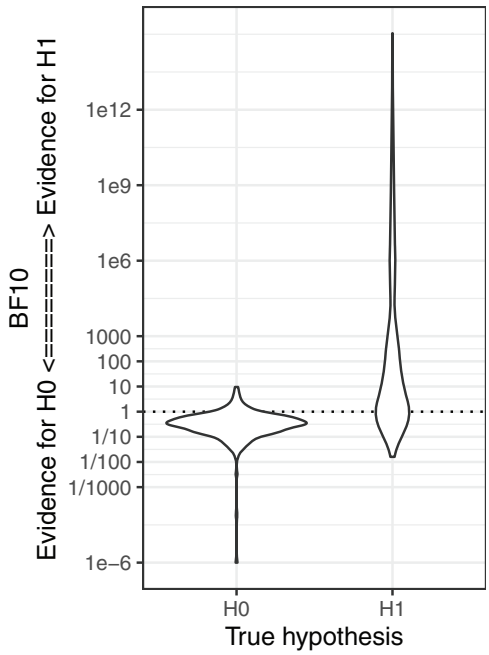
Now, we can investigate our question of interest in SBC: We can look at how likely each model was chosen a posteriori on average and compare these average posterior model probabilities (see Table 2, "Mean"; in addition, their 95% binomial confidence intervals) to the prior model probabilities that were in fact used to simulate the data (i.e., 50% each).

The results (see Table 2) show that the average posterior model probability for H1 versus H0 was at roughly 50%. This result directly corresponds to the prior model probability of 50%. The confidence intervals include the prior of 50%. This SBC analysis therefore, for this specific and simple example case, did not indicate any signs of significant bias. This is important calibration information for the bridge sampling approach because it has not been clear so far whether bridge sampling yields unbiased estimates for the types of multilevel models often used in research practice. These results are therefore encouraging and support the application of bridge sampling for computation of Bayes factors and posterior probabilities for our case study. However, much more extensive simulation studies are required to investigate this point more generally, which is beyond the scope of this article.

Importantly, the SBC analysis supported the Bayes factor estimates and *could not* detect a difference between average posterior model probabilities and the prior model probabilities, suggesting that the Bayes factor estimates for this analysis are valid. However, this will not always be the outcome of SBC. Next, we will investigate a case where SBC shows a problem with posterior model probabilities.

An Example Where Bayes Factor Estimates Are Not Accurate When the Models Are Mis-Specified. We again

Figure 3
Distribution of Bayes Factors (BF10), Shown as Violin Plots, as a Function of Which Hypothesis was True in the Simulations From the SBC



Note. One outlier data point for H0 with a BF10 of -3e-86 resulted from an unstable marginal likelihood (i.e., the bridge sampling did not converge) and was removed for visualization.

Table 2
Average Posterior Probability for the H1, p(H1|y), Together With 95 Percent Frequentist Binomial Confidence Intervals

Quantity	2.5% quantile	Mean	97.5% quantile
p(H1|y)	45.53	49.91	54.3

investigate the same experimental design as above, with the same priors for the parameters and the same observational model as in the previous section. The only thing that differs in this analysis is that we leave out the random slopes in the estimation procedure. In R syntax, H1 can thus be written as `log(rt) ~ 1 + x + (1 | subj)` and H0 can be written as `log(rt) ~ 1 + (1 | subj)`. It is known from analysis of frequentist tools that leaving out random slopes from a linear mixed-effects model can lead to an increase in α, the probability of a type I error (Barr et al., 2013; Matuschek et al., 2017). Here, we are interested in whether leaving out random slopes from a corresponding Bayesian multilevel model leads to a bias in posterior model probabilities. In line with the frequentist results, we expect that posterior model probabilities for H1 should be inflated when neglecting random slopes. To investigate this, we perform the exact same SBC analysis as before, but only leaving out the random slopes from the fitted models.

The result (see Table 3) shows that now, as expected, the average posterior probability for the alternative hypothesis (H1) of 58.57% is higher than its prior probability of 50%. This increase is supported by the 95% confidence intervals, which do not overlap with 50%. Moreover, a frequentist intercept-only logistic regression shows that the average posterior model probability is highly significantly different from the prior probability of 50% ($b = .35$, $SE = .09$, $z = 3.81$, $p = .0001$). This shows that leaving out random slopes from a multilevel model when the data do contain random variation of the effect across subjects can lead to severe biases in the posterior model probabilities, as pointed out by Barr et al. (2013) for frequentist linear mixed-effects models. SBC is a useful tool that can detect such biases. It is therefore highly recommended to use SBC to calibrate one's Bayes factor estimates for a specific study, model, and priors.

As a brief comment, one can also simulate data based on the assumption that there are no random slopes, that is, that the random variance of slopes is exactly zero. In R syntax, this would be expressed as: `log(rt) ~ 1 + x + (1 | subj)`. We can then fit Bayesian models which estimate the variance of random slopes and compute a Bayes factor based on this, that is, in R-syntax comparing the H1, `log(rt) ~ 1 + x + (1 + x | subj)`, to the H0, `log(rt) ~ 1 + (1 + x subj)`. This implies investigating evidence for the fixed effect x, while estimating random slopes of x across subjects although this random variance is in fact zero. Interestingly, in this case the average posterior does not diverge from the prior 50% (see Table 4).

This shows that it is a good idea to estimate random slopes in Bayesian linear mixed-effects models, even if it is unclear whether the true random slope variance is zero. However, note that the average posterior is slightly below 50% (i.e., 48%), thus, a potential smaller bias (i.e., slightly too conservative estimates with a higher probability for the H0) might be observed when larger numbers of simulations are used.

Table 3

Average Posterior Probability for the H1, p(H1|y), Together With 95 Percent Frequentist Binomial Confidence Intervals, for a Mis-Specified Model Lacking Random Slopes

Quantity	2.5% quantile	Mean	97.5% quantile
p(H1|y)	54.21	58.57	62.84

Table 4

Average Posterior Probability for the H1, p(H1|y), Together With 95 Percent Frequentist Binomial Confidence Intervals, for a Mis-Specified Model, With Random Slopes in the Model But Not in the Data

Quantity	2.5% quantile	Mean	97.5% quantile
p(H1|y)	43.79	48.16	52.54

Issue 3: Bayes Factor Estimates Can Vary Dramatically Due to Sampling Variability

A very different source limiting the robustness of Bayes factor estimates derives from the variability that is observed in the data, that is, among subjects, items, and residual noise. Every time a new replication run is conducted, using different subjects and items, it will lead to different outcomes of the statistical analysis. This limit to robustness is well known in frequentist analyses, as the "dance of *p*-values" (Cumming, 2014), where over repeated replications, *p*-values are not consistently significant across studies. Instead, the results yield highly different *p*-values each time a study is repeated, and this can even be observed when simulating data from some known truth and rerunning analyses on simulated data sets. This same type of variability should also be present in Bayesian analyses (also see https://daniellakens.blogspot.com/2016/07/dance-of-bayes-factors.html). Here we investigate this type of variability in Bayes factor analyses.

One way to investigate how variable the outcome of Bayes factor analyses can be (given that the Bayes factor is computed in a stable and accurate way) is to investigate the variability of Bayes factors across the prior predictive simulations performed in the SBC above, as is illustrated in Figure 3, and which shows that estimated Bayes factors can be highly variable and often provide evidence for the false hypothesis.

One potential critique of the result shown in Figure 3 may be that the prior assumptions underlying our specific SBC simulations might be rather artificial, because we used a simple artificial example that may not closely reflect the real situation in many studies in the cognitive sciences. (The example was designed for pedagogical purposes and for fast computations, not to closely match real cognitive data sets. In a real SBC analysis, the priors and the experimental design should be closely informed by domain expertise.) Here we are therefore interested in how much variability in Bayes factors we should expect in a typical data set from the cognitive sciences. One question therefore is how to choose the priors to yield realistic simulated data sets that are comparable to what one would expect in real experimental observations, and that would thus provide realistic estimates of the variability of Bayes factors in the cognitive sciences. To approximate such realistic data sets, one possibility is to define the priors based on a previous real empirical data set, by estimating the model posterior on this prior data, and using the model posterior as our prior for simulating data and for analyzing variability of Bayes factors. We expect that this procedure (effectively implementing posterior predictive analyses) should produce realistic estimates of the variability in Bayes factors that would be expected if we would run this identical study again in

several replication attempts using the exact same experimental design.

Examples: Facilitatory Interference Effects. To investigate this, we will look at some fairly typical experimental example studies from the cognitive sciences. We look at studies that investigated cognitive mechanisms underlying a well-studied phenomenon in sentence comprehension. The example we consider here is the agreement attraction configuration below, where the ungrammatical sentence (2) seems more grammatical than the equally ungrammatical sentence (1):

1. The key to the cabinet are in the kitchen.

2. The key to the cabinets are in the kitchen.

Both sentences are ungrammatical because the subject does not agree with the verb in number: The verb ("are") does not agree in number with the subject of the sentence ("key"). Sentences such as (2) are often found to have shorter reading times at the verb ("are") compared with (1) (for a meta-analysis see Jäger et al., 2017). Such shorter reading times are sometimes referred to as "facilitatory interference" (Dillon, 2011). One proposal explaining the shorter reading times is that the attractor word (here, cabinets) agrees locally in number with the verb, leading to an illusion of grammaticality. This is an interesting phenomenon as the plural versus singular feature of the attractor noun ("cabinet/s") is not the subject, and therefore does not need to agree with the verb. That agreement attraction effects are consistently observed indicates that the human language processing system is not strictly obeying the rules of grammar when assembling the parse. An account of agreement attraction effects in language processing that is based on a full computational implementation (which is in the ACT-R framework; Taatgen et al., 2006), explains such agreement attraction effects in ungrammatical sentences as a result of retrieval-based working memory mechanisms (Engelmann et al., 2019; cf. Hammerly et al., 2019). Agreement attraction in ungrammatical sentences has been investigated many times in similar experimental setups with different dependent measures such as self-paced reading and eye-tracking. We choose this phenomenon for analysis here because it provides an example of a relatively robust effect in cognitive science, and because multiple data-sets on agreement attraction are publicly available.

In this section, we look at the data variability of Bayes factors (and associated effect estimates) using posterior predictive simulations across several different scenarios. First, we investigate a study by Lago et al. (2015) using priors (in the model fitting, not in the simulation of data) derived from a meta-analysis, where the prior mean differs from zero, and where the data provide some evidence for an effect (Case 1). Then, we look at this same data set using a more neutral prior that is centered on zero (Case 2). Next, we use data from a study (Wagers et al., 2009) where the overall effect of interest is close to zero (Case 3). These two data sets are of rather small size (30–60 subjects), which is often the case and rather typical in the cognitive sciences. Here, we also investigate the variability of Bayes factors (and associated effect estimates) in a large sample study (Jäger et al., 2020), which used 181 subjects (Case 4), and yields much more stable Bayes factor estimates. Next, we go one step further by looking not at simulated replications of a study, but at 11 real empirical replication studies of the same experimental effect. As in the simulated data shown earlier, the real empirical data results also show strong variability of the Bayes factor across studies, little evidence within each single study, but strong evidence when pooling across individual studies.

Case 1: Lago et al. (2015)

First, we investigate facilitatory agreement attraction effects by looking at a self-paced reading study by Lago et al. (2015). We estimate a fixed effect for the experimental condition agreement attraction (x; i.e., sentence type), against a null model where the fixed effect of sentence type is excluded. Note that for the agreement attraction effect of sentence type, we use sum contrast coding (i.e., –1 and +1). We run a multilevel model with the following formula in `brms`: `rt ~ 1+x + (1+x | subj) + (1+x | item)` , where rt is reading time, we have random variation associated with subjects and with items, and we assume that reading times follow a log-normal distribution: family = lognormal().

In the analysis of the real empirical data, we use results from a meta-analysis (Jäger et al., 2017) to obtain priors for the effect size of the factor x agreement attraction. We describe how we obtained the priors in detail in a long version of this article, which is available on arXiv (arXiv:2103.08744). For extensive discussion of prior specification in such designs, see Nicenboim et al. (2021).

Next, we take a look at the population-level results from the Bayesian modeling (see Table 5). They show that for the fixed effect x, capturing the agreement attraction effect, the 95% credible interval ranges between –.046 and –.015 on the log ms scale. This suggests that the effect may have the expected negative direction, reflecting shorter reading times in the plural condition. Importantly, such estimation is not really answering the question how much evidence there is that the parameter is needed to explain the data (see Rouder et al., 2018; Wagenmakers et al., 2020) because we did not specify the null hypothesis of zero effect explicitly. Instead, Bayes factors are needed to investigate this issue.

To estimate Bayes factors, run the model again, now without the parameter of interest, that is, the null model, which essentially fixes β to exactly zero: `rt ~ 1 + (1+x | subj) + (1+x | item)`. Then, compute the log marginal likelihoods and Bayes factors using bridge sampling (Gronau et al., 2017, 2020). This procedure yields the Bayes factor BF_{10}, that is, the evidence of the alternative over the null.

The procedure described above yields a Bayes factor of 6.74, suggesting that there is some support for the alternative model. That is, the Bayes factor on the real empirical data provides evidence for the alternative hypothesis that there is a difference between the experimental conditions, that is, a facilitatory effect in the plural condition of the size derived from the meta-analysis. Under the criteria suggested by Jeffreys (1939), the Bayes factor provides only moderate evidence for an effect of sentence type on reading times.

Table 5
Population-Level Fixed Effect Estimates for the Bayesian Linear Mixed-Effects Model

Fixed effect	Estimate	Est. Error	Q2.5	Q97.5
Intercept	6.015	0.056	5.903	6.127
x	−0.031	0.008	−0.046	−0.015

Next, we use the posterior of this model for our prior predictive simulations to investigate the data variability of Bayes factors. Figure 4 visualizes the simulated data via density plots for the observed data (black) and for 100 posterior simulated data sets (shown in color/grey). It shows that the simulated data seem fairly in line with the empirically observed data, at least for this simple density plot.

The question that we are interested in is, how variable will Bayes factors be in these simulated data (in which we use the model posterior to inform our prior). We can compute Bayes factors on each of these simulated data to investigate whether there is evidence for agreement attraction effects. Of great interest to us is then the question of how variable the results of these Bayes factor analyses will be across different simulated replications of the same study. We now perform this analysis for 50 different artificial data sets.

We can now visualize the distribution of Bayes factors (BF_{10}) across the artificial data sets by plotting a histogram. Values larger than one in this histogram indicate evidence for the alternative model (H1) that agreement attraction effects exist (i.e., the sentence type effect is different from zero and as specified by the meta-analysis), and Bayes factor values smaller than one indicate evidence for the null model (H0) that no agreement attraction effect exists (i.e., the difference in reading times between experimental conditions is zero).

The results (see Figure 5) show that the Bayes factors are quite variable. Although all data sets are simulated from the same posterior predictive distribution, the Bayes factor results are as different as providing moderate evidence for the null model ($BF_{10} < 1/3$) or providing strong evidence for the alternative model ($BF_{10} > 10$). The bulk of the simulated data sets provide moderate or anecdotal evidence for the alternative model. That is, much like the "dance of *p*-values" (Cumming, 2014), this analysis reveals a "dance of the Bayes factors" (https://daniellakens.blogspot.com/2016/07/dance-of-bayes-factors.html) with simulated repetitions of the same study. The variability in these results shows that a typical cognitive or psycholinguistic data set is not necessarily highly informative for drawing firm conclusions about the hypotheses in question.

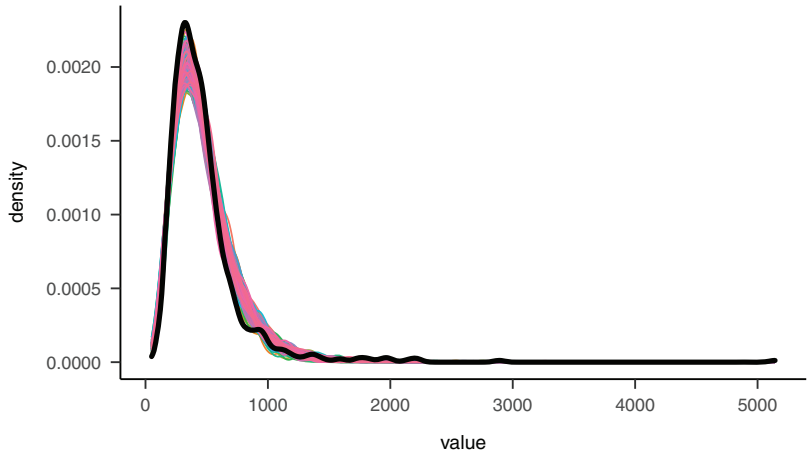
What is driving these differences in the Bayes factors between simulated data sets? One obvious reason why the outcomes may be so different is that the difference in reading times between the two sentence types, that is, the experimental effect that we wish to make inferences about, may vary based on the noise and uncertainty in the posterior predictive simulations. It is therefore interesting to plot the Bayes factors from this simulated data set as a function of the difference in simulated reading times between the two sentence types as estimated in the Bayesian model. That is, we extract the estimated mean difference in reading times at the verb between plural and singular attractor conditions from the fixed effects of the Bayesian model, and plot the Bayes factor as a function of this difference (together with 95% credible intervals).

The results (displayed in Figure 6, left panel) show that the mean difference in reading times between experimental conditions varies across posterior predictive simulations. This indicates that the experimental data and design contain a limited amount of information about the effect of interest. Of course, if the (simulated) data is not stable, Bayes factor analyses based on this simulated data cannot be stable across simulations either. Accordingly, as is clear from Figure 6, the magnitude of the difference in mean reading times between experimental conditions is indeed a main driving force for the Bayes factor calculations.

An important observation from Figure 6 (left panel) is that as the difference between reading times becomes more negative, that is, as the plural noun condition (i.e., "cabinets" in the example; sentence 2) is read faster than the singular noun condition (i.e., "cabinet"; example sentence 1), the Bayes factor BF10 increases to larger and larger values, indicating that the evidence in favor of the alternative model increases. By contrast, when the difference between reading times becomes *less* negative, that is, the plural condition (sentence 2) is not read much faster than the singular condition (sentence 1), then the Bayes factor BF10 decreases to values smaller than 1. Importantly, this behavior occurs because we are using our informative priors from the meta-analysis, where the prior mean for the agreement attraction effect is not centered at a mean of zero, but has a negative value (i.e., a prior mean of −.027 on the log millisecond scale). Therefore,

Figure 4

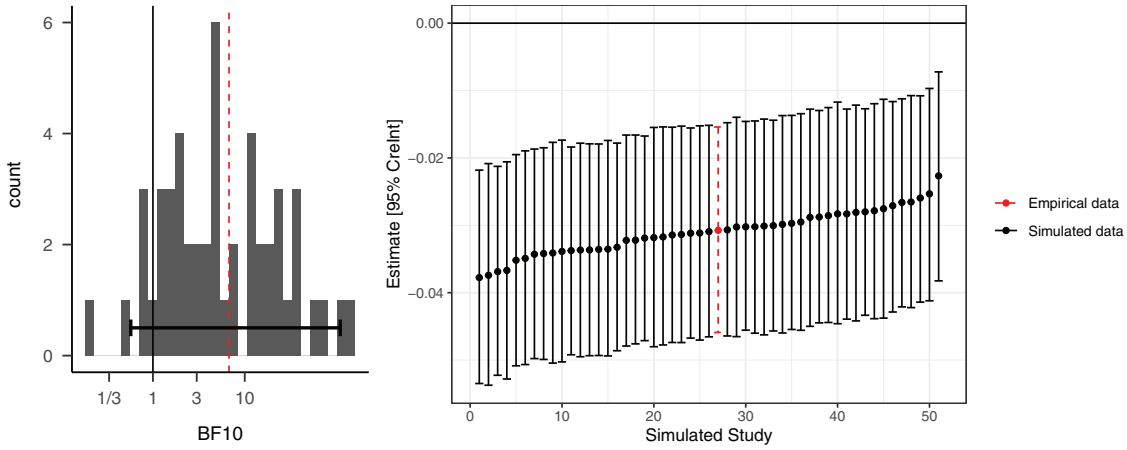
Density Plots for Observed Data (Black) and for 100 Posterior Artificial Data Sets (Shown in Color/Grey) Simulated From the Fitted Bayesian Model Displayed in the Table 5



Note. See the online article for the color version of this figure.

Figure 5

Distribution of Bayes Factors and Effect Estimates in the Simulated Data Sets and in the Empirical Data



Note. Left panel. Histogram of Bayes factors (BF10) of the alternative model over the null model in 50 simulated data sets. The vertical solid black line shows equal evidence for both hypotheses; the dashed red line shows the Bayes factor computed from the empirical data; the horizontal error bar shows 95 percent of all Bayes factors. Right panel: Estimates of the facilitatory effect of retrieval interference and 95 percent credible intervals across all simulations (black, solid lines) and the empirically observed data (red, dashed line). See the online article for the color version of this figure.

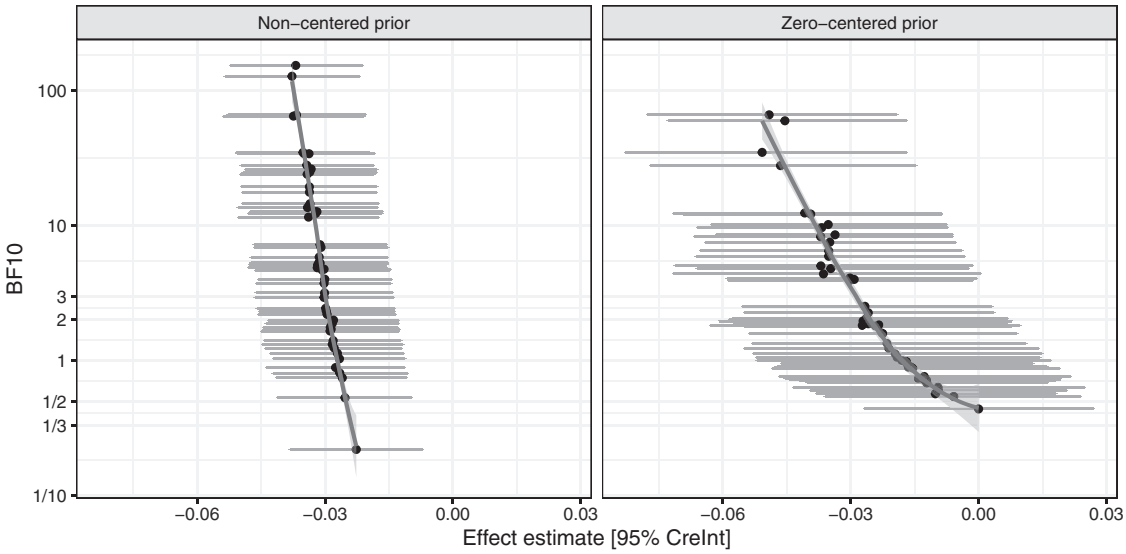
differences in reading times that are *less* negative/more positive than this prior mean are more in line with a null model of no effect. This also leads to the striking observation that the 95% credible intervals are quite consistent and none of them overlap with zero, whereas the Bayes factor results are far more variable. Computing Bayes factors for such a prior with a nonzero mean asks the very specific question of whether the data provide more evidence for the effect size obtained from the meta-analysis compared to the absence of any effect.

Case 2: Using a Prior Centered on Zero

We now use the same simulated data again. However, in the analysis of each simulated data set, we use a different prior for the agreement attraction effect than before: We do not use the prior informed by the meta-analysis (with a nonzero prior mean), but instead we use a centered prior, where the prior mean is zero. That is, we are agnostic with respect to the direction of the effect (using a prior standard deviation of .3; see the sensitivity analysis below).

Figure 6

Bayes Factor (BF10) as a Function of the Effect Estimate (With 95 Percent Credible Intervals) for 50 Simulated Studies



Note. Left panel: Same results as shown in the previous figure, with a noncentered prior from a meta-analysis. Right panel: Centered prior with mean 0 and standard deviation 0.3. See the online article for the color version of this figure.

For this prior centered on zero the Bayes factors now show a slightly different result. As is displayed in Figure 6 (right panel), Bayes factors now follow a hockey-stick function. For large negative differences between reading times (i.e., right panel, left side), Figure 6 again shows evidence in favor of the alternative hypothesis. When the estimated difference between reading times approaches zero (right panel, right side), Bayes factors show support for the null hypothesis. However, this support for the null is now much less pronounced (only anecdotal, i.e., BF10 $>$ 1/3). This is the case as the alternative hypothesis (H1) now specifies a prior mean of zero for the effect size, such that even an estimated effect size of zero can still be explained by the alternative model, and the null model is not so much better in explaining the data.

Case 3: A Study With an Effect Size Close to Zero

Next, we show an example case where the effect size in the original empirical study used to define the prior is very close to zero, that is, there is no difference between experimental conditions (Wagers et al., 2009; Experiment 3, singular). We use the mean-centered prior (prior M = 0, prior standard deviation = .3) for model fitting and Bayes factor computations. Figure 7 shows that the Bayes factor gets positive, providing some support for the alternative model not only for negative estimated effect sizes, but also for positive estimated effect sizes. Any difference between experimental conditions - negative or positive - can support the alternative model.

Case 4: A Large Sample Study

Last, the previous studies had relatively limited sample sizes, for example, the study by Lago et al. (2015; Experiment 1) had 32 subjects, and the study by Wagers et al. (2009; Experiment 3, singular) had data from 60 subjects. We now want to see how stable Bayes factors are in a situation where the sample size is relatively large. For this, we perform the same Bayes factor analysis for data from a study by Jäger et al. (2020), which contains eye-tracking reading time (total reading time) data on agreement attraction from 181 subjects. The results are displayed in Figure 8. The figure shows that with 181 subjects, the Bayes factor is quite stable. That is, across 50 simulated data sets, the Bayes factor computed on the simulated data always ranges between 1.3 and 1.7. An important insight here is that in large sample studies, the "dance of the Bayes factor" is very limited to a narrow range, and Bayes factors are quite stable. This may in part be the case because the posterior predictive distribution was so narrow that all the simulated data sets were very much alike.

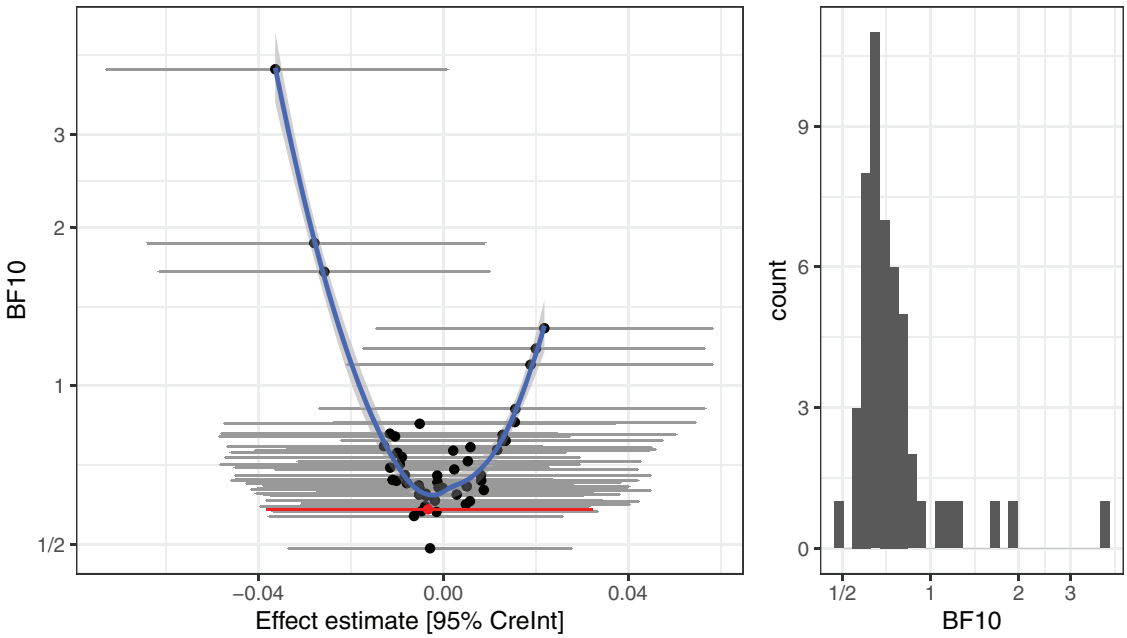
How Consistent is the Bayes Factor Across Multiple Studies?

Above, we investigated variability of Bayes factors over simulated replications of a given empirical study. Here, we go one step further. Instead of relying on simulated replications of the same experiment, we take real data from real empirical replications of the same type of experimental study. This allows us to investigate the extent to which the results from Bayes factor analyses vary from study to study, even if the same experimental effect is investigated.

We obtained the experimental data from a set of different studies that all investigate (inter alia) agreement attraction in ungrammatical sentences during sentence processing (Avetisyan et al., 2020; Dillon et al., 2013; Lago et al., 2015; Wagers et al., 2009). They all investigate reading time on a target word, measured via self-paced reading or via eye-tracking. There are of course

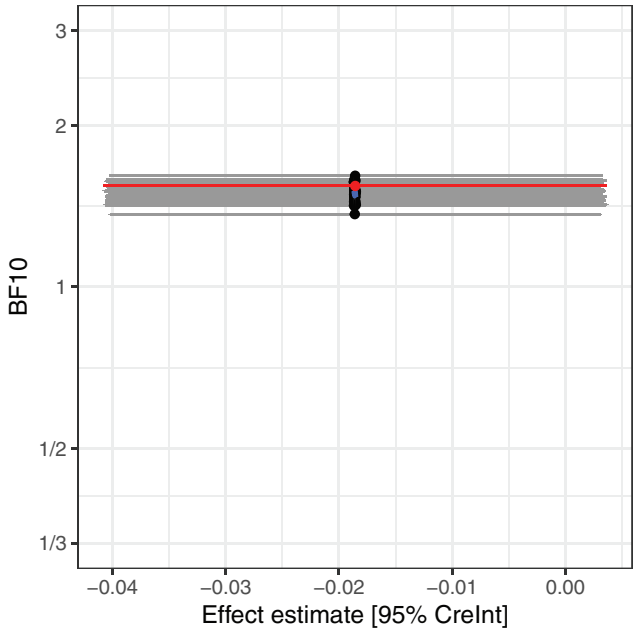
Figure 7

Results From a Study With No Facilitatory Effect (Wagers et al., 2009, Experiment 3, Singular)



Note. Centered prior with mean 0 and standard deviation 0.3. Bayes factor (BF10) as a function of the effect estimate (with 95 percent credible intervals) for 50 simulated studies. See the online article for the color version of this figure.

Figure 8
Analysis for a Large-Sample Study (Jäger et al., 2020)



Note. Centered prior with mean 0 and standard deviation 0.3. Bayes factor (BF10) as a function of the effect estimate (with 95 percent credible intervals) for 50 simulated studies. See the online article for the color version of this figure.

important differences between the studies: some investigate English, others Spanish or Armenian; and the syntactic configurations differ across studies. However, they all investigate the same phenomenon (agreement attraction), and are therefore trying to estimate the same effect. Importantly, agreement attraction is generally thought to be a robust empirical phenomenon (Phillips et al., 2011); this example therefore provides an example case for an empirically well-established effect in the cognitive sciences.

For all of these studies, we focus on the question of whether there is evidence for a difference in mean reading times between sentence types. Again, we use Bayesian modeling using `brms` for posterior estimation, assuming a zero-centered prior for the agreement attraction effect. As before Bayes factors are computed using bridge sampling. A sensitivity analysis is performed for each of the data sets separately (for details on how to perform and interpret a sensitivity analysis, see the next section entitled Issue 4: Bayes Factors Can Be Highly Sensitive to Priors).

Figure 9 visualizes the results of this analysis. It shows that the evidence in support of agreement attraction effects is very weak for each of the analyzed studies. One study (Wagers et al., 2009; Experiment 4) shows at least moderate evidence ($BF_{10} > 3$) for an interference effect of small size (the prior standard deviation of .040 shows the largest Bayes factor). Moreover, several of the other studies also show some evidence for small agreement attraction effects, but this evidence is anecdotal at best, with maximal Bayes factors ranging between 1 and 3.

What the analysis consistently shows, however, is that (a) in all studies the estimated effect is in the expected direction, that is, it is negative (except in Avetisyan et al., 2020; where there is a numerically positive effect); and (b) all studies provide evidence against a large agreement attraction effect. For the largest studied prior standard deviation of .4, the results show at least moderate evidence for the null model and against the alternative for 10 out of the 11 data sets, and three data sets actually provide strong evidence ($BF_{10} < 1/10$) against such a large prior effect size.

Moreover, the analysis reveals large variability in the results across data sets. Although the analysis of some data sets suggests tentative evidence for the alternative model, supporting agreement attraction effects in sentence comprehension (e.g., Wagers et al., 2009; Experiment 4), other data sets show no evidence for agreement attraction effects at all (e.g., Wagers et al., 2009; Experiment 3, singular).

The analysis shows that no data set provides strong evidence for H1. This might be quite surprising to the reader, given that agreement attraction effects are generally thought to be a robust empirical phenomenon (Phillips et al., 2011). What we illustrate here is that individual studies may in fact carry quite limited information about a fairly standard experimental effect. Indeed, standard experimental designs and sample sizes may be insufficiently sensitive to accurately detect a typical cognitive effect such as agreement attraction. Evidence synthesis through meta-analyses will be needed to make clear inferences about the effect of agreement attraction (an example meta-analysis investigating prediction effects in ERP data is discussed in Nicenboim et al., 2020).

Meta-analyses can be performed using Bayesian modeling, and again, Bayes factors can be used to quantify the evidence a meta-analysis provides in favor of some hypothesis. We illustrate this point here. First, we run frequentist linear mixed effects models for each of the data sets (using the function `lmer()` from the R package `lme4`). We fit frequentist models here because most published research tends to report frequentist models. From each analysis, we save the estimated agreement attraction effect and its associated standard error.

Based on these estimates, we perform a Bayes factor meta-analysis. Here, we estimate one Bayesian model assuming the alternative hypothesis (H1) for each of the different values of the prior standard deviation. Each H1 model is specified in `brms` as `b | SE(SE) ~ 0 + Intercept + (1 | expt)`, indicating that we model the variability in frequentist effect estimates (`b`) and their standard errors (`sd(SE)`) by using an intercept (Intercept) as well as random intercepts across experiments ((1 | expt)). Each of these alternative models (H1) is then compared to the same null model (H0), which in `brms` is encoded as `b | SE(SE) ~ 0 + (1 | expt`; i.e., assuming an intercept of zero), and which is identical to the H1 models otherwise. The comparison is done by running bridge sampling on each model and comparing marginal likelihoods to obtain Bayes factors.

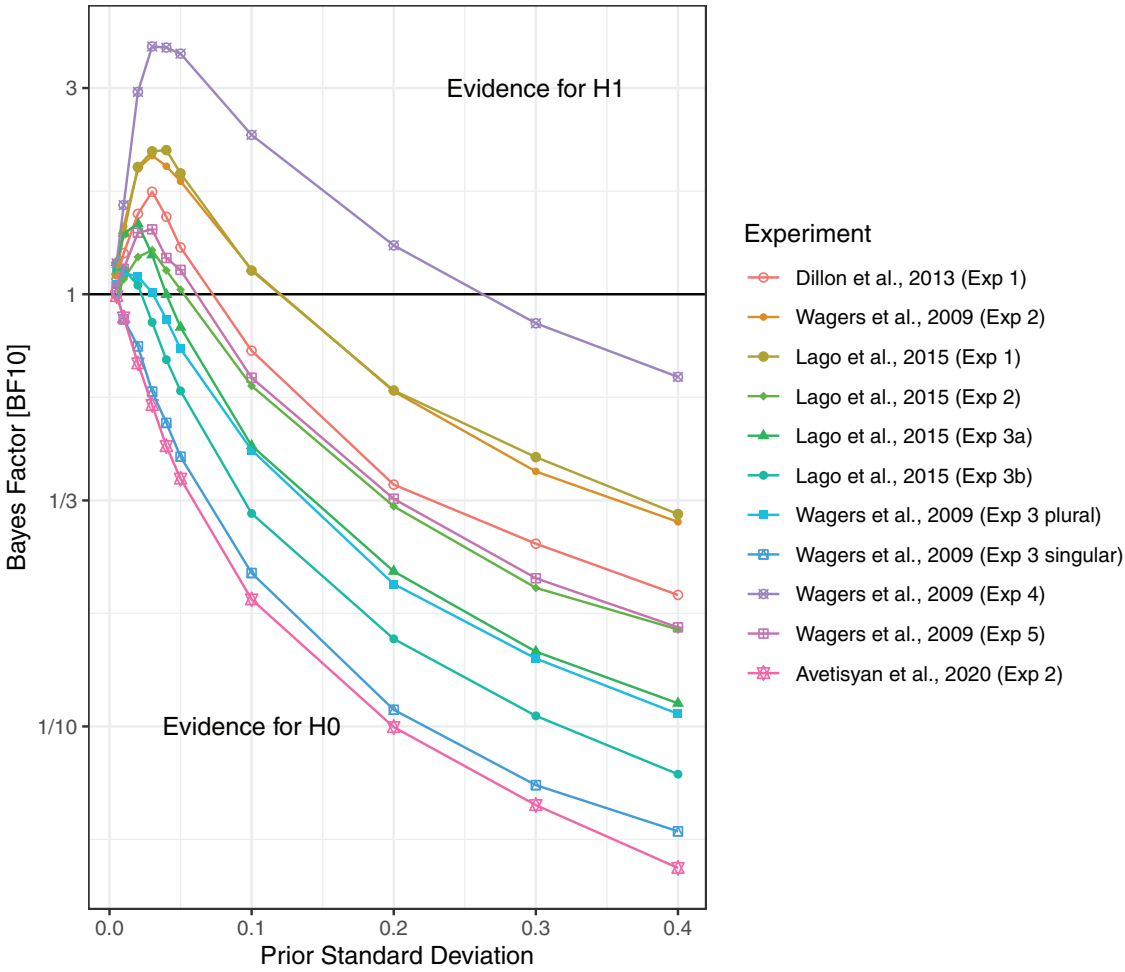
The results from this meta-analysis using a sensitivity analysis with the Bayes factor (see Figure 10) shows that once we synthesize the available evidence from the different studies, there is extreme evidence ($BF_{10} > 100$) for the alternative hypothesis that agreement attraction effects exist. Thus, while the individual studies, each considered separately, do not provide much evidence for the effect, combining studies into a Bayesian meta-analysis clearly shows there is evidence for the effect existing given the available data.

Issue 4: Bayes Factors Can Be Highly Sensitive to Priors

In the above example, there was good prior information from a meta-analysis about the free model parameter β. However, what happens if we are not sure about the prior for the model

Figure 9

Prior Sensitivity Analyses for Different Empirical Data Sets, Each Implementing a Replication Study of Interference Effects of Number Attraction in Sentence Comprehension



Note. For each empirical study (indicated by different colors and shapes), the Bayes factor (BF10) of the alternative model against the null model is shown for different prior standard deviations for the size of the experimental effect. See the online article for the color version of this figure.

parameter? It might happen that we compare the null model with a very "bad" alternative model, because our prior for β is not appropriate.

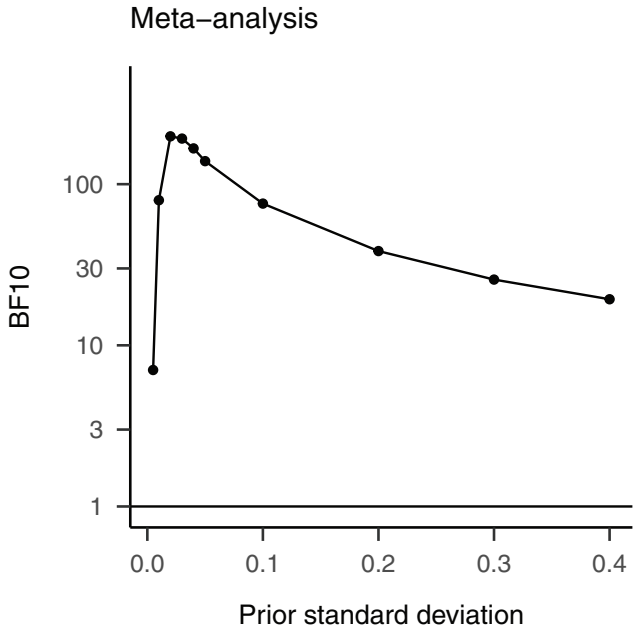
To deal with this situation, many authors use or recommend default prior distributions, where the priors for the model parameters are fixed (e.g., at the scale of an effect size), and are independent of the scientific problem in question, and of potential subjective perspectives on it (Morey & Rouder, 2011; Navarro, 2015; Rouder et al., 2009; Zellner & Siow, 1980). While Rouder et al. (2009) provide default priors that are appropriate to generic situations, they also (p. 235) state: "simply put, principled inference is a thoughtful process that cannot be performed by rigid adherence to defaults." In other words, they point out that it is important to consider alternative values for the prior; a sensitivity analysis is crucial component of a Bayes factor-based analysis.

Sensitivity analysis refers to examining different alternative models, where each model uses different prior assumptions. This way, it's possible to investigate the extent to which the Bayes factor results depend on, or are sensitive to, the prior assumptions (for an example from psycholinguistics, see Nicenboim et al., 2020). Next, we will perform a sensitivity analysis for the agreement attraction effect for the data by Lago et al. (2015).

For all of the priors for β (i.e., the effect of sentence type x) we assume a normal distribution with a mean of zero since we do not want to make any assumption a priori about the direction of the effect. What differs between the different priors is their standard deviation (using different values ranging from $SD = .005$ to $SD = .4$). A large standard deviation allows for very large effect sizes a priori, whereas a small standard deviation implies the assumption that we expect the effect not to be very large a priori.

Note that a sensitivity analysis is a case of inference over model space (i.e., with many different models), where one reports the entire model posterior instead of choosing any particular model (i.e., a particular prior). Importantly, the difference between

Figure 10
Sensitivity Analysis for the Bayesian Meta-Analysis



Note. The Bayes factor (BF10) as a function of the prior standard deviation provides extreme evidence in favor of the effect.

inference and decision making is critical here. The posterior provides continuous evidence about models with different priors, but it does not support decision making, that is, selection of individual models, without using a utility function. We will discuss this critical distinction further below.

We run the brms model for each of 10 different priors. Then, the Bayes factor is computed using bridge sampling against the null model with the effect of the predictor x being assumed to be 0.

We plot the Bayes factors BF_{10}, that is, the evidence for the alternative model over the null model, as a function of the prior standard deviation (see Figure 11). The results show that there is very little evidence for an agreement attraction effect. The Bayes factors provide anecdotal evidence for the alternative model for small prior standard deviations (the maximum lies at a standard deviation of .03). At the same time there is anecdotal evidence against agreement attraction effects for models with a larger prior standard deviation (i.e., against a priori large effect sizes). Indeed, Bayes factors explicitly penalize models with wide priors if the data aren't consistent with large effect sizes. Note that these results do not directly support a decision to pick the model with standard deviation of .03 as the best model (without using utility functions)—the evidence in posterior inference is continuous rather than discrete.

The reason that the conclusion differs (sometimes dramatically) as a function of the prior is that priors are never uninformative when it comes to Bayes factors. The wide priors specify that we expect very large effect sizes (with some considerable probability), and there is relatively little evidence in the data for such large effect sizes.

Indeed, very recently Uri Simonsohn criticized Bayes factors because they might provide evidence in favor of the null and against a very specific alternative model, when the researchers only knew the direction of the effect (see https://datacolada.org/78a). This can happen when very uninformative vague priors are used (Montero-Melis et al., 2019), and provides a major motivation for using more informed prior distributions.

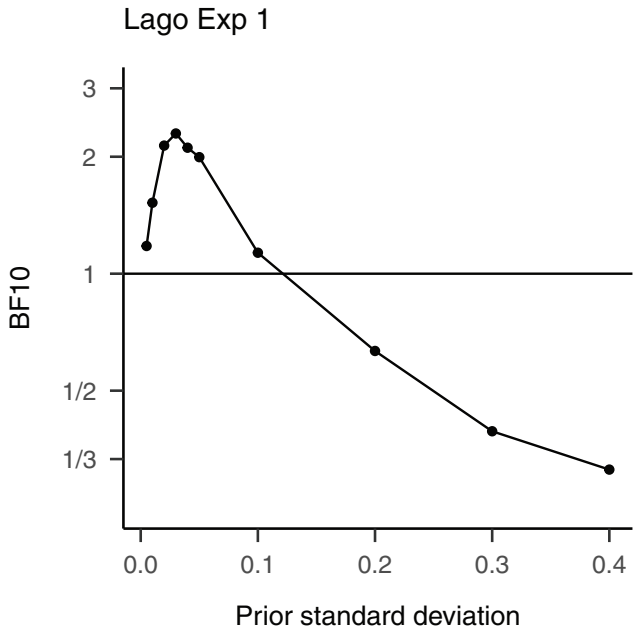
Overall, the outcome of the sensitivity analysis basically shows that the Bayes factors lie somewhere between 3 and 1/3, which all indicate inconclusive results.

Issue 5: Discovery Claims Using Bayes Factors Require Decision Theory

Bayes factors tell us how much evidence the data provide in favor of one model or another. That is, they allow us to perform inferences on the model space, that is, to determine how much each hypothesis is consistent with the data. Based on this evidence, it is also possible to perform decisions about selecting one hypothesis or the other, for example, to declare discovery based on a Bayes factor analysis. Several heuristics have been proposed on how such decisions can be made. For example, Jeffreys (1939) proposed a scale of how to put continuous evidence into discrete categories; these categories can then be used for decision-making. One common heuristic sometimes used in basic research is to treat Bayes factors that are larger than 10 (or smaller than 1/10) as grounds for declaring discovery. Another heuristic that is often used in machine learning is to select the model with the highest posterior probability. However, these are just heuristics, they are not principled ways to derive decisions from evidence.

To perform principled decisions based on Bayesian analyses, utility functions are needed. The utility of different possible actions, that is, the value of the consequences when accepting and

Figure 11
Sensitivity Analysis: Bayes Factor (BF10) as a Function of the Prior Standard Deviation



Note. The data are from the Lago et al. study, Experiment 1 (discussed as Cases 1 and 2 above).

acting based on one hypothesis or another, can differ quite dramatically in different situations. For example, for a researcher trying to implement a life-saving therapy, falsely rejecting this new therapy could have high negative utility (negative utility is loss), whereas falsely adopting the new therapy may have little negative consequences. By contrast, falsely claiming a new discovery in fundamental research may have bad consequences (high loss), whereas falsely missing a new discovery claim may be less problematic if further evidence can be accumulated. Thus, the performance of decision-making procedures can be determined only in the context of utility functions appropriate to a given analysis.

For example, in the cognitive sciences, if one claims a discovery based on a decision process, this can yield a true discovery (TD), which would have positive value, for example, a utility of U_{TD} = 1. However, a discovery claim can also be false (FD), yielding a possibly negative utility of $U_{FD} = -1.5$. Second, an alternative outcome of a decision-making process is to not claim discovery, but to reject it. Again, this can be a true negative (TN), which may have positive utility (e.g., $U_{TN} = .5$). However, the decision not to claim discovery can also be false (false negative, FN; i.e., missing a true new discovery), which might have a negative utility (e.g., of $U_{FN} = -.5$). Note that the utilities that we chose here are arbitrary, and other values could be chosen as well. In the cognitive sciences, decision making might, in general, be premature. If we cannot construct useful utilities then we probably shouldn't be trying to make decisions. Reporting inferences directly and avoiding discovery claims avoids having to worry about well-motivated utility functions. One research goal in the cognitive science would be to develop a procedure of how such utilities can be conceived in a way that is not arbitrary, but theoretically motivated.

Bayesian Decision Making Processes

To perform decisions in Bayesian analyses, the implementation of Bayesian decision making processes (Gelman et al., 2013; Robert, 2007) is necessary, which convert inferential information, such as the continuous Bayes factor or continuous posterior model probabilities, into discrete decisions.

Even if meaningful utility functions are available, there are still two important caveats associated with discrete decisions: first, in practice, we often work with estimators of Bayes factors rather than with true Bayes factors. Such estimators can be noisy and this noise will influence the decision making process. As the second caveat, because the inferential information varies with observations, so too will the decisions. Thus, random noise in the data can lead to very different inferences, and thus to very different decisions, simply based on chance.

Robust decision making requires sufficiently good experimental design to reduce the variation of the inferences, and hence the decisions, as much as possible. At the very least we have to quantify that variation to understand how stable a decision-making process will actually be. Indeed, because of the variation inherent in decisions, again, often making *no* decision may actually be the best approach! If one just reports the inferences (i.e., Bayes factors), others can make their own decisions using their own utility functions in combination with the full information in the reported inference.

Not only inference, but also decision making will vary depending on the data, and may perform well or badly (in terms of utilities) depending on what the data look like. The problem of course is that before running a study we do not know what the data looks like and what the possible outcomes of a study will be. Therefore, we need to quantify how those utilities can vary across different possible data sets using calibration studies (Betancourt, 2018). These can be implemented by simulating data based on some priors and models, where we thus know which model or hypothesis was true in the data simulation. Conveniently, the same simulations from the SBC can also be used to calibrate decision making processes based on Bayes factors. Then we can run a Bayesian decision procedure on each of the simulated data sets.

For example, we can look at simulated data sets where the true model (i.e., the model sampled from the model prior) corresponds to a null hypothesis (H0), perform decisions based on the Bayesian evidence, and obtain the false discovery rate (FDR), that is, how often the Bayes factor supports an alternative model (H1) when in fact the null hypothesis (H0) is true. This is a Bayesian equivalent of frequentist Type I error probability (α). Likewise, we can look at the simulated data sets where the true model (i.e., the model sampled from the prior) corresponds to an alternative hypothesis (H1), compute Bayesian evidence, perform decisions, and obtain the true discovery rate (TDR), that is, the probability of choosing the alternative hypothesis (H1) when it is actually true. This is a Bayesian equivalent of frequentist power analyses.

As discussed above, one immediate heuristic for turning Bayes factors into decisions is a threshold (e.g., of $BF_{10} = 10$), which can be used to determine true and false discovery rates. A principled alternative to such heuristics is to define a utility function. Interestingly, based on this, one can try out different Bayes factor thresholds (e.g., $BF_{10} = \{5, 10, 15\}$) and compute the total utility for each. Then, it is possible to choose the optimal threshold value to yield optimal total utility. This procedure may even compensate for experimental design limitations, which are sometimes unavoidable.

Using SBC Simulations to Calibrate Decisions

Above, we had used SBC to calibrate the continuous evidence obtained by computing Bayes factors. An alternative way to look at the results from the SBC analysis is to use thresholds on the Bayes factor to make discrete decisions, such as the decision to declare discovery.

In a first approach, here we use one conventional threshold sometimes used in practice, that is, that a Bayes factor larger than 10 provides (strong) evidence for the alternative hypothesis (H1), a Bayes factor smaller than 1/10 provides (strong) evidence for the null hypothesis (H0), and a Bayes factor between 1/10 and 10 provides "moderate" or "anecdotal" evidence for either hypothesis. For simplicity, we here use the thresholds of 10 and 1/10. We can look at what decisions the Bayes factors support by looking at the simulated data from the SBC. We investigate decisions based on whether H0 or H1 was actually used to simulate the artificial data.

The results (see Table 6) show that when H0 was actually true in the data simulation, the Bayes factor provided no strong evidence ($10 > BF_{10} > 1/10$) in 87% of cases, and provided evidence for H0 in only 13% of simulations. When H0 was true, it never decided for H1, reflecting a false discovery rate (FDR) of zero. Likewise, when H1 was actually true in the data simulation, the Bayes factor provided no strong evidence in 52% simulations, and provided evidence for H1 in 48% of cases (i.e., the TDR). However, it never

Table 6
Percentages of Supported Hypotheses (H0, H1, No Hypothesis) as a Function of Which Hypothesis Was True in the Simulations (H0 versus H1)

True hypothesis	Evidence for H0	No evidence	Evidence for H1
H0	13	87	0
H1	0	52	48

provided evidence for H0. These results show that for this example case of a small artificial data set with rather strong effect sizes, the Bayesian decision rule is often uncertain about the true hypothesis, but that it does not decide for the false hypothesis.

An alternative Bayesian decision-rule that is sometimes used in practice is to choose the model that has the highest posterior probability (note that this does not involve the possibility to be undecided). For the present data set this shows (see Table 7) a false discovery rate of 9% and a true discovery rate of 67%, again suggesting that the effect size and experimental design in the artificial data set were, arguably, sufficient for detecting a true effect from the data with a reasonable accuracy.

Bayesian Calibration of Frequentist Analysis Methods for the Same Data

It is possible to compare the results from the Bayesian calibration of Bayes factor analyses with corresponding analyses of frequentist analysis tools for the simulated data. In frequentist analyses, H0 is rejected if the p-value, that is, the probability for obtaining an effect as extreme as observed or stronger under the null hypothesis, is small. In frequentist statistics, a finding is considered statistically significant if the p-value is smaller than some threshold, which is conventionally $p < .05$ (see Benjamin et al., 2018; for an alternative threshold of $p < .005$); H0 is not rejected otherwise, that is, if $p > .05$. Unlike in Bayesian data analysis, frequentist null hypothesis significance testing often favors one hard cut-off value. Based on such a cut-off, when H0 was used to generate artificial simulated data, we can compute the number of times that H0 is falsely rejected, that is, compute the α error rate. Moreover, in the cases where H1 was used to simulate the artificial data, we can compute how often the frequentist model rejects H0 correctly, reflecting statistical power. For this, we fit a frequentist linear mixed-effects model using the lmer function to each of the 500 simulated data sets.

We can now apply the $p < .05$ cut-off, and compute how often H0 is rejected when H0 is true, and how often H0 is rejected when H1 is true (note that $p > .05$ is not a good cut-off to accept H0, but simply to fail to reject it). The results (see Table 8) show that when the null hypothesis is actually true (first row of the table), the empirical alpha error is estimated as 3%, which is reasonably close to the expected value of 5%. When the alternative hypothesis is true (i.e., second row of the table), statistical power is estimated to lie at 59%. This result shows that the effect size in our artificial example is large enough to detect it in the small data set with intermediate (but not with good) power.

Importantly, this Bayesian calibration analysis is different from a standard frequentist simulation analysis of the α and β errors. In this Bayesian analysis, we assume uncertainty about the exact effect size because the priors are specified as distributions. In frequentist analyses, one usually assumes a single fixed effect size and computes α and β errors for exactly this effect size, possibly without considering uncertainty that exist about the precise effect size.

The results from the calibration analyses show similarities and differences in the calibration between the Bayesian decision rule and corresponding frequentist null hypothesis significant testing. The Bayesian and frequentist decisions were similar as both had a fairly good chance of detecting a true H1 from the data (for the Bayes factor decision: 48%; for the posterior probability decision: 67%; for the frequentist decision: 59% of the true H1 were detected). However, by construction the Bayes factor decision rule distinguishes between cases where there is no evidence and cases where support for H0 can be observed. The Bayes factor decision rule ($BF_{10} < 1/10$) provided correct support for H0 in 13% of the simulations, which is very low sensitivity for detecting a null hypothesis. The frequentist analysis (and the decision rule based on posterior probabilities) by construction does not distinguish between situations of "no evidence" versus "evidence for H0."

Principled Decisions Using Utility Functions

The decision-rules that we have studied in the previous sections (based on Bayesian and frequentist analyses) relied on conventions about thresholds that would determine a decision, for example, on whether to declare discovery. However, as noted before, these conventions provide no principled approach on how to perform decisions in a Bayesian setup. Utility functions can be defined to specify the value of the consequences that originate from a given decision rule. Utility refers to the value that is associated with possible choices under given truths. The threshold we used above, that is, to choose the model with the highest posterior probability, is optimal if all of the possible decision-truth combinations have equal utility. However, in practice, the different combinations may have different utilities, which necessitates the definition of utility functions.

Here we define an exemplary utility function to support discrete decision-making. Let's assume that we make a decision between two options: claiming a discovery or not claiming a discovery. Then let's assume that a true discovery (TD) has utility of $U_{TD} = 10$ and that

Table 7
Percentages of Supported Decisions (No Discovery, Discovery) as a Function of Which Hypothesis Was True in the Simulations (H0 versus H1)

True hypothesis	No discovery	Discovery
H0	0.91	0.09
H1	0.33	0.67

Table 8
Percentages of Supported Frequentist Decisions (Don't Reject H0, Reject H0) as a Function of Which Hypothesis Was True in the Simulations (H0 versus H1)

True hypothesis	Don't reject H0	Reject H0
H0	97	3
H1	41	59

failing to claim a true discovery (i.e., a false negative, FN) has utility $U_{FN} = -5$. Moreover, we assume that a false discovery (FD) has utility (loss) of $U_{FD} = -50$, whereas a correct negative result (true negative, TN) has utility $U_{TN} = 5$. These numbers seem rather arbitrary for the kind of basic research applications in the cognitive sciences that we have in mind. A key problem here is that it is not clear how to choose these numbers appropriately. However, thresholds for *p*-values or for labeling results from Bayes factor analyses are also arbitrary, and are only fixed by convention. Importantly, such threshold conventions define implicit utility functions, which may or may not be relevant to a given problem. Utility analyses explicitly quantify the consequences of different possible actions. Specific utility functions could be agreed upon by research communities. One problem with utility functions is that it is currently unclear what procedure could be used to quantify such utilities. That is, how can we quantify the utility of a false positive published finding, for example, measured by the number of false citations. Research is needed in the cognitive sciences to investigate how utilities can be quantified and linked to evidence, yielding procedures for their definition. Alternatively, given that clear and good utility functions are hard to derive, one viable approach is to not make any decisions, but rather to communicate continuous evidence.

Next, we can compute the average expected utility given a certain decision threshold. For this, we define an index matrix *TA* ("truth-action"), where each column indicates one combination of truth and actions. For example, Column 1 would indicate all cases in the simulations where H0 was true (i.e., the data was simulated based on H0), and the decision procedure decided to claim discovery (i.e., false discovery). Column 2 would indicate cases where H0 was true and no discovery was claimed (true negative). Column 3 would indicate cases where H1 was true and discovery was claimed (true discovery), and Column 4 indicates cases where H1 was true and no discovery was claimed (false negative). Each row of the index matrix *TA* corresponds to one simulated data set from the SBC, and for each simulated data set the matrix indicates via a 1 which truth-action combination was realized in the SBC simulations with a given decision-rule, whereas all other truth-action combinations are marked with a 0. Table 9 shows the first rows of the index matrix *TA*.

Moreover, we define a vector of utilities u for these four different possible truth-action combinations: $u = \{-50,5,10,-5\}$ (see Table 10). Based on these definitions, we can now compute the average expected utility (averaged across all N simulated data sets) as:

$$\text{average expected utility} = \frac{1}{N}\sum_{n=1}^{N} TA \times u = \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{4} TA_{n,j} \cdot u_j \tag{10}$$

In this example, the average expected utility for a decision-rule that chooses the hypothesis with the highest posterior probability is 2.50. Here, we used the posterior probabilities for decision making. An alternative approach is to use decision rules based on Bayes factors instead. Let's use the threshold $BF_{10} \geq 10$ for a discovery claim. Based on this new discovery threshold, we can compute a new index matrix *TA*, where for each data set there will be a new decision about whether discovery will be claimed ($BF_{10} \geq 10$) or not ($BF_{10} < 10$). We again use the same utility function as used before (see Table 10). Based on this *TA* matrix and utility function, we can again compute the average expected utility, now for this new decision-rule.

Table 9
First 10 Rows of the Truth-Action Matrix TA, Which Codes the Combination of the True Hypothesis (H0 versus H1) and the Action Taken (Declare Discovery Versus Declare No Discovery) for Each Simulated Data Set

Data	True H0/ discov.	True H0/no discov.	True H1/ discov.	True H1/no discov.
1	0	1	0	0
2	0	1	0	0
3	0	1	0	0
4	0	1	0	0
5	0	1	0	0
6	0	0	1	0
7	1	0	0	0
8	0	1	0	0
9	0	1	0	0
10	0	1	0	0

Now, the expected average utility is 3.56 and thus higher than before. We can vary the discovery (Bayes factor) *threshold* to select the threshold with the highest average utility. This means that we can try out different threshold values and compute the average expected utility based on each threshold. Ideally, we should choose a threshold for discovery that provides maximal utility, averaged across all simulated data sets. We perform this analysis here: We choose a lot of different values for the discovery thresholds (i.e., Bayes factor thresholds varying from 1 to 100) and compute the average utility for each of these discovery thresholds. The results from this analysis are shown in Figure 12. They show that for decision-rules using a low value for the Bayes factor threshold, the average utility is low. Intuitively, this means that if we claim discovery based on very little evidence (e.g., $BF_{10} > 2$), we have low utility because there might be a high chance of a false discovery, and a low chance for a true negative, which yields low utility. The largest average utility is obtained for a Bayes factor threshold of 7. For thresholds larger than 7, average utility declines again. Intuitively, average utility may be low because if we use a very high discovery threshold, then we have a lot of false negatives, and only few true discoveries, which on average again yields low utility.

Based on this analysis, one could thus call a discovery when the Bayes factor reaches a value of at least 7. In this case, the number of false positives and false negatives is likely low, while the number of true positives and true negatives is still high. That is, this yields an optimal policy for deciding on discovery given the utilities specified above and given the simulated data sets, models, and priors.

The TDR and FDR for this threshold are shown in Table 11. The results are very close to the results with a Bayes factor threshold of 10. Again, the FDR is 0. However, the TDR is now a bit larger and takes a value of 52%. Note that while analysis of TDR and FDR provide a good first approach, more elaborate rates can be defined when a "no decision" option is possible.

Discussion

We provided a discussion of the Bayesian quantification of evidence in favor of one of two alternative hypotheses and investigated

Table 10
Utilities for the Four Truth (H0/H1)—Action (No/Discovery) Combinations

	u
True H0, discovery claim	−50
True H0, no discovery claim	5
True H1, discovery claim	10
True H1, no discovery claim	−5

the performance of Bayes factors with respect to prior assumptions, effective sample size, simulation-based calibration, data variability, and utility functions. We implemented competing hypotheses in hierarchical Bayesian models using the R package `brms`, and tested these hypotheses against each other by estimating Bayes factors approximately using bridge sampling.

The results illustrate the strong dependence of Bayes factors on the prior assumptions, which calls for the use of (a) (weakly) informative priors (Schad, Betancourt, et al., 2021), and of (b) prior sensitivity analyses, to investigate Bayes factors for different prior assumptions about the size of the effect. Our results moreover illustrate challenges and limitations in the performance of Bayes factor analyses. First, we studied theoretical aspects of Bayes factor estimation. We showed that Bayes factors can be estimated in an unstable way because a very large effective sample size (of the MCMC sampler) is needed in order to obtain stable results from the bridge sampling algorithm. Moreover, we noted that even if Bayes factor approximations are stable, because bridge sampling does not come with strong guarantees, it is unclear whether approximate Bayes factor estimates are accurate, that is, whether approximate Bayes factor estimates correspond to the true Bayes factor. We showed how simulation-based calibration can be used to investigate whether Bayes factor estimates are accurate for a given application case (i.e., model, priors, and experimental design). Our results moreover showed that if the model is misspecified (leaving out random slopes), then posterior inference is incorrect, echoing the conclusions in Barr et al. (2013) for frequentist linear mixed models.

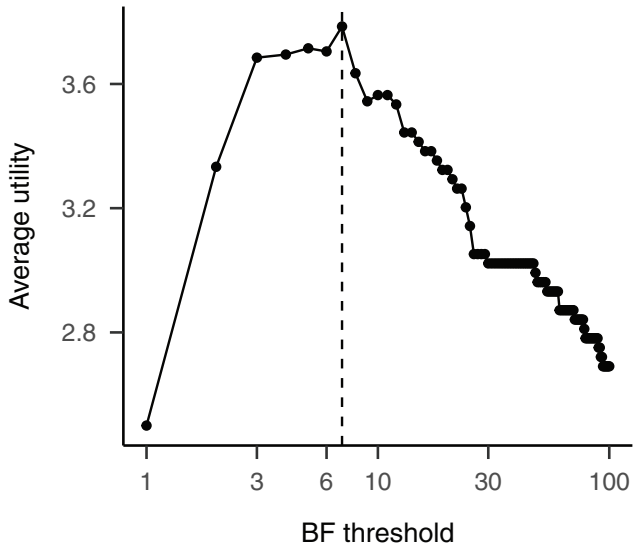
Second, analyses of artificial and of real replication data showed that the results from Bayes factor analyses—just like *p*-values in frequentist analyses - can considerably vary across different repeated replication attempts. However, for a range of different real empirical studies, the results also show some robustness against drawing strong conclusions. This suggests that some typical linguistic or psychological experiments may not be sufficiently powered to provide strong evidence for or against the small effect sizes that may be realistic to expect and that are of theoretical interest. Importantly, using Bayesian statistics and the Bayes factor does not solve this problem because low-powered studies will most likely yield inconclusive results in a Bayes factor analysis. As with frequentist approaches, for Bayesian analyses studies with larger sample sizes or stronger effect sizes may therefore also be needed, for example by sharing data across labs and through meta-analyses, to overcome such situations of low power.

Third, we studied decision making based on Bayesian analyses and saw how decisions can widely vary with the data. We discussed some heuristics for performing decisions, and illustrated how utility functions can be used to obtain optimal decisions.

Based on these challenges to the robustness of Bayes factors and the resulting inferences and decisions, we formulated a Bayes factor workflow, where simulation-based calibrations can be used to investigate these different issues for a given application case. This workflow then allows us to judge the extent to which inferences and decisions based on Bayes factors are robust for the given application case.

Taken together, in principle Bayes factor analyses provide a useful tool that can be used to investigate evidence for different hypotheses in the cognitive sciences. We showed how Bayes factors can misbehave based on estimation error, data variation, and poor Bayesian decision-procedures, partially reflecting that fact that widespread limitations in experimental design can also limit the conclusions that can be drawn based on individual data sets. We propose a Bayes factor workflow to identify these potential problems for a given application case. When used with care and appropriately calibrated, Bayes factors provide a useful approach for quantifying evidence and supporting decision-making on discovery claims in the cognitive sciences. At least in the types of data that we have investigated here, the Bayes factor cannot in general be used as a blunt instrument like the p-value, with some default prior specified for the target parameter.

Figure 12
Utility for Claiming Discovery as a Function of the Critical BF Cut-Off



Note. Note that with more (than 500) simulations, the line should become smooth.

Table 11
Percentages of Supported Decisions (No Discovery Claim, Discovery Claim) as a Function of Which Hypothesis Was True in the Simulations (H0 Versus H1), for an Optimal Discovery BF Threshold

True hypothesis	No discovery	Discovery
H0	100	0
H1	48	52

References

Aitkin, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society: Series B (Methodological)*, *53*(1), 111–128. https://doi.org/10.1111/j.2517-6161.1991.tb01812.x

Avetisyan, S., Lago, S., & Vasishth, S. (2020). Does case marking affect agreement attraction in comprehension? *Journal of Memory and Language*, *112*, 104087. https://doi.org/10.1016/j.jml.2020.104087

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., . . . Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Bennett, C. H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, *22*(2), 245–268. https://doi.org/10.1016/0021-9991(76)90078-4

Betancourt, M. (2019). *Probabilistic modeling and statistical inference*. GitHub repository. https://github.com/betanalpha/knitr_case_studies/tree/master/modeling_and_inference

Betancourt, M. (2020a). *Markov chain Monte Carlo*. GitHub repository. https://github.com/betanalpha/knitr_case_studies/tree/master/markov_chain_monte_carlo

Betancourt, M. (2020b). *Towards a principled Bayesian workflow (RStan)*. GitHub repository. https://github.com/betanalpha/knitr_case_studies/tree/master/principled_bayesian_workflow

Betancourt, M. (2016). *Diagnosing suboptimal cotangent disintegrations in Hamiltonian Monte Carlo*. arXiv Preprint arXiv:1604.00695.

Betancourt, M. (2017). *A conceptual introduction to Hamiltonian Monte Carlo*. arXiv. http://arxiv.org/abs/1701.02434

Betancourt, M. (2018). *Calibrating model-based inferences and decisions*. arXiv. https://arxiv.org/abs/1803.08393

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

Chow, S.-M., & Hoijtink, H. (2017). Bayesian estimation and modeling: Editorial to the second special issue on Bayesian data analysis. *Psychological Methods*, *22*(4), 609–615. https://doi.org/10.1037/met0000169

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

de Heide, R., & Grünwald, P. D. (2021). Why optional stopping can be a problem for Bayesians. *Psychonomic Bulletin & Review*, *28*(3), 795–812. https://doi.org/10.3758/s13423-020-01803-x

Dickey, J. M., & Lientz, B. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*(1), 214–226. https://doi.org/10.1214/aoms/1177697203

Dillon, B. W. (2011). *Structured access in sentence comprehension* [PhD thesis]. University of Maryland.

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, *69*(2), 85–103. https://doi.org/10.1016/j.jml.2013.04.003

Doorn, J. V., Aust, F., Haaf, J. M., Stefan, A., & Wagenmakers, E.-J. (2021). Bayes factors for mixed models. *PsyArXiv*. https://doi.org/10.31234/osf.io/y65h8

Engelmann, F., Jäger, L. A., & Vasishth, S. (2019). The effect of prominence and cue association on retrieval processes: A computational account. *Cognitive Science*, *43*(12), e12800. https://doi.org/10.1111/cogs.12800

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, *25*(1), 5–34. https://doi.org/10.3758/s13423-017-1262-3

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. https://doi.org/10.3758/s13423-017-1317-5

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A Statistics in Society*, *182*(2), 389–402. https://doi.org/10.1111/rssa.12378

Ge, H., Xu, K., & Ghahramani, Z. (2018). Turing: A language for flexible probabilistic inference. *Proceedings of Machine Learning Research*, *84*, 1682–1690. https://doi.org/10.17863/CAM.42246

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC Press.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97. https://doi.org/10.1016/j.jmp.2017.09.005

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). Bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, *92*(10), 1–29. https://doi.org/10.18637/jss.v092.i10

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, *44*(1), 133–152. https://doi.org/10.1006/jmps.1999.1280

Hammerly, C., Staub, A., & Dillon, B. (2019). The grammaticality asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence. *Cognitive Psychology*, *110*, 70–104. https://doi.org/10.1016/j.cogpsych.2019.01.001

Heck, D. W., Boehm, U., Böing-Messing, F., Bürkner, P.-C., Derks, K., Dienes, Z., Fu, Q., Gu, X., Karimova, D., Kiers, H., Klugkist, I., Kuiper, R. M., Lee, M. D., Leenders, R., Leplaa, H. J., Linde, M., Ly, A., Meijerink-Bosman, M., Moerbeek, M., & Dienes, Z. (in press). A review of applications of the Bayes factor in psychological research. *Psychological Methods*. https://psyarxiv.com/cu43g

Hendriksen, A., de Heide, R., & Grünwald, P. (2021). Optional stopping with Bayes factors: A categorization and extension of folklore results, with an application to invariant situations. *Bayesian Analysis*, *16*(3), 961–989. https://doi.org/10.1214/20-BA1234

Hoijtink, H., & Chow, S.-M. (2017). Bayesian hypothesis testing: Editorial to the special issue on Bayesian data analysis. *Psychological Methods*, *22*(2), 211–216. https://doi.org/10.1037/met0000143

Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, *94*, 316–339. https://doi.org/10.1016/j.jml.2017.01.004

Jäger, L. A., Mertzen, D., Van Dyke, J. A., & Vasishth, S. (2020). Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study. *Journal of Memory and Language*, *111*, 104063. https://doi.org/10.1016/j.jml.2019.104063

Jeffreys, H. (1939). *Theory of probability*. Clarendon Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Lago, S., Shalom, D., Sigman, M., Lau, E. F., & Phillips, C. (2015). Agreement processes in Spanish comprehension. *Journal of Memory and Language*, *82*, 133–149. https://doi.org/10.1016/j.jml.2015.02.002

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*(1), 1–7. https://doi.org/10.1016/j.jmp.2010.08.013

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal*

of Multivariate Analysis, *100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*(6), 362–375. https://doi.org/10.1016/j.jmp.2008.03.002

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*(4), 325–337. https://doi.org/10.1023/A:1008929526011

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32. https://doi.org/10.1016/j.jmp.2015.06.004

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, *6*(1996), 831–860.

Montero-Melis, G., Paridon, J. V., Ostarek, M., & Bylund, E. (2019). *Does the motor system functionally contribute to keeping words in working memory? A pre-registered replication of Shebani and Pulvermüller (2013, Cortex)* PsyArXiv. https://psyarxiv.com/pqf8k/

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of bayes factors for common designs*. https://CRAN.R-project.org/package=BayesFactor

Morey, R., & Rouder, J. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*(4), 406–419.

Mulder, J., & Wagenmakers, E.-J. (2016). Eds.' 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. https://doi.org/10.1016/j.jmp.2016.01.002

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79–95. https://doi.org/10.3758/BF03210778

Navarro, D. J. (2015). *Learning statistics with R*. https://learningstatisticswithr.com

Navarro, D. J. (2019). Between the devil and the deep blue sea: Tensions between scientific judgement and statistical model selection. *Computational Brain & Behavior*, *2*(1), 28–34. https://doi.org/10.1007/s42113-018-0019-z

Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational Ideas - Part II. *Language and Linguistics Compass*, *10*(11), 591–613. https://doi.org/10.1111/lnc3.12207

Nicenboim, B., Schad, D. J., & Vasishth, S. (2021). *An introduction to Bayesian data analysis for cognitive science*. https://vasishth.github.io/bayescogsci/book/

Nicenboim, B., Vasishth, S., & Rösler, F. (2020). Are words pre-activated probabilistically during sentence comprehension? Evidence from new data and a Bayesian random-effects meta-analysis using publicly available data. *Neuropsychologia*, *142*, 107427. https://doi.org/10.1016/j.neuropsychologia.2020.107427

Oelrich, O., Ding, S., Magnusson, M., Vehtari, A., & Villani, M. (2020). When are Bayesian model probabilities overconfident? *arXiv*. https://arxiv.org/abs/2003.04026

Phillips, C., Wagers, M. W., & Lau, E. F. (2011). Grammatical illusions and selective fallibility in real-time language comprehension. In J. T. Runner (Ed.), *Experiments at the interfaces* (Vol. 37, pp. 147–180). Emerald Bingley.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, pp. 1–10). Technische Universität Wien.

Rabe, M. M., Kliegl, R., & Schad, D. J. (2021). *Designr: Balanced factorial designs*. https://maxrabe.com/designr

Robert, C. (2007). *The Bayesian choice*. Springer-Verlag.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4

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. https://doi.org/10.3758/s13423-017-1420-7

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. https://doi.org/10.3758/PBR.16.2.225

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in python using pymc3. *PeerJ Computer Science*, *2*, e55. https://doi.org/10.7717/peerj-cs.55

Schad, D. J., & Vasishth, S. (2019). The posterior probability of a null hypothesis given a statistically significant result. *arXiv Preprint*. https://arxiv.org/abs/1901.06889

Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods*, *26*(1), 103–126. https://doi.org/10.1037/met0000275

Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., & Vasishth, S. (2021). *Bayes factors*. https://osf.io/y354c/

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, *110*, 104038.

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Sorensen, T., Hohenstein, S., & Vasishth, S. (2016). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *Quantitative Methods for Psychology*, *12*(3), 175–200. https://doi.org/10.20982/tqmp.12.3.p175

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes factor design analysis using an informed prior. *Behavior Research Methods*, *51*(3), 1042–1058. https://doi.org/10.3758/s13428-018-01189-8

Taatgen, N. A., Lebiere, C., & Anderson, J. R. (2006). Modeling paradigms in ACT-R. In R. Sun (Eds.), *Cognition And multi-agent interaction: Ftom cognitive modeling to social simulation* (pp. 29–52). Cambridge University Press.

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., & Gelman, A. (2018). Validating Bayesian inference algorithms with simulation-based calibration. *arXiv*. https://arxiv.org/abs/1804.06788

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

Tendeiro, J. N., & Kiers, H. A. (2021). On the white, the black, and the many shades of gray in between: Our reply to van Ravenzwaaij and Wagenmakers (2021). *PsyArXiv*. https://doi.org/10.31234/osf.io/tjxvz

Tendeiro, J. N., Kiers, H. A., & van Ravenzwaaij, D. (2021). Worked-out examples of the adequacy of Bayesian optional stopping. *Psychonomic Bulletin & Review*. Advance online publication. https://doi.org/10.3758/s13423-021-01962-5

van Ravenzwaaij, D., & Wagenmakers, E.-J. (2021). Advantages masquerading as "issues" in Bayesian hypothesis testing: A commentary on Tendeiro and Kiers (2019). *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000415

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (2018). Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, *25*(1), 1–4. https://doi.org/10.3758/s13423-018-1443-8

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*(6), 491–498. https://doi.org/10.1016/j.jmp.2010.07.003

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, *71*, 147–161. https://doi.org/10.1016/j.wocn.2018.07.008

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2020). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*, *16*(2), 667–718.

Wagenmakers, E.-J., Lee, M. D., Rouder, J. N., & Morey, R. D. (2020). The principle of predictive irrelevance or why intervals should not be used for model comparison featuring a point null hypothesis. In C. W. Gruber (Ed.), *The theory of statistics in psychology* (pp. 111–129). Springer.

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. https://doi.org/10.1016/j.cogpsych.2009.12.001

Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*(2), 206–237. https://doi.org/10.1016/j.jml.2009.04.002

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos de Estadistica Y de Investigacion Operativa*, *31*(1), 585–603. https://doi.org/10.1007/BF02888369

Received April 29, 2021
Revision received October 18, 2021
Accepted November 8, 2021 ■