A Review of Applications of the Bayes Factor in Psychological Research

Heck, D.W.; Boehm, U.; Böing-Messing, F.; Bürkner, P.-C.; Derks, K.; Dienes, Z.; Fu, Q.; Gu, X.; Karimova, D.; Kiers, H.A.L.; Klugkist, I.; Kuiper, R.M.; Lee, M.D.; Leenders, R.; Leplaa, H.J.; Linde, M.; Ly, A.; Meijerink-Bosman, M.; Moerbeek, M.; Mulder, J.; Palfi, B.; Schönbrodt, F.D.; Tendeiro, J.N.; van den Bergh, D.; Van Lissa, C.J.; van Ravenzwaaij, D.; Vanpaemel, W.; Wagenmakers, E.-J.; Williams, D.R.; Zondervan-Zwijnenburg, M.; Hoijtink, H.

DOI
10.31234/osf.io/cu43g
10.1037/met0000454

Publication date
2023

Document Version
Author accepted manuscript

Published in
Psychological Methods

(in press at Psychological Methods)

A Review of Applications of the Bayes Factor in Psychological Research

Daniel W. Heck[1], Udo Boehm[2], Florian Böing-Messing[3,4], Paul-Christian Bürkner[5], Koen Derks[6], Zoltan Dienes[7], Qianrao Fu[8], Xin Gu[9], Diana Karimova[10], Henk A. L. Kiers[11], Irene Klugkist[8], Rebecca M. Kuiper[8], Michael D. Lee[12], Roger Leenders[10,3], Hidde J. Leplaa[8], Maximilian Linde[11], Alexander Ly[2,13], Marlyne Meijerink-Bosman[10], Mirjam Moerbeek[8], Joris Mulder[10], Bence Palfi[7], Felix D. Schönbrodt[14], Jorge N. Tendeiro[11,15], Don van den Bergh[2], Caspar J. Van Lissa[8], Don van Ravenzwaaij[11], Wolf Vanpaemel[16], Eric-Jan Wagenmakers[2], Donald R. Williams[17], Mariëlle Zondervan-Zwijnenburg[8], and Herbert Hoijtink [8]

[1]University of Marburg

[2]University of Amsterdam

[3]Jheronimus Academy of Data Science

[4]Eindhoven University of Technology

[5]Aalto University

[6]Nyenrode Business University

[7]University of Sussex

[8]Utrecht University

[9]East China Normal University

[10]Tilburg University

[11]University of Groningen

[12]University of California Irvine

[13]CWI Amsterdam

[14]Ludwig-Maximilians-University München

[15]Hiroshima University

[16]KU Leuven

[17]University of California Davis

Version: January 7, 2022

Author Note

Worked examples and illustrations with data and analysis files are available at the Open Science Framework, `https://osf.io/k9c5q/`. The manuscript was uploaded to PsyArXiv and ResearchGate for timely dissemination (`https://psyarxiv.com/cu43g`).

The first author Daniel W. Heck proposed the idea for this paper. The last author Herbert Hoijtink initiated the paper. They jointly collected all contributions and constructed the paper and accompanying website. The names of all contributing authors are listed in alphabetical order. While working on this paper, the last author was supported by a fellowship from the Netherlands Institute for Advanced Study in the Humanities and Social Sciences (NIAS-KNAW).

Correspondence concerning this article should be addressed to Daniel W. Heck, Department of Psychology, University of Marburg, Gutenbergstraße 18, 35032 Marburg, Germany (dheck@uni-marburg.de) or to Herbert Hoijtink, Department of Methodology and Statistics, Utrecht University, P.O. Box 80140, 3508 TC, Utrecht, The Netherlands (h.hoijtink@uu.nl).

Abstract

The last 25 years have shown a steady increase in attention for the Bayes factor as a tool for hypothesis evaluation and model selection. The present review highlights the potential of the Bayes factor in psychological research. We discuss six types of applications: Bayesian evaluation of point null, interval, and informative hypotheses, Bayesian evidence synthesis, Bayesian variable selection and model averaging, and Bayesian evaluation of cognitive models. We elaborate what each application entails, give illustrative examples, and provide an overview of key references and software with links to other applications. The paper is concluded with a discussion of the opportunities and pitfalls of Bayes factor applications and a sketch of corresponding future research lines.

Translational Abstract

The last 25 years have shown a steady increase in attention for the Bayes factor as a tool for hypothesis evaluation and model selection. The Bayes factor provides a method for quantifying the relative evidence for two competing hypotheses that are both instantiated by specific statistical models with prior distributions on the parameters. This general approach can be used to address many specific, theoretically relevant research questions. The present review highlights the potential of the Bayes factor in psychological research. We discuss six types of applications: whether a randomized experiment has an effect or not (point null hypothesis), whether an effect is inside or outside a range of negligible effect sizes (interval hypothesis), whether a set of means follows a specific order (informative hypothesis), whether a set of studies jointly corroborate a theoretical claim (evidence synthesis), which variables are most relevant for prediction (variable selection), and which model provides the best account of latent processes (cognitive modeling). We elaborate what each application entails, give illustrative examples with reproducible files for the software R and JASP, and provide an overview of key references and software with links to other applications. We concluded with a discussion of the opportunities and pitfalls of the Bayes factor.

Keywords: Bayes Factor, Evidence, Hypothesis Testing, Model Selection, Theory Evaluation

A Review of Applications of the Bayes Factor in Psychological Research

Introduction

The paper by Kass and Raftery (1995) can be seen as the starting point for an increased interest in hypothesis evaluation and model selection using the Bayes factor. Now, 25 years later, this paper reviews the current state of affairs with a focus on applications of the Bayes factor in psychological research. The target audience are psychological researchers who consider using the Bayes factor for the analysis of their data. After introducing the Bayes factor, an overview of six types of applications will be given. Each type will be introduced, an example is given, and the interested reader is directed to the supplementary repository on the Open Science Framework (OSF, `https://osf.io/k9c5q/`) which contains a separate folder for each section containing vignettes with additional examples, references, and links to software. The website will connect users and developers of Bayes factors and will thus support the application of Bayes factors in psychological research.

The Bayes Factor

To ensure accessibility of this paper to a wide range of psychological researchers, the concepts (which will in this section be printed in bold face) that are relevant when using the Bayes factor will be introduced using a simple example. Readers who want to know more about the statistical underpinnings of the Bayes factor are first of all referred to Kass and Raftery (1995), Edwards et al. (1963), and Myung and Pitt (1997) for general introductions, and secondly, for specific applications, to the references given in the corresponding sections that follow below.

Our simple example is inspired by the experiments with respect to extrasensory perception (ESP) presented in Bem (2011). Imagine that each of $n = 40$ persons looks at the backside of two cards, one card hiding the number 7, the other hiding an erotic picture, and, subsequently, guesses which card hides the erotic picture. The **data** resulting from this experiment can be summarized as $x = 26$, that is, 26 persons picked the erotic picture as predicted by Bem's hypothesis of "precognitive detection of erotic stimuli."

The Bayes factor quantifies the support in the data for two competing statistical **models**. Examples of models are the various types of analysis of variance, regression, and structural equation models. Each model makes specific predictions for the data and must be specified *before* the data are inspected. If one does not believe in ESP, the data resulting from our imaginary experiment can be modeled using:

$$\mathcal{M}_1 : x \sim \text{Binomial}(n = 40, \theta = .50). \tag{1}$$

This model assumes that the number of successes follows a binomial distribution with $n = 40$ trials and probability θ of guessing "correctly" equal to .50, meaning that the choice for one or the other card is completely random and unsystematic. If one does believe in ESP, the model can be

$$\mathcal{M}_2 : x \sim \text{Binomial}(n = 40, \theta \neq .50), \tag{2}$$

that is, the probability θ of choosing the card with the erotic stimulus differs from .50.

Frequentist inferences (often referred to as classical statistics) could now continue by estimating the parameter θ and computing a 95% confidence interval. For the example at hand, the estimate would be $\hat{\theta} = 26/40 = .65$ with a confidence interval of $[.48, .79]$. Since the critical test value is within this interval, the model $\mathcal{M}_1$ assuming $\theta = .50$ cannot be rejected at a significance level of $\alpha = 5\%$.

Bayesian inference requires, for both models, the specification of a **prior distribution** for the parameter θ. A prior distribution specifies for each model which values of the parameter θ are considered to be more and less likely before seeing the data. $\mathcal{M}_1$ completely specifies θ as being exactly equal to .50. Therefore, the corresponding prior distribution also specifies that $\theta = .50$ is the only option. However, $\mathcal{M}_2$ does not completely specify θ; it does state that $\theta \neq .50$, but a prior distribution is needed to quantify the uncertainty regarding the expected effect size of ESP.

Researchers can either assume a subjective or a default prior distribution. A **subjective prior distribution** reflects the expectations of a researcher. A researcher believing in ESP might specify $\theta \sim \text{Uniform}[.50, .60]$, meaning that all values of θ in the interval $[.50, .60]$ are equally plausible a priori. This prior is illustrated in Figure 1 (first

row, second panel) and reflects the believe that, if ESP exists, it is expected to be weak, resulting in a probability of choosing the card with the erotic stimuli larger than .50, but not much larger.In general, subjective prior distributions are an extension of models: in the example at hand, they change the comparison of $\mathcal{M}_1$ with $\mathcal{M}_2$ to the comparison of $\theta = .50$ with $\theta \sim \text{Uniform}[.50, .60]$ (subsequently referred to as $\mathcal{M}_{2a}$). A **default prior distribution** is not tailored to reflect the prior believe of a researcher. It is chosen such that the resulting Bayes factors are well calibrated, that is, provide a statistically well-founded comparison of $\theta = .50$ with $\theta \neq .50$. For the example at hand, a default prior distribution could be $\theta \sim \text{Uniform}[0, 1]$, that is, each value in the interval from 0 to 1 is equally likely (subsequently referred to as $\mathcal{M}_{2b}$). Note that, as will be elaborated in the next section, the scale of many default prior distributions can be modified according to the researchers' expectations about the range of plausible effect sizes. To some degree, this re-introduces the use of subjective expectations to specify the prior distribution.

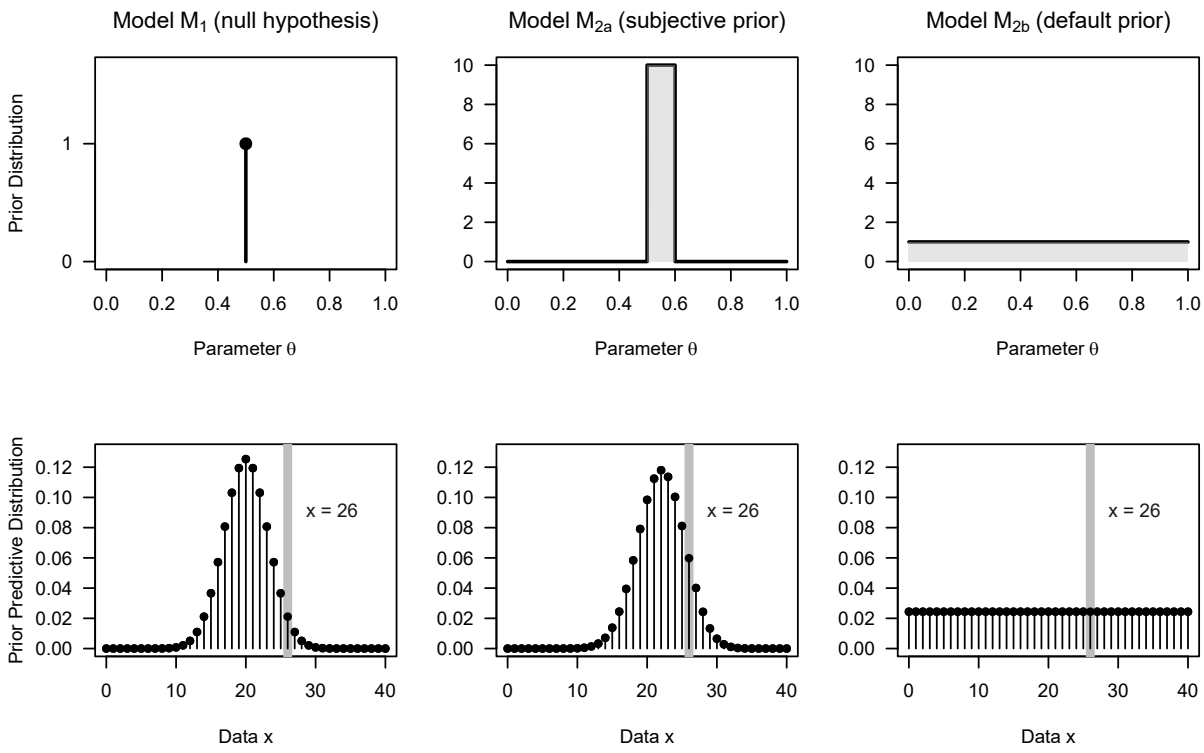
Now that prior distributions have been specified, we can assess the predictions of each model before observing any data. For this purpose, the **prior predictive distribution** provides the probability of observing a specific number of successes ($x = 0$, $x = 1, \ldots, x = 40$) conditional on a model and prior. This distribution is shown for all three models in the second row of Figure 1. Whereas $x = 20$ and $x = 22$ are the most likely observations under model $\mathcal{M}_1$ and $\mathcal{M}_{2a}$, respectively, all possible observations are equally likely under $\mathcal{M}_{2b}$. Based on the prior predictive distribution, we can directly obtain the **marginal likelihood** of the actually observed data given each model. For the example at hand, the marginal likelihood $P(x = 26 \mid \mathcal{M})$ is the probability of observing $x = 26$ "correct" guesses out of $n = 40$ trials given a specific model $\mathcal{M}$ with some prior distribution (highlighted in gray in the prior predictive distributions in Figure 1). The larger the marginal likelihood, the better this combination of model and prior predicts the data.

The **Bayes factor** is the ratio of the marginal likelihoods of the observed data for two models,

$$\text{BF}_{1,2a} = \frac{P(x = 26 \mid \mathcal{M}_1)}{P(x = 26 \mid \mathcal{M}_{2a})}, \tag{3}$$

Figure 1

Prior Distributions and Prior Predictive Distributions of Three Models.



Note. The first row shows three models with different prior distributions for the probability θ of choosing the card with the erotic stimulus in the ESP experiment. The second row shows the corresponding prior predictive distribution of observing x "correct" guesses out of $n = 40$ trials with the marginal likelihood $P(x = 26 \mid \mathcal{M})$ of each model highlighted in gray.

and thus compares the ability of $\mathcal{M}_1$ and $\mathcal{M}_{2a}$ to predict the actually observed outcome. If $\text{BF}_{1,2a}$ equals 9, the support in the data is 9 times larger for $\mathcal{M}_1$ than for $\mathcal{M}_{2a}$. If $\text{BF}_{1,2a}$ equals 0.2, the inverse $\text{BF}_{2a,1} = 1/\text{BF}_{1,2a} = 5$ shows that the support in the data is five times larger for $\mathcal{M}_{2a}$ than for $\mathcal{M}_1$. If the observed $\text{BF}_{1,2a}$ equals 108.45, 9.02, or 2.75, there is convincing, moderate, or unconvincing support for $\mathcal{M}_1$ over $\mathcal{M}_{2a}$, respectively [1]. For the data at hand (26 correct guesses out of 40 trials), $\text{BF}_{2a,1} = 2.83$ implies that the support in the data for the model $\mathcal{M}_{2a}$ is about three times stronger than for $\mathcal{M}_1$. This

[1] Note that these are subjective interpretations of observed Bayes factor values and not fixed reference values with specific verbal labels as presented in, for example, Jeffreys (1939) and Kass and Raftery (1995).

is illustrated graphically in the second row of Figure 1 showing that the data $x = 26$ (highlighted in gray) are about three times more likely under $\mathcal{M}_{2a}$ than under $\mathcal{M}_1$. In contrast, $\text{BF}_{2b,1} = 1.16$ implies that the support in the data for the models $\mathcal{M}_{2b}$ and $\mathcal{M}_1$ is about equal. This illustrates the effect of using a subjective prior that is corroborated by the data versus using a vague default prior: Compared to the null model assuming $\theta = .50$, one obtains some support for an ESP effect between .50 and .60 ($\mathcal{M}_{2a}$), but about equal support for an ESP effect between 0 and 1 ($\mathcal{M}_{2b}$), respectively. Hence, using a very wide and unspecific prior under the alternative hypothesis has the potential to hurt the chances of finding evidence for an effect. In general, the Bayes factor penalizes complex models (e.g., models with many parameters or vague priors) if the increase in complexity does not pay off in terms of a better fit, thus achieving an optimal trade-off between model fit and complexity (cf. Occam's razor; Myung & Pitt, 1997).

As exemplified, the Bayes factor is a continuous measure of the relative support for two models. This is also shown by the fact that the Bayes factor $\text{BF}_{1,2a}$ is simply the multiplicative factor required to update the **prior model odds** of $\mathcal{M}_1$ versus $\mathcal{M}_{2a}$ to the corresponding **posterior model odds**,

$$\underbrace{\frac{P(\mathcal{M}_1 \mid x = 26, n = 40)}{P(\mathcal{M}_{2a} \mid x = 26, n = 40)}}_{\text{Posterior model odds}} = \underbrace{\text{BF}_{1,2a}}_{\text{Bayes factor}} \times \underbrace{\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_{2a})}}_{\text{Prior model odds}} . \tag{4}$$

Hence, the Bayes factor quantifies how to update one's beliefs in the models in light of new data. Importantly, the Bayes factor is independent of the **prior model probabilities** $P(\mathcal{M}_1)$ and $P(\mathcal{M}_{2a})$ which represent the beliefs in the models *before* seeing the data. If researchers specify the prior model probabilities, the Bayes factor can be used to compute the **posterior model probabilities** $P(\mathcal{M}_1 \mid x = 26, n = 40)$ and $P(\mathcal{M}_{2a} \mid x = 26, n = 40)$ which represent the beliefs in the models *after* seeing the data (for details, see Equation 7). For example, if $P(\mathcal{M}_1 \mid x = 26, n = 40) = .80$ and $P(\mathcal{M}_{2a} \mid x = 26, n = 40) = .20$, the probability that $\mathcal{M}_1$ is the best of the two models equals 80%. The complement of the posterior model probabilities can be interpreted as **Bayesian error probabilities**, meaning that choosing the model $\mathcal{M}_1$ comes with a Bayesian (that is, conditional on the observed data) error probability of $1 - .80 = .20$.

Equation 4 shows that posterior model probabilities depend on the relative support in the data for the models (i.e., the Bayes factor) and an evaluation of the prior probability of the models *before* the data are inspected. There are again two types of prior model probabilities: default and subjective. "Default" usually implies that a priori each model under consideration is equally likely.[2] For the example at hand this would imply that $P(\mathcal{M}_1) = P(\mathcal{M}_{2a}) = 1/2$. Default prior model probabilities can be used if each of the models under consideration is considered to be an equally plausible description of the population from which the data are sampled. For the example at hand, these prior model probabilities would result in $P(\mathcal{M}_1 \mid x = 26, n = 40) = .26$ versus $P(\mathcal{M}_{2a} \mid x = 26, n = 40) = 1 - .26 = .74$ when assuming a subjective prior distribution on the probability θ, and $P(\mathcal{M}_1 \mid x = 26, n = 40) = .46$ versus $P(\mathcal{M}_{2b} \mid x = 26, n = 40) = 1 - .46 = .54$ when assuming a default prior distribution. This shows that posterior model probabilities result in identical conclusions as the Bayes factor when all models are a priori equally likely: compared to assuming $\theta = .50$, there is some versus about-equal support for ESP.

However, especially in the example at hand, a subjective choice of the prior model probabilities might be more reasonable (Rouder & Morey, 2011; Wagenmakers et al., 2011). Extraordinary claims such as "ESP exists" should be met with a priori skepticism in the form of, for example, $P(\mathcal{M}_1) = .90$ and $P(\mathcal{M}_{2a}) = 1 - .90 = .10$, implying that a priori, the existence of ESP is considered to be rather unlikely. In this case, only if the support in the data (the Bayes factor) in favor of $\mathcal{M}_{2a}$ is extraordinary, will the Bayesian error probability associated with a decision in favor of "ESP exists" be small. For the example at hand, these prior model probabilities would result in posterior model probabilities of $P(\mathcal{M}_{2a} \mid x = 26, n = 40) = .24$ when assuming a subjective prior on θ and $P(\mathcal{M}_{2b} \mid x = 26, n = 40) = .11$ when assuming default prior on θ, meaning that there is not much support for the existence of ESP irrespective of which prior distribution is used.

[2] This is not always the case (Scott & Berger, 2006). In linear regression, for instance, there are default rules for distributing prior mass over rival models that do not lead to a uniform assignment (see Section "Variable Selection and Model Averaging").

Overview

The concepts introduced in the previous section —data, model, prior distribution, prior predictive distribution, marginal likelihood, Bayes factor, prior model probabilities, posterior model probabilities, and Bayesian error probabilities— will reappear in the six sections that follow. Each section discusses applications of the Bayes factor that are relevant and used in psychological research. The topics are Bayesian evaluation of point null, interval, and informative hypotheses, Bayesian evidence synthesis, Bayesian variable selection, and Bayesian evaluation of cognitive models. Each section has been coordinated by one or more experts in the respective field. To give credits where they are due, the names of the coordinators and the contributors are given in the first line of each section. All sections have the same structure: First, the application at hand and the main references are introduced in a subsection titled "What is Bayesian . . . ". Second, an "Illustrative Example" is presented. Finally, the subsection "Further Information" provides an overview of key references, software, and other illustrative applications available at the OSF repository.

Null Hypotheses

Coordinators: Henk Kiers, Jorge Tendeiro. *Contributors:* Qianrao Fu, Felix Schönbrodt.

What is Bayesian Evaluation of Point Null Hypotheses

A point null hypothesis is an *extremely precise* statement involving one or more population parameters. For instance, one way to study the difference between two population group means (say, μ_1 and μ_2) consists of "evaluating" the point hypothesis $\mathcal{H}_0 : \mu_1 - \mu_2 = 0$, which states that the group means are exactly equal to each other. The null hypothesis thus implies, for instance, that the means are *exactly equal* across control and treatment groups, or across drug A and drug B. Bayesian evaluation of the null hypothesis consists of comparing it to an alternative hypothesis ($\mathcal{H}_1$). A straightforward alternative would seem to be "The difference in means is not zero." One could then, given the data, compare the support for each of these hypotheses by means of the Bayes factor.

Whereas the null hypothesis is sufficiently specified for this purpose, the alternative hypothesis must, however, be specified more precisely. One way would be to

specify a particular nonzero value such as $\mathcal{H}_1 : \mu_1 - \mu_2 = 0.8$. [3] Alternatively, the now standard procedure, proposed by Jeffreys (1935), formulates the alternative hypothesis itself in a probabilistic way. An example is "$\mathcal{H}_1 : (\mu_1 - \mu_2)$ is a random value distributed according to the standard normal distribution." In Bayesian evaluation of point null hypotheses, Bayes factors are computed to express relative support of the null hypothesis versus the *particular* alternative hypothesis specified. This is expressed by the ratio of the likelihood of the data under one model versus the likelihood of the data under the other model (see Equation 3).

The Bayes factor is a versatile measure that can be computed for different model comparison situations. The only requirement is that both models are specified sufficiently so as to compute the likelihoods under the respective models. Alternative examples are models on proportions, correlations, or (sets of) regression weights (hence, they also work in situations where ANOVA models are being compared). Apart from their universal usability, they are often used as an improvement on null hypothesis significance testing. The latter can only be used to reject the null hypothesis but not to support it. Using Bayes factors, in contrast, based on the ratio of evidence for both models, one can in fact decide to reject either model, not just one of them. In practice, rejecting $\mathcal{H}_0$ is usually taken as (tentatively) accepting that there is strong evidence of an effect. Analogously, one could come up with the reversed conclusion, that is, that there is little evidence of an effect if the specific $\mathcal{H}_1$ is rejected.

The practical usefulness of Bayes factors involving point null hypotheses is not uncontended. While we have detailed our concerns elsewhere (Tendeiro & Kiers, 2019; Kiers & Tendeiro, 2019), here we will briefly explain some of our main concerns for the *practical* application of null hypothesis Bayesian testing (for a commentary, see van Ravenzwaaij & Wagenmakers, in press). Firstly, we think that one should carefully consider whether a point null hypothesis adequately reflects one's theory. For instance, does one truly expect that a particular population parameter is *exactly* equal to zero?

[3] For ease of notation, the hypotheses in the text refer to the *unstandardized* difference of two means. In practice, priors are usually defined for *standardized* mean differences.

While in some cases one might indeed think so, in other cases one might rather entertain the hypothesis that the population parameter is "close to zero," or not larger than a specified minimally relevant effect size (these hypotheses will be discussed in the next section on Interval Hypotheses). The models compared should obviously align with relevant hypotheses and one should carefully consider whether the point null hypothesis or a less commonly considered interval hypothesis should be employed.

Secondly, when evaluating the Bayes factor one should carefully think what the alternative hypothesis actually means and *what* comparison is actually being evaluated. In case of $\mathcal{H}_1 : \mu_1 - \mu_2 = 0.8$, it is quite clear that a Bayes factor BF_{01} of, for instance, 8.2 expresses that there is more support for the model specifying the effect is 0 in the population than specifying it to be 0.8. However, if the null hypothesis would be compared to $\mathcal{H}_2 : \mu_1 - \mu_2 = 0.3$ we might find $\text{BF}_{02} = 0.10$, and hence conclude that out of the three models, $\mathcal{H}_2$ has most support as a model describing the phenomenon. The current "standard" procedure for Bayesian evaluation of the null hypothesis is more complex, as it specifies $\mathcal{H}_1$ in terms of a probability distribution for $(\mu_1 - \mu_2)$. When using such a default prior (usually, a Cauchy distribution on the standardized mean difference; Rouder et al., 2009), researchers have to specify a scale factor that determines the width of the prior distribution, that is, whether smaller or bigger effect sizes are expected under the alternative hypothesis. So, if the obtained Bayes factor for the comparison of $\mathcal{H}_0$ against a $\mathcal{H}_1$ with a specific scale factor would be $\text{BF}_{01} = 11.2$, it would mean that there is much more support for the null hypothesis than for *this specific* $\mathcal{H}_1$. Hence, again, this does not imply that there is much support *in general* for the null hypothesis, but only *compared to* the particular alternative specified.

Above we mentioned two practical concerns about Bayesian evaluation: the plausibility of the null hypothesis and the particular alternative specified. In our paper (Tendeiro & Kiers, 2019), we mentioned several others, although they are partly consequences of the above. In our opinion, practitioners should know about these concerns in order to use and interpret Bayes factors for null hypothesis testing properly. Concerns about the specification of the alternative hypothesis can be mitigated by careful

sensitivity analysis, in which Bayes factors are computed under various plausible choices of the prior (Kass & Raftery, 1995; Myung & Pitt, 1997; Sinharay & Stern, 2002). Comparing these Bayes factors, one gets insight in to what extent the results depend on the choice of the prior. For example, one can plot the Bayes factor BF_{01} as a function of the scale factor of a default prior (such "Bayes factor robustness checks" are available in `JASP`; Wagenmakers et al., 2018). It is interesting to take into account that not just variations of standard "default" priors could be considered, but also the use of very different priors under the alternative hypothesis could be of interest (e.g., particular non-zero values of the effect size as discussed above; or informative prior distributions, see Gronau, Ly, & Wagenmakers, 2020).

Illustrative Example

Data. Lin et al. (2019) performed an observational study aimed at unravelling patient-level characteristics associated with hospital readmission after earlier heart failure hospitalization. They focused on the distressed personality construct (Type D personality) which is a joint disposition towards negative affectivity and social inhibition.[4]

Model. For the illustration purposes of this section, we focus on part of the results reported in Table 1 of Lin et al. (2019). Lin et al. (2019) assessed the evidence in the data concerning the mean difference between Type D personality and non-Type D personality for outcome variables BMI (body mass index) and NAS (negative affectivity score). Two independent samples t-tests were performed. Table 1 summarizes the results (under "Frequentist"). From a purely significance testing perspective, we conclude that there is not enough evidence in the data to reject the null hypothesis of equal group mean values for BMI ($p = .784$), but there is enough evidence to reject the null hypothesis for NAS ($p < .001$).

Bayes factor, Prior and Posterior Model Probabilities. In classical statistics, it has frequently been mentioned that not rejecting $\mathcal{H}_0$ does not equate to *supporting* it. Bayes factors allow support for either hypothesis under comparison (unlike p-values), but the support for one hypothesis should always be considered as relative to

[4] The data are freely available at `https://doi.org/10.1371/journal.pone.0215726`.

Table 1

Two Independent Samples t-Tests from Lin et al. (2019).

	Sample		Frequentist test			Bayes factor			Posterior of effect size	
	Non-Type D Mean (SD)	Type D Mean (SD)	t	df	p	BF_{01}	$p(\mathcal{H}_0 \mid D)$	$p(\mathcal{H}_1 \mid D)$	Median	95% CI
BMI	26.0 (4.9)	25.7 (4.4)	0.28	214	.784	4.2	.81	.19	0.05	$(-0.35, 0.45)$
NAS	5.5 (4.2)	15.4 (3.5)	-10.93	220	$< .001$	10^{-20}	.00	1.00	-2.36	$(-2.85, -1.87)$

Note. Sample sizes 199 (Non-Type D) and 23 (Type D); there are 6 missing BMI values. BMI = body mass index; NAS = negative affectivity score; $p(\mathcal{H}_i \mid D)$ = posterior probability of $\mathcal{H}_i$ $(i = 0, 1)$ for equal prior model odds; 95% CI = central 95% credible interval when assuming a Cauchy prior distribution with scale $r = 0.707$ under $\mathcal{H}_1$.

the other, so one should always take into account that there may be better hypotheses that are not being tested. In case of comparing well-established rival hypotheses, this is not an issue. However, in the practice of point null hypothesis testing where the alternative is actually very strictly specified while theory merely predicts "there is an effect", it is important to realize that other specifications might actually get more support. In the next two sections, it will be shown that both interval and informative hypotheses can be used to specify substantially meaningful alternative hypotheses.

To illustrate, we compared the support for $\mathcal{H}_0$ against the default $\mathcal{H}_1$ as specified by the `R` package `BayesFactor` (Morey & Rouder, 2018) assuming that the effect size is distributed as a Cauchy distribution with scale factor $r = 0.707$. Table 1 reports the ensuing BF_{01} values of the null versus the alternative hypothesis. As can be seen, in case of BMI, the Bayes factor $\text{BF}_{01} = 4.2$ indicates some support *for* $\mathcal{H}_0$ compared to the particular alternative used. For variable NAS, the Bayes factor $\text{BF}_{01} = 10^{-20}$ indicates decisive support for *this specific* alternative hypothesis of different group means over the hypothesis that there is no difference, so the latter can safely be discarded (while it is still necessary to rely on parameter estimation to learn which particular value would be supported most). Assessing the robustness of the results, a sensitivity analysis shows that scale factors between $r = 0.5$ and $r = 1.5$ lead to values of the Bayes factor BF_{01} ranging from 3.2 to 7.9 for BMI and 10^{-19} to $5 \cdot 10^{-20}$ for NAS.

In our paper (Tendeiro & Kiers, 2019), we argued that Bayesian estimation of a

posterior probability distribution is much more informative than the pairwise comparison of models via Bayes factors. Other scientists strongly value theory building and comparison via testing, and for them the Bayes factor is a versatile tool (van Ravenzwaaij & Wagenmakers, in press). In cases of point null hypothesis testing, where actually the rival model is not that well established, testing can usefully be complemented by an extensive or simple way of estimating effect sizes. When on the basis of the Bayes factor, one is convinced that the null hypothesis should be rejected, indeed still little is known about the actual effect size. Therefore, in such situations, it is always good to inspect credible intervals, or better still the full posterior distribution, to give an indication of the size and directions of the effects (see last panel of Table 1).

Further Information

The interested reader is first of all referred to Rouder et al. (2009) who introduced the Bayesian *t*-test used in the application above. Hoijtink et al. (2019) present a tutorial with respect to the Bayesian evaluation of null hypotheses exemplified using ANOVA. However, these are by far not the only applications of null hypothesis Bayesian testing. Additional examples can be found on the OSF website corresponding to this paper: Testing the equality of within-group variances in the ANOVA context; testing the equality of population means using robust estimates of these means in the ANOVA context; testing the equality of correlation coefficients; and testing the parameters of network models. Additionally, the OSF folder "Null Hypotheses" provides vignettes illustrating Bayesian updating (repeated addition of data and recomputation of the Bayes factor) and sample size determination (loosely spoken, Bayesian power analysis). The main software resources for the evaluation of null hypotheses are the `R` package `BayesFactor` (Morey & Rouder, 2018) which can be used for *t*-tests, various types of ANOVA's, multiple regression and contingency tables; the `R` packages `bain` (Gu et al., 2019) and `BFpack` (Mulder et al., 2019) which can handle the evaluation of null hypotheses in the context of virtually any statistical model; and `JASP` (JASP Team, 2020) which offers an easy-to-use GUI for the Bayesian evaluation of many standard models in psychology (based on packages such as `bain` and `BayesFactor`).

Interval Hypotheses

Coordinators: Don van Ravenzwaaij, Maximilian Linde. *Contributors:* Koen Derks, Zoltan Dienes, Bence Palfi.

Bayesian Evaluation of Superiority, Non-Inferiority, and Equivalence Designs

In clinical psychology, new treatments and medications are developed and evaluated on a regular basis. For instance, in an attempt to combat depression, new antidepressants are developed from time to time.[5] Typically, efficacy of these medications gets assessed through the evaluation of two contrasting hypotheses of which either one or both are *interval hypotheses*.

The most common design type involving an interval hypothesis is the so-called *superiority design*, in which a point null hypothesis (e.g., the medicine has no effect) is contrasted with an alternative hypothesis (e.g., the medicine has some positive effect). In classical frequentist statistical inference, this design is typically analyzed using a one-sided t–test. Depending on whether or not a control condition is present, the test can either be a one-sample test or a two-sample test. If high scores on the dependent variable represent "superiority", the null hypothesis is $\mathcal{H}_0 : \ \delta = 0$ and the alternative hypothesis is $\mathcal{H}_1 : \ \delta > 0$. A schematic depiction of this design is shown in the top row of Figure 2. If low scores on the dependent variable represent "superiority", we have $\mathcal{H}_0 : \ \delta = 0$ and $\mathcal{H}_1 : \ \delta < 0$.

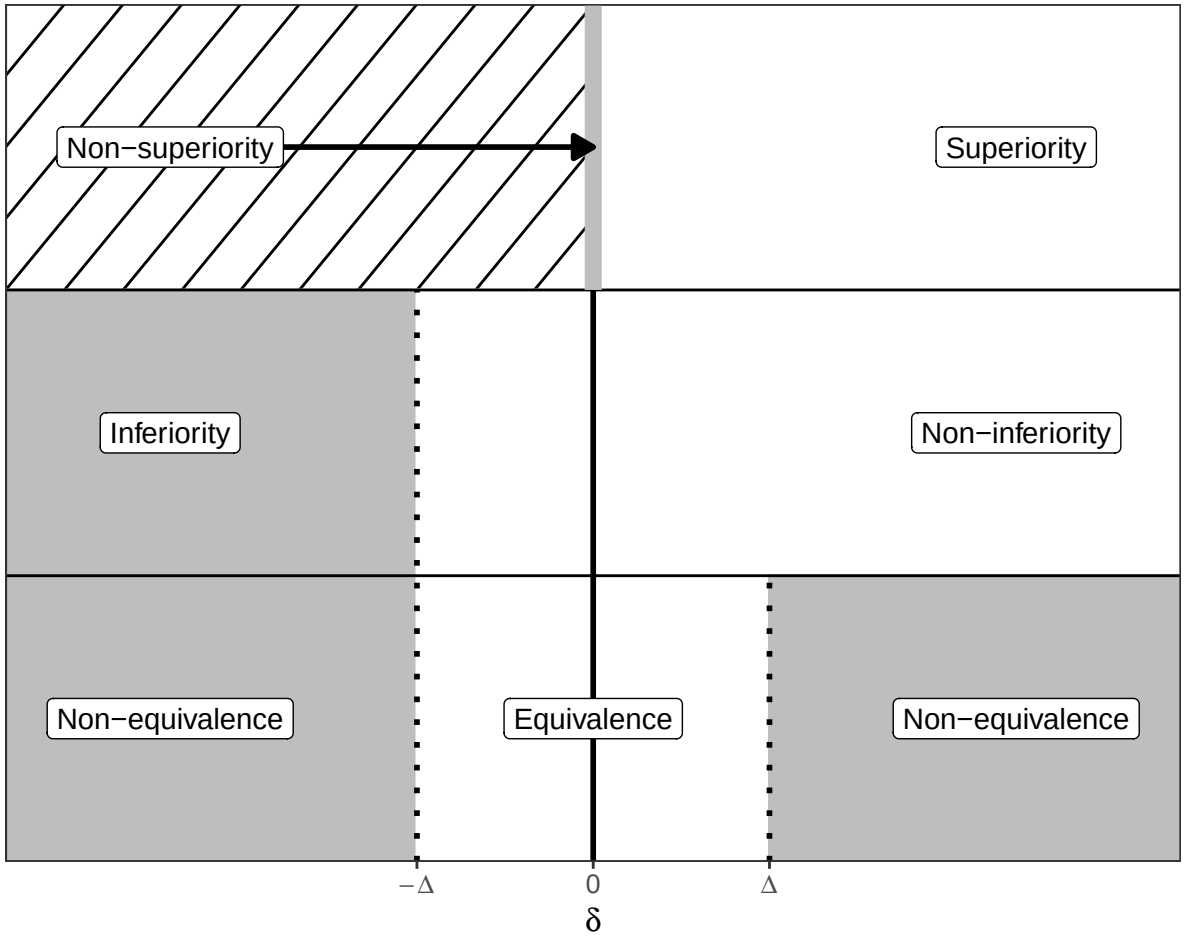
There are, however, cases when the clinician seeks to deviate from the superiority design. New medications are not always developed with the goal in mind of them being *more* effective than existing alternatives. Instead, a new medication might be preferable because different patients might respond favorably to different medications, a new medication might have fewer side effects than an existing one (e.g., Chadwick & Vigabatrin European Monotherapy Study Group, 1999), be cheaper (see, e.g., Kaul & Diamond, 2006, for a discussion), or easier to administer (e.g., Van de Werf et al., 1999).

A second design type in which the contrasting hypotheses are complementary

[5] For a list of recently endorsed psychotropic drugs, see `www.centerwatch.com/directories/1067-fda-approved-drugs/topic/108-psychiatry-psychology`.

Figure 2

Schematic Depiction of the Superiority, Non-Inferiority, and Equivalence Designs.



Note. The equivalence interval is $[-\Delta, +\Delta]$, the non-inferiority margin is Δ. Grey areas represent the null hypotheses whereas white areas represent the alternative hypotheses. In the figure, high scores on the dependent variable represent "superior" or "non-inferior" outcomes. See text for details.

interval hypotheses is the *non-inferiority design* (Senn, 2008; Piaggio et al., 2012). In these designs, the goal is to demonstrate that a new treatment or drug is "not worse" than an existing alternative. To do this, it is necessary to establish a non-inferiority margin Δ prior to observing the data. This margin reflects the amount the new medication can be worse than the existing alternative that is still considered tolerable. Once the data has been observed, classical statistical inference proceeds by conducting a one-sided t–test where, in case high scores on the dependent variable represent "non-inferiority", the null hypothesis is $\mathcal{H}_0 : \ \delta < -\Delta$ and the alternative hypothesis is $\mathcal{H}_1 : \ \delta > -\Delta$. A schematic depiction of this design is shown in the middle row of Figure 2. If, on the other hand, low scores on the dependent variable represent "non-inferiority", the null hypothesis is $\mathcal{H}_0 : \ \delta > \Delta$ and the alternative hypothesis is $\mathcal{H}_1 : \ \delta < \Delta$.

A third design type in which the contrasting hypotheses are complementary interval hypotheses is the *equivalence design*. These designs are valuable when one may wish to find evidence in favor of the null hypothesis, for instance in case of a replication attempt. The null hypothesis is typically defined as a small interval $[-\Delta, +\Delta]$ around zero and gets contrasted to an alternative hypothesis of all values not captured by the null hypothesis (but see the contribution by Palfi & Dienes, available on OSF, for a different specification of the alternative hypothesis). A schematic depiction of this design is shown in the bottom row of Figure 2.

In van Ravenzwaaij et al. (2019), computation of Bayes factors for each of these three designs is discussed (see also Linde & van Ravenzwaaij, 2019a). Calculation of Bayes factors follows the work of Rouder et al. (2009), Morey and Rouder (2011), and Gronau, Ly, and Wagenmakers (2020). Hypotheses are specified in terms of a population effect size δ and the prior for δ is a default Cauchy prior centered on zero. An advantage of such default priors is *model selection consistency*. This means that data generated under one of the models under consideration should lead to a Bayes factor for this model converging to infinity as sample size increases.[6] A specific advantage of the default Cauchy prior over a popular default alternative, the normal distribution, is *information*

[6] This does not hold for overlapping interval hypotheses.

consistency. This means that different sequences of data with a specific, fixed sample size for which likelihood ratios go to infinity should have corresponding Bayes factors that also go to infinity (for more details, see Bayarri et al., 2012; Consonni et al., 2018; van Ravenzwaaij & Ioannidis, 2019). Intuitively, this means that the Bayes factor can become arbitrarily large, meaning that the evidence possibly provided by the data is not limited by an upper bound.

For superiority designs specifically, the default Cauchy prior is truncated at zero, such that negative values of δ have zero density, and the null hypothesis is a point null (i.e., $\delta = 0$). For equivalence designs, both the null hypothesis and the alternative hypothesis are intervals and the default Cauchy prior is truncated accordingly. Similarly, in non-inferiority designs, both hypotheses are interval hypotheses and computation of Bayes factors uses truncated Cauchy priors centered on zero (for alternative applications of Bayes factors based on complementary hypotheses, see also Hoijtink, 2012, Section 1.2.3 and the contribution by Derks, available on OSF). Note that this means that contrary to the other two designs, the center of the prior for the non-inferiority design is not aligned with the null hypothesis value (which is centered on the non-inferiority margin $-\Delta$). In the next section, we present an illustrative example of a Bayes factor calculation for such a non-inferiority design using published data on cognitive behavioral therapy.

Illustrative Example

Andersson et al. (2013) studied whether internet-delivered cognitive behavioral therapy (ICBT) in the treatment of depression symptoms is non-inferior to "standard" cognitive behavioral therapy (CBT). In case of equal efficacy, the authors state that ICBT may be preferable over CBT because (1) ICBT is mainly delivered in text form, allowing patients the possibility of repeating parts of the treatment at will; and (2) some patients may consider the face-to-face aspect of psychotherapy stressful, which might negatively impact the treatment outcomes.

Data. The independent variable is group membership (ICBT versus CBT), sample sizes are $n = 32$ for the ICBT group and $n = 33$ for the CBT group post-treatment. The dependent variable is the post-treatment score on the

Montgomery-Asberg Depression Rating Scale. Mean scores are 13.6 (SD = 9.8) for the ICBT group and 17.1 (SD = 8.0) for the CBT group indicating a descriptively lower level of depression in the ICBT group. Strictly speaking, pre-treatment scores should be taken into account, but as we only have access to the descriptive summary statistics from the published article, we are not able to conduct the appropriate analysis. For purposes of this demonstration, we note that the pre-treatment scores in both groups are nearly identical, so an analysis on post-treatment scores alone will likely provide a reasonable proxy.

Model. We compare the mean depression scores of the two groups using a one-sided t test with $\mathcal{H}_0 : \delta > \Delta$ and $\mathcal{H}_1 : \delta < \Delta$ where Δ denotes the non-inferiority margin. Note that the hypotheses correspond to a situation where low scores on the dependent variable represent "non-inferiority." The non-inferiority margin defined by the authors as a clinically important treatment difference was 2 unstandardized units. For reference, note that the sample was restricted to patients with depression scores between 15 and 35 at baseline (SD= 4.8) and that a remission is defined by a score ≤ 10. The prior for the (standardized) population effect size δ under the alternative is a truncated Cauchy$(0,\ 1/\sqrt{2})$ distribution with the two values denoting the location and scale parameter, respectively.[7] The analysis was carried out using the `baymedr` package (Linde & van Ravenzwaaij, 2021).

Bayes Factor, Prior and Posterior Model Probabilities. The support in the data for $\mathcal{H}_1$ (non-inferiority) is 80 times stronger than the support for its complement $\mathcal{H}_0$ (inferiority), as indicated by a Bayes factor of $\text{BF}_{10} \approx 79.59$. Using equal prior model probabilities for $\mathcal{H}_0$ and $\mathcal{H}_1$, a preference for $\mathcal{H}_1$ comes with a Bayesian error probability of $1/(79.59 + 1) \approx .01$ (cf. Equation 4). The overall conclusion is that $\mathcal{H}_1$ is the preferred hypothesis, meaning that internet-delivered cognitive behavioral therapy is non-inferior to "standard" cognitive behavioral therapy. The code and data used to evaluate the

[7] The scale parameter $r = 1/\sqrt{2} = 0.707$ is chosen for conventional reasons. A sensitivity analysis with a range of scale parameters in the interval $[0.3,\ 1.5]$ was conducted, showing that although the Bayes factor fluctuates ($\min \text{BF}_{10} = 49.13$; $\max \text{BF}_{10} = 105.39$), the overall conclusion does not change as r varies.

hypotheses of interest can be found on the OSF website corresponding to this paper.

Further Information

The interested reader is first of all referred to van Ravenzwaaij et al. (2019). On the OSF website corresponding to this paper, additional examples illustrate the evaluation of interval hypotheses for equivalence testing with complementary hypotheses and for the application of non-inferiority testing in the context of financial statement audits. The main software resources for the evaluation of interval hypotheses are the `R` packages `baymedr` (Linde & van Ravenzwaaij, 2021), `BayesFactor` (Morey & Rouder, 2018), `bain` (Gu et al., 2019; partly also implemented in `JASP` which offers an easy-to-use GUI; JASP Team, 2020), and `BFpack` (Mulder et al., 2019).

Informative Hypotheses

Coordinator: Herbert Hoijtink. *Contributors:* Florian Böing-Messing, Diana Karimova, Rebecca Kuiper, Roger Leenders, Caspar van Lissa, Marlyne Meijerink-Bosman, Mirjam Moerbeek, Joris Mulder, Donald R. Williams.

What is Bayesian Evaluation of Informative Hypotheses?

Questions have been raised with respect to the usefulness of null and alternative hypotheses in psychological research (e.g., Cohen, 1994; Royal, 1997, pp. 79-81; Wainer, 1999). The null hypothesis does often not represent the expectations that researchers have. Consider the experiment presented by Williams and Bargh (2008) to test whether the perception of spatial distance affects people's thoughts and feelings. Participants were presented with a visual grid with a clearly demarcated center. Whereas the "**c**lose" group was instructed to focus on two coordinates located closely to the center, the "**i**ntermediate" and "**f**ar" groups were instructed to focus on coordinates located farther from the center. Subsequently, participants were asked to rate their attachment to friends, family, and home-town on a 7-point Likert scale. For this experiment, the null hypothesis that the mean ratings are identical for the three groups, $\mathcal{H}_0 : \mu_c = \mu_i = \mu_f$, does *not* represent the expectations of Williams and Bargh (2008). Moreover, if $\mathcal{H}_0$ is "rejected" in favor of the unspecific alternative hypothesis $\mathcal{H}_a : \mu_c \neq \mu_i \neq \mu_f$, little is learned because of the "something is going on but I don't know what" character of $\mathcal{H}_a$.

These issues are addressed by the use of informative hypotheses (Hoijtink et al., 2019; van Lissa et al., 2021) which represent researchers' expectations in terms of specific, theory-based relationships of the parameters such as means or regression slopes. The informative hypothesis entertained by Williams and Bargh (2008) clearly is $\mathcal{H}_i : \mu_c > \mu_i > \mu_f$ and their "alternative" hypothesis is its complement $\mathcal{H}_c :$ not $\mathcal{H}_i$ which encompasses all other possible orders of the three group means. The "alternative" hypothesis does not have to be the complement of $\mathcal{H}_i$, it can also be any other informative hypothesis $\mathcal{H}_{i'}$ representing a competing expectation, or the null hypothesis $\mathcal{H}_0$ if it is considered to be a plausible representation of the state of affairs in the population of interest.

Usually, informative hypotheses are formulated using equality and inequality constraints ("equal to", "smaller than", or "larger than") on the parameters of the model at hand. Examples are $\mathcal{H}_1 : (\theta_1 - \theta_2) > (\theta_2 - \theta_3) > (\theta_3 - \theta_4)$ which, if the θs represent the mean of an outcome at four consecutive measurement occasions, states that the difference in means decreases over time; $\mathcal{H}_1 : (\theta_1 > 0.60)$ & ... & $(\theta_{10} > 0.60)$ which, if the θs represent standardized factor loadings, states that each is larger than 0.60; $\mathcal{H}_1 : (\theta_{11} - \theta_{12} > \theta_{21} - \theta_{22})$ & $(\theta_{11} > \theta_{12})$ & $(\theta_{11} > \theta_{21})$ which, if the θs represent the means in a 2×2 analysis of variance, specifies a specific form for the interaction effect; and $\mathcal{H}_1 : \frac{1}{3} \cdot \theta_1 > \frac{1}{9} \cdot \theta_2$ which, if θ_1 and θ_2 represent the means in two conditions in which persons can score either 0 to 3 points or 0 to 9 points, respectively, states that the weighted average score in Condition 1 is larger than that in Condition 2. However, there has also been attention for non-linear constraints (Klugkist et al., 2010). For example, if the θs represent the cell probabilities in a 2×2 contingency table, $(\theta_{11} \cdot \theta_{22})/(\theta_{12} \cdot \theta_{21}) > 1$ specifies that the odds-ratio is larger than 1, that is, there is a positive association between the levels of both categorizations. In summary, informative hypotheses are a means for researchers to formally represent their expectations.

The examples given above concern the evaluation of informative hypotheses in the context of various models such as repeated-measures analysis, factor analysis, ANOVA, and contingency tables. However, informative hypotheses can also be of interest and

Table 2

Data by Lucas (2003) on the Institutionalization of Female Leadership.

Group	N	Mean	SD
(1) Random male leader	30	2.33	1.86
(2) Random female leader	30	1.33	1.15
(3) Skilled male leader	30	3.20	1.79
(4) Skilled female leader	30	2.23	1.45
(5) Institutionalized female leader	30	3.23	1.50

Note. Individual-level data were simulated using the above descriptives presented in Lucas (2003).

evaluated in many other contexts, for example, with respect to the parameters of structural equation models and generalized linear (mixed) models, with respect to the group variances in an ANOVA, and with respect to a matrix of correlations.

Illustrative Example

Data. Lucas (2003) studied the effect of the institutionalization of female leadership on perceived leadership. He randomly assigned 150 persons to five groups: (1) a group with a randomly selected male leader; (2) a group with a randomly selected female leader; (3) a group with a male leader selected on account of proven skills; (4) a group with a female leader selected on account of proven skills; and, (5) a group in which a movie was used to institutionalize female leadership after which a female leader was selected on account of proven skills. For each person in each group, the influence of their leader was measured using the number of times out of 10 in which the person changed their response to match the response of the leader. Descriptive statistics are presented in Table 2.

Model. The basic model is an analysis of variance model in which the influence scores have a group-specific mean μ_g (for $g = 1, ..., 5$) and a common variance σ^2. Competing models are created via the superimposition of informative hypotheses. Here, we test three hypotheses. The first hypothesis $\mathcal{H}_1 : \mu_5 = \mu_3 > \mu_4 > \mu_1 > \mu_2$ states that

Table 3

Evaluation of Informative Hypotheses for Lucas (2003).

Hypotheses	BF_{ic}	$P(\mathcal{H}_i)$	$P(\mathcal{H}_i \mid \text{data})$
$\mathcal{H}_1 : \mu_5 = \mu_3 > \mu_4 > \mu_1 > \mu_2$	70.07	.25	.93
$\mathcal{H}_2 : (\mu_1, \mu_3) > (\mu_2, \mu_4, \mu_5)$	0.11	.25	.00
$\mathcal{H}_3 : (\mu_3, \mu_4, \mu_5) > (\mu_1, \mu_2)$	5.85	.25	.05
$\mathcal{H}_u : \mu_1, \mu_2, \mu_3, \mu_4, \mu_5$		.25	.01

Note. Bayes factors were computed using the `R` package `bain`. BF_{ic} denotes the Bayes factor of the hypothesis in the row versus its complement.

institutionalized female leadership is equally influential as skills-based male leadership, in addition to the combination of male leaders being more influential than female leaders and skills-based leaders being more influential than randomly chosen leaders. The second hypothesis $\mathcal{H}_2 : (\mu_1, \mu_3) > (\mu_2, \mu_4, \mu_5)$ assumes that any type of male leadership is more influential than any type of female leadership. Finally, the third hypothesis $\mathcal{H}_3 : (\mu_3, \mu_4, \mu_5) > (\mu_1, \mu_2)$ states that skills-based leaders are more influential than randomly chosen leaders.

In Bayesian evaluation of informative hypotheses, a default prior distribution is used for the unconstrained hypothesis $\mathcal{H}_u : \mu_1, \mu_2, \mu_3, \mu_4, \mu_5$ in which all group means are estimated as free parameters. The prior distributions for the informative hypotheses $\mathcal{H}_1$, $\mathcal{H}_2$, and $\mathcal{H}_3$, are then obtained by restricting the domain of $\mathcal{H}_u$ (i.e., the space of all possible parameter values) to those parameter values that are in agreement with each of these hypotheses. This implies that, instead of specifying informative prior distributions on the parameters, the user specifies informative hypotheses involving equality and inequality constraints (for details, see Hoijtink et al., 2019).

Bayes factor, Prior, and Posterior Model Probabilities. As can be seen in Table 3, the support in the data for $\mathcal{H}_1$ is about 70 times stronger than the support for its complement "not $\mathcal{H}_1$." There is some support for $\mathcal{H}_3$ over its complement and no support for $\mathcal{H}_2$. To derive posterior model probabilities, the set of competing models also

includes the unconstrained hypothesis $\mathcal{H}_u$ as a fail-safe hypothesis which will be preferred if none of the three substantive hypotheses $\mathcal{H}_1$, $\mathcal{H}_2$, and $\mathcal{H}_3$ are supported by the data. Using equal prior model probabilities, Table 3 shows that the data support $\mathcal{H}_1$ the strongest, and that a preference for $\mathcal{H}_1$ comes with a Bayesian error probability of $1 - .93 = .07$. The overall conclusion is that $\mathcal{H}_1$ is the preferred hypothesis, meaning that institutionalization of female leadership renders an average influence score equal to that of men who are appointed on account of proven skills, that skills-based leaders are more influential within each gender group, and that men are more influential than women within each skill group (this matches with the group means displayed in Table 2). The codes and data used to evaluate the hypotheses with `R` or `JASP` can be found on the OSF website corresponding to this paper.

Further Information

The interested reader is first of all referred to Hoijtink et al. (2019) and van Lissa et al. (2021) who present tutorials with respect to the Bayesian evaluation of informative hypotheses with a focus on ANOVA and structural equation modeling, respectively. On the OSF website corresponding to this paper, additional examples illustrate how to evaluate informative hypotheses about within-group variances in an ANOVA context; about population means using robust estimates in the ANOVA context; about parameters of network models; and about parameters in structural equation models, cluster randomized trials, correlation matrices, and partial correlation networks. Moreover, one example presents an AIC-based alternative for the Bayes factor called the GORIC (Kuiper et al., 2011). The main software resources for the evaluation of informative hypotheses are the `R` packages `bain` (Gu et al., 2019; partly also implemented in `JASP`, JASP Team, 2020) and `BFpack` (Mulder et al., 2019) which have been used for most of the examples provided, and the `R` package `multinomineq` which can be used for the evaluation of linear constraints on the probabilities of contingency tables and multinomial models (Heck & Davis-Stober, 2019).

Bayesian Evidence Synthesis

Coordinators: Irene Klugkist, Mariëlle Zondervan-Zwijnenburg. *Contributors:* Daniel W. Heck, Hidde J. Leplaa.

What is Bayesian Evidence Synthesis

To reach solid scientific conclusions, one cannot rely on a single study. Research needs to be replicated and results from multiple studies need to be aggregated to provide the overall current state of knowledge on a specific theory or hypothesis. The need for replication and aggregation of results from related studies has been widely recognized for decades and attention recently increased after the publication of what is known as the reproducibility project in psychology (Open Science Collaboration, 2015).

The predominant approach for the aggregation of quantitative results from multiple studies is meta-analysis. In meta-analysis, evidence from multiple studies is synthesized assuming either one true population effect (fixed-effect meta-analysis) or a distribution of effects (random-effects meta-analysis). The Bayes factor can be used to test these different assumptions and to obtain a model-averaged estimate of the overall effect size (Gronau et al., 2021). However, irrespective of the statistical approach being used, meta-analysis assumes that each study provides an effect size that is comparable across studies and can be included in a joint model. This limits the use of meta-analysis for the aggregation of results from studies that are theoretically related but methodologically highly diverse.

Bayesian evidence synthesis (BES), on the other hand, does not need any similarity between studies to be synthesized except that they all contain information on the same theoretical concepts. The goal is to use the available information optimally for the evaluation of hypotheses concerning these concepts. In that sense, BES is especially useful for synthesizing results from indirect or conceptual replications, that is, studies that essentially investigate the same central hypothesis but with variations in the design of the study, like, for instance, differences in instruments, statistical models, or context of the study. Results that agree across different methodologies and contexts jointly provide stronger support for the underlying central theory. BES can be used to synthesize

evidence from such diverse studies and is therefore a more flexible tool for aggregation than (Bayesian) meta-analysis.

There are two approaches to BES of original and replication studies. In each approach, the Bayes factor plays a central role. One approach is to use the original study to specify one or more informative hypotheses. These hypotheses can be based on the psychological theory described in the study, on the results of the study, or on both. The informative hypotheses are then evaluated with data of a replication study which provides Bayes factors that express the level of evidence for the hypotheses (Leplaa et al., 2020).

Another approach uses all available studies, original and replication(s), to aggregate the evidence for a predefined set of theoretical hypotheses. This is enabled by the application of Equation 4. Starting with prior model odds of one (i.e., equal probabilities for two hypotheses before any study is included), the BF resulting from the first study provides posterior odds representing the support for the hypotheses in study 1. These odds can subsequently be used as the prior odds for the second study. Updating with the BF resulting from study 2 provides new posterior odds that now contain the aggregated evidence from study 1 and 2. This process can be repeated for each new study and the final posterior odds represent the aggregated evidence over all studies investigating the same theory (Kuiper et al., 2013). Overall, this simply means that the Bayes factors of all studies are multiplied. The resulting overall level of evidence represents to what extent the theory is supported in all studies.

Importantly, the result of the BES procedure outlined above is not the same as the result of sequential updating or sequential testing. In the Bayesian context, for instance, Schönbrodt et al. (2017) describe sequential hypothesis testing in which hypotheses are continuously evaluated with the Bayes factor while increasing the sample size as often as desired. It is important to note that sequential analyses provide the total amount of evidence in *all* observations, irrespective of whether all observations are part of one large study or result from multiple smaller studies. This is not the case in BES. Each study for which evidence is synthesized has its own parameters, and, for each study, the central theoretical hypothesis leads to a separate statistical hypothesis for the parameter(s) of

that specific study. As an example, the central hypothesis "there is no treatment effect" can, for three different studies, lead to the study-specific null hypotheses:

$$\mathcal{H}_0^{(1)} : \mu = 0; \quad \mathcal{H}_0^{(2)} : \eta = 0; \quad \mathcal{H}_0^{(3)} : \xi = 0, \tag{5}$$

with corresponding alternative hypotheses stating that there is an effect. The parameters μ, η, and ξ represent the study-specific operationalizations of the treatment effects. The product of the Bayes factors per study summarizes to what extent the central null hypothesis is supported more than the central alternative *in all studies simultaneously* when assuming distinct and statistically independent parameters per study. Note that in the synthesis itself, the sample sizes of the studies do not play a direct role; only the Bayes factors or posterior model probabilities are involved. However, larger sample sizes often lead to more pronounced evidence for or against hypotheses. If a data set (small or large) results in a posterior model probability close to zero, this hypothesis consequentially has a low probability of being supported by all studies simultaneously.

An example wherein it is very natural to ask if a certain expected relation is supported in each replication is in the context of multiple $N = 1$ studies. Summarizing over all respondents, for instance, using sequential updating, would provide the level of evidence for the hypotheses at the population level. However, if a hypothesis is true at the population level, there is no guarantee that it holds for every person, which may be the question of interest. If sufficient data are available, individual Bayes factors can be computed for each person and they can be combined to provide evidence for the question to what extent a hypothesis is supported for each of the respondents while assuming separate, independent parameters per person (Klaassen et al., 2018).

Also for conceptual replications, the outlined BES-procedure provides the information of interest, that is, to what extent is a central hypothesis supported in all variations of the study. The illustrative example presented next combines the outlined approaches: study data is used to derive informative hypotheses that are evaluated with new data, but also, Bayes factors for each study are aggregated over three different studies.

Illustrative Example

Data. Zondervan-Zwijnenburg et al. (2020) studied the relation between parental age and offspring behavior problems. In this example, we focus on the relation between maternal age and offspring's self-reported externalizing and internalizing behavior problems. Data was available from three Dutch cohort studies: Generation R (Gen-R; Kooijman et al., 2017), the Research on Adolescent Relationships - Young cohort (RADAR-Y; Branje & Meeus, 2018), and the Tracking Adolescents' Individual Lifestyles Survey (TRAILS; Oldehinkel et al., 2015). All these studies had independently collected data on the variables of interest. The questionnaires used to assess these constructs, however, varied between studies.

Model. Regression models were used to consecutively predict internalizing and externalizing behavior problems by maternal age. First, each data set was randomly split into two halves: one to generate hypotheses and one to evaluate them. In each of the cohorts, various regression models (e.g., linear, quadratic, spline) were conducted on the first half of the data for hypothesis generation. The analyses led to the following hypotheses concerning linear (i.e., β_{age}) and quadratic (i.e., β_{age^2}) relations between maternal age and child behavior problems:

$\mathcal{H}_0$: $\beta_{\text{age}} = 0, \beta_{\text{age}^2} = 0$ (i.e., no effect at all).

$\mathcal{H}_1$: $\beta_{\text{age}} < 0, \beta_{\text{age}^2} = 0$ (i.e., linear decrease).

$\mathcal{H}_2$: $\beta_{\text{age}} < 0, \beta_{\text{age}^2} > 0$ (i.e., quadratic, decelerated decrease).

$\mathcal{H}_u$: $\beta_{\text{age}}, \beta_{\text{age}^2}$ (i.e., any linear or quadratic trend).

Bayes factor, Prior and Posterior Model Probabilities. For each cohort, the second half of the data was used to compute Bayes factors for all hypotheses assuming the default priors of the `R` package `bain` (Gu et al., 2019) which use a fraction of information from the data as introduced by O'Hagan (1995). The associated posterior model probabilities were used to compare the four hypotheses. Subsequently, BES was applied to synthesize the evidence as explained in the previous section. The OSF

repository provides `R` code for synthesizing posterior model probabilities across studies.[8]

As can be seen in Table 4, for externalizing problems, Gen-R prefers $\mathcal{H}_2$, RADAR shows substantial support for $\mathcal{H}_0$ and $\mathcal{H}_2$, while TRAILS prefers $\mathcal{H}_0$. The synthesized posterior model probabilities, however, clearly prefer $\mathcal{H}_0$: there is strong evidence against a linear or quadratic relation between maternal age and child self-reported externalizing problems. $\mathcal{H}_0$ is robustly supported by *all cohorts simultaneously*. For internalizing problems, all cohorts prefer $\mathcal{H}_0$, although some probability is allocated to other hypotheses. The synthesized result encompasses clear, robust support for $\mathcal{H}_0$: there is strong evidence against a linear or quadratic relation between maternal age and child self-reported internalizing problems. Overall, the analyses show that increased age at giving birth does not have a detrimental effect on child problem behaviors.

Table 4

Separate and Synthesized Posterior Model Probabilities for Maternal Age in Relation to Offspring Child-Reported Problem Behavior (Zondervan-Zwijnenburg et al., 2020).

Cohort	Externalizing Problems				Internalizing Problems			
	$\mathcal{H}_0$	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{H}_u$	$\mathcal{H}_0$	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{H}_u$
Gen-R	.22	.18	*.49*	.13	*.86*	.09	.04	.01
RADAR-Y	*.43*	.07	.38	.12	*.81*	.16	.02	.01
TRAILS	*.83*	.15	.02	.01	*.93*	.06	.01	.00
Synthesized	**.93**	.02	.04	.00	**1.00**	.00	.00	.00

Note. Numbers in italic font represent the highest posterior model probability per cohort. Numbers in bold font represent the highest synthesized results.

[8] The original data and full analysis cannot be made available as open access. Interested readers are referred to Zondervan-Zwijnenburg et al. (2020) to gain access to the data and the code for computing the posterior model probabilities for each cohort.

Further Information

The seminal paper on Bayesian evidence synthesis using the product of Bayes factors approach is that of Kuiper et al. (2013) who describe a BES application in sociological research. The example above is part of the analyses in Zondervan-Zwijnenburg et al. (2020) synthesizing different cohort studies to evaluate psychological theories. For this work, the regression analyses were conducted in the default environment of `R` (R Core Team, 2019), while Bayes factors and posterior model probabilities were obtained with the `R` package `bain` (Gu et al., 2019). The OSF website corresponding to the present paper provides two more illustrations: the first shows how BES can be used for the evaluation of the question whether a replication study corroborates an original study; and the second illustrates (standard) Bayesian meta-analysis using Bayes factors and model averaging (Gronau et al., 2021) through the `R` package `metaBMA` (Heck et al., 2019) which is also available in `JASP`.

Bayesian Variable Selection and Model Averaging

Coordinators: Alexander Ly, Don van den Bergh. *Contributors:* Paul Bürkner, Xin Gu.

What is Bayesian Variable Selection?

Variable selection is about inferring the importance of covariates in predicting an outcome variable of interest. For instance, the outcome variable Y_i might be the average happiness score of a country i and the covariates $X_{i1}, X_{i2}, \ldots, X_{ip}$ could represent social demographic measurements such as the Wealth (W) and the Life expectancy (Le). The goal is to infer:

Q1 *Which* social demographic features predict Y?

Q2 What is *the extent* with which each feature affects Y?

Once we have identified the features that predict Y, we can develop and test psychological theories and make policy for the better. As there can be multiple features that predict Y, an intervention should first focus on those features that affect happiness the most.

With p covariates, there are 2^p possible combinations of selecting predictors in a regression model to answer the question Q1. For example, if $p = 2$, then the $2^2 = 4$

competing hypotheses are $\mathcal{H}_0$: no covariate predicts Y, $\mathcal{H}_1$: only the first covariate is a predictor, $\mathcal{H}_2$: only the second covariate is a predictor, and $\mathcal{H}_3$: both covariates predict Y. These possible answers to the question Q1 correspond to the following models:

$$\begin{aligned}
&\mathcal{M}_0 : Y_i = \beta_0 + \epsilon_i, && \mathcal{H}_0 : \beta_1 = 0, \beta_2 = 0 \\
&\mathcal{M}_1 : Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i, && \mathcal{H}_1 : \beta_1 \neq 0, \beta_2 = 0 \\
&\mathcal{M}_2 : Y_i = \beta_0 + \beta_2 X_{i2} + \epsilon_i, && \mathcal{H}_2 : \beta_1 = 0, \beta_2 \neq 0 \\
&\mathcal{M}_3 : Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, && \mathcal{H}_3 : \beta_1 \neq 0, \beta_2 \neq 0
\end{aligned} \quad \Longleftrightarrow \quad (6)$$

where β_0 is the intercept, β_p the coefficient of the pth covariate, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ a normally distributed error term. These models correspond to precise null hypotheses displayed in the right column of Equation 6.

There are now two approaches to address Q1: (i) By selecting a single best model/hypothesis from Equation 6, or (ii) by model averaging. To do so in a Bayesian framework, *both* strategies require Bayes factors and prior model probabilities $P(\mathcal{M}_j)$. The Bayes factor given data in favor of $\mathcal{M}_j$ over the null model $\mathcal{M}_0$ is denoted by BF_{j0}. Once we observed data, we can rearrange Equation 4 to update the prior model probabilities and obtain posterior model probabilities,

$$P(\mathcal{M}_j \mid \text{data}) = \frac{\text{BF}_{j0}\, P(\mathcal{M}_j)}{\sum_{j'=0}^{2^p-1} \text{BF}_{j'0}\, P(\mathcal{M}_{j'})}, \quad (7)$$

where the Bayes factor of the null model to itself is always one (i.e., $\text{BF}_{00} = 1$).

One strategy to select a single best model is to choose the model for which the posterior model probability $P(\mathcal{M}_j \mid \text{data})$ is highest. Note that for $p = 2$ such that $P(\mathcal{M}_0 \mid \text{data}) = .24$, $P(\mathcal{M}_1 \mid \text{data}) = .26$, $P(\mathcal{M}_2 \mid \text{data}) = .25$, $P(\mathcal{M}_3 \mid \text{data}) = .25$, the posterior probability of the best model being correct is only 26%, thus the Bayesian error probability is 74%.

An alternative strategy is model averaging. In that case, the prior and posterior model probabilities in Equation 7 serve as weights to evaluate the importance of each variable across the models. To simplify matters, suppose that each model is equally probable a priori, that is, $P(\mathcal{M}_j) = .25$ for each of the four models when $p = 2$. Table 5

shows all the models for $p = 2$ with the prior model probabilities and the bottom line shows the prior importance of each variable. The prior inclusion probability of the first covariate is calculated by simply summing the prior model probabilities of all models in which it is present. This is indicated by a check mark in the column β_1, and thus, $P(\text{inclusion of } \beta_1) = P(\mathcal{M}_1) + P(\mathcal{M}_3) = .25 + .25$. Similarly, the posterior inclusion probability of β_1 is $P(\text{inclusion of } \beta_1 \mid \text{data}) = P(\mathcal{M}_1 \mid \text{data}) + P(\mathcal{M}_3 \mid \text{data})$.

Model averaging can also be used to estimate the extent to which each covariate affects the outcome variable Y across all of the considered models. For this purpose, we average the 2^p posterior distributions of β_p within model $\mathcal{M}_j$ weighted by the corresponding posterior model probabilities. The resulting model-averaged posterior distribution of β_p combines the uncertainty of testing between the models with the uncertainty of estimating the parameters within each model. The following section elaborates on this procedure with a concrete example.

Table 5

Model Space and Prior Model Probability for $p = 2$ Predictors.

Model $\mathcal{M}_j$	$P(\mathcal{M}_j)$	β_0	β_1	β_2
$\mathcal{M}_0$	.25	✓	—	—
$\mathcal{M}_1$	.25	✓	✓	—
$\mathcal{M}_2$	.25	✓	—	✓
$\mathcal{M}_3$	.25	✓	✓	✓
Inclusion probability		1.00	.50	.50

Note. Tick marks indicate which regression parameters β_p are included in model $\mathcal{M}_j$. β_0 is the intercept included in all models.

Illustrative Example

Data. We use `JASP` (JASP Team, 2020) to put theory into practice and consider the problem of predicting the average happiness score by country. Each year Gallup, a research-based consulting company, conducts a series of interviews with inhabitants of

multiple countries. During those interviews, the interviewees' happiness is measured using the Cantril Self-Anchoring Striving Scale (Glatzer and Gulyas, 2014), and questions are asked about several other demographic factors.[9] Altogether, a score for Wealth (W), Life expectancy (Le), Social support (Ss), Freedom (F), Generosity (G), and Perception of Corruption (PoC) is obtained per country, which is subsequently related to the average happiness of the interviewees from said country.

Model. Since all variables in the data set are approximately continuous, the obvious model choice is linear regression. The data set contains six predictors, thus there are $2^6 = 64$ models to consider.

Bayes factor, Prior and Posterior Model Probabilities. Since there are 64 models, choosing the prior model probabilities is a delicate matter. An intuitive choice is that all models are equally likely a priori, that is, $P(\mathcal{M}_j) = 1/64$. However, a drawback is that the prior distribution on the *number* k of active covariates is then not uniform.

To illustrate this, consider the previous example with $p = 2$ predictors and uniform prior model probabilities $P(\mathcal{M}_j) = .25$ (see Table 5). The prior probability for including $k = 1$ predictor is thus $P(\mathcal{M}_1) + P(\mathcal{M}_2) = .50$. However, the prior probability for including $k = 0$ predictors is only .25, and likewise, the prior probability for including $k = 2$ predictors is also .25. Similarly, for $p = 6$ covariates, there is an a priori bias to include $k = 3$ or $k = 4$ predictors (with prior probabilities of .312 and .234, respectively) as opposed to including $k = 0$ or $k = 6$ predictors (with a prior probability of .016 each). This shows that uniform model probabilities result in an increased prior probability of selecting about half of the predictors. To circumvent this problem, we use a beta-binomial prior on the model space so that the prior probability of including any number k of predictors is constant and equals $1/(p + 1)$.[10] For instance, if $p = 6$, then the set of models with $k = 5$ predictors gets a prior model probability of $1/7 \approx .143$. As 6 out

[9] Details about the data collection can be found on Gallup's website: `http://www.gallup.com/178667/gallup-world-poll-work.aspx`. Gallup's own report is available at `http://worldhappiness.report/ed/2018/`. A `JASP` file including the data and the Bayesian analysis is available on the OSF repository.

[10] More precisely, we assume a beta-binomial distribution with parameters $\alpha = 1$ and $\beta = 1$.

of the 64 models have $k = 5$ active covariates, each of these models gets a prior model probability of $.143/6 \approx .024$. The simulation studies in Scott and Berger (2006) show that model averaging with these beta-binomial prior model probabilities results in a good control of the false discovery rate.

Table 6 shows the 10 out of 64 models with the highest posterior model probability. The fourth column shows the Bayes factor BF_{j1} for each model relative to the best model (i.e., the model in the first row). It is important to note that this Bayes factor provides a pairwise comparison. For instance, the Bayes factor in the second row implies that the data are about 4.7 times more likely under the second than the first model. Based on the pairwise comparison only, the second model would be preferred over the first model. However, as there are 64 models, we have multiple comparisons which are accounted for by the beta-binomial prior on the models. Hence, based on both the prior model probabilities and the evidence in the data in terms of Bayes factors, there is a slight preference for the model in the first compared to the model in the second row (i.e., the posterior model probabilities are .296 versus .232, respectively).

If we were to select a single model, one choice would be the model with all six predictors. Note that such choice is not without substantial doubt, as the Bayesian error probability is then $1 - .296 \approx 70\%$. As an alternative strategy, we may consider the model-averaged results. As the model comparison table grows rapidly with the number of predictors, it can be cumbersome to draw model-averaged inference from Table 6 directly by summing over the posterior model probabilities of the models that include the covariate of interest. `JASP` performs these computations under the hood. The results are shown in Table 7 and can be used to address the question Q2: What is *the extent* with which each feature affects Y? Here, the columns show the prior and posterior inclusion probability, the inclusion Bayes factor, the model-averaged posterior mean and standard deviation of the coefficient β_p, and a 95% model-averaged credible interval. In particular, the posterior mean of 0.309 can be used as a best guess for the magnitude of the coefficient of Wealth, and the 95% credible interval $[0.139, 0.529]$ serves as a model-averaged measure of uncertainty regarding this estimate.

Table 6

Comparison of Regression Models for Predicting the Average Happiness Score per Country.

Model $\mathcal{M}_j$	$P(\mathcal{M}_j)$	$P(\mathcal{M}_j \mid \text{data})$	BF_{j1}	R^2
1. W + Le + Ss + F + Ge + Poc	.143	.296	1.000	.802
2. W + Le + Ss + F + Poc	.024	.232	4.701	.799
3. W + Le + Ss + F	.010	.203	10.293	.794
4. W + Le + Ss + F + Ge	.024	.195	3.957	.799
5. W + Ss + F	.007	.027	1.798	.781
6. W + Ss + F + Poc	.010	.015	0.757	.786
7. W + Ss + F + Ge	.010	.014	0.702	.786
8. W + Ss + F + Ge + Poc	.024	.008	0.163	.789
9. Le + Ss + F + Ge	.010	.002	0.120	.780
10. Le + Ss + F + Poc	.010	.002	0.110	.780

Note. The covariates are Wealth (W), Life expectancy (Le), Social support (Ss), Freedom (F), Generosity (G), and Perception of Corruption (PoC).

In this example, the results from picking a single best model versus the model-averaged results are in agreement if for the latter all covariates with a inclusion Bayes factor larger than one are included. However, Table 7 shows that the data provide little evidence to actually include the covariates Generosity and Perception of Corruption.

Further Information

A general description of Bayesian testing, model averaging, and reporting is provided by van Doorn et al. (2021), Jeffreys (1939), Ly et al. (2016a, 2016b), and Ly et al. (2020), whereas a more specialized account of Bayesian linear regression is given by Bayarri et al., (2012), Li and Clyde (2018), Liang et al., (2008), Rouder and Morey (2012), and Zellner and Siow (1980). For more on Bayesian model averaging we refer to Hinne et al. (2020), Hoeting et al. (1999), Scott and Berger (2006), van den Bergh et al. (2019, 2020, in press), and Wasserman (2000). Key software implementations in `R` are the

Table 7

Model Averaging and Posterior Summaries of Regression Coefficients.

				Posterior		95% CI	
Coefficient	$P(\text{incl})$	$P(\text{incl} \mid \text{data})$	$\text{BF}_{\text{inclusion}}$	Mean	SD	Lower	Upper
Intercept	1.000	1.000	1.000	5.346	0.044	5.259	5.432
Wealth	.500	.992	118.555	0.309	0.099	0.123	0.524
Life expectancy	.500	.936	14.726	0.032	0.014	0.000	0.054
Social support	.500	.999	753.675	2.337	0.568	1.218	3.452
Freedom	.500	.999	1868.243	1.715	0.409	0.914	2.529
Generosity	.500	.518	1.076	0.231	0.319	−0.085	0.980
Perception of Corruption	.500	.556	1.252	−0.250	0.313	−0.954	0.050

Note. $P(\text{incl})$ and $P(\text{incl} \mid \text{data})$ show the prior and posterior inclusion probability of each predictor, respectively. The inclusion Bayes factor $\text{BF}_{\text{inclusion}}$ quantifies the change from prior to posterior inclusion probability. The last four columns refer to the model-averaged posterior distribution.

`BAS` package (Clyde, 2018; Clyde et al., 2011), which includes model averaging, and the `BayesFactor` package (Morey & Rouder, 2018), which also includes methods for ANOVA models. Both these packages are integrated in `JASP` (JASP Team, 2020), a free, open-source, and easy-to-use software package which was used to run the analysis above. The OSF repository contains additional illustrations on Bayesian one-sided variable selection and on Bayesian evaluation of hypotheses in multilevel models.

Bayesian Evaluation of Cognitive Models

Coordinator: Daniel W. Heck. *Contributors:* Udo Boehm, Michael D. Lee, Wolf Vanpaemel.

What is Bayesian Evaluation of Cognitive Models?

Whereas many theories in psychology are stated in verbal form, cognitive modeling aims at developing and testing mathematical and statistical accounts of fundamental capacities such as memory, categorization, decision making, and other aspects of behavior and cognition (Farrell & Lewandowsky, 2018; Lee & Wagenmakers, 2013). By providing

formalized representations of psychological theories, cognitive models do not only describe and explain established psychological phenomena, but also facilitate the derivation of novel and precise predictions that can be tested empirically (Kellen, 2019; Suppes, 1966; Vanpaemel, 2020). In contrast to standard statistical models such as regression, cognitive models include theoretically meaningful parameters that have a direct psychological interpretation in terms of latent processes. Once a cognitive model has been empirically established as an appropriate account of the process under consideration, it allows the measurement of latent psychological processes (Batchelder, 1998).

The application of Bayes factors in cognitive modeling serves three major goals. First, the canonical use of Bayes factors in cognitive modeling is to test alternative substantive theories against each other (Lee & Wagenmakers, 2013). For instance, in recognition and working memory, it is of interest to test threshold models which assume a discrete number of memory states against signal detection models which assume a latent continuous memory strength (e.g., Kellen & Klauer, 2014). To test competing theories, each account needs to be cast in the precise mathematical form of a cognitive model by specifying an assumed mechanism by which latent processes generate observable behavior. The Bayes factor then quantifies the relative evidence for each of the possibly non-nested models, thus providing a direct test of the psychological theories under scrutiny (Myung et al., 2000). Conceptually, the Bayes factor selects the model with the highest average predictive accuracy for the observed behavior. Thereby, it achieves an optimal trade-off between the fit of a model and its complexity, namely, its ability to fit any data that can possibly be observed (Myung & Pitt, 1997).

A second major goal of using Bayes factors in cognitive modeling are hypothesis tests on the inferred model parameters. Both for exploratory model development and for confirmatory model validation, researchers are often interested in testing the effect of experimental manipulations or continuous covariates on latent processes. Because parameters in cognitive models have a substantive interpretation, hypotheses can be directly specified on the model parameters. For instance, researchers can test whether the parameter measuring memory strength increases for study items that are presented

multiple times, or for individuals with higher intelligence. To address such research questions, the Bayes factor can be used to test whether the relevant parameters are affected by an experimental manipulation or associated to a continuous covariate, respectively (Boehm et al., 2018; Wagenmakers et al., 2010). Statistically, this approach is closely related to the standard application of Bayes factors to null hypothesis testing of directly-observable variables. The key difference is that in cognitive modeling, substantively meaningful parameters are being tested rather than descriptive statistics of the raw behavioral data (e.g., accuracy or mean response time) which may not represent valid and process-pure measures of the latent constructs.

Hypothesis tests on model parameters are also crucial for establishing the construct validity of a cognitive model. For a cognitive model to be valid, it must be shown that there is a one-to-one mapping between the model parameters and the corresponding latent cognitive constructs.[11] The validity of the parameters is usually established by selective influence (Batchelder & Riefer, 1999), that is, by testing whether a theoretically motivated manipulation influences only the relevant parameter but not the remaining parameters. For instance, in memory models, the repeated presentation of study items should result in an increase of the corresponding memory-strength parameter but not affect any of the response-bias parameters. The Bayes factor is ideally suited to quantify the evidence for such a specific pattern of selective influence by comparing a restricted version of the cognitive model against an unconstrained, more general version (e.g., Bott et al., 2020; Heathcote et al., 2015; see also Section “Informative Hypotheses”).

As a third major goal, Bayes factors are often used in cognitive modeling for the classification of participants. Instead of assuming that all individuals are best described by a single model, mixture models assume that different individuals are best described by different models. For instance, in choice experiments, it is of interest to test whether participants rely on different decision strategies (e.g., Lee, 2016). More generally, mixture models assume that behavioral data are a combination of two or more qualitatively

[11] Another necessary, but not sufficient condition for a model to be valid, is that it must fit the observed data (see Roberts & Pashler, 2000).

different cognitive processes. By specifying how exactly each latent process generates observable outcomes, latent mixture models have the ability to infer which parts of the data belong to which of these different processes. Statistically speaking, when implementing a set of cognitive models, inferences about latent group membership are usually made using discrete parameters that act as indicator variables. Treating each of the alternative models as mixture components for the data, the posterior of each indicator parameter has a natural interpretation in terms of posterior odds between the models. By factoring out the prior, the indicator parameters can also be interpreted in terms of Bayes factors between the models.

The Bayes factor is uniquely suited for evaluating cognitive models for two reasons. First, cognitive models usually do not only differ in the number of free parameters but also in the functional form of how the parameters are formally linked to generate predictions. For instance, many models of recognition memory assume two separate parameters for memory strength and response bias, even though each model makes fundamentally different assumptions on how these two parameters jointly determine the accuracy of responses. Unlike popular model-selection indices such as AIC and BIC, the Bayes factor takes this functional flexibility of models into account: If two models fit equally well, the Bayes factor will prefer the more parsimonious model that makes the more precise predictions (Myung & Pitt, 1997). Second, in the context of cognitive modeling, a strong argument can be made for using informative, subjective prior distributions instead of default, objective priors (Lee & Vanpaemel, 2018; Vanpaemel, 2010). Since the parameters of a cognitive model have a direct theoretical interpretation, it is much easier to specify an informed prior distribution compared to standard, off-the-shelf statistical models that are applied across many different contexts. Hence, researchers can express a theoretical commitment by assigning higher prior probability to parameter values that are plausible given the specific theory and application.

Illustrative Example: Multinomial Models of the Weapon Identification Task

Data. Rivers (2017) studied the effect of stereotypes on the identification of stereotype-congruent and stereotype-incongruent objects in a weapon-identification task.

Participants ($N = 82$) were presented with images on which they had to identify a target (tool or weapon). In each trial, the image was preceded by one of three primes (either an image of a white or black face, or a neutral outline of a face). The analysis below focuses on the frequencies of correct (+) and incorrect (−) responses in each of the 3 (prime: white, black, neutral) × 2 (target: tool, weapon) within-subjects conditions (in total, 216 trials per individual). Moreover, Rivers (2017) implemented a between-subject manipulation by assigning individuals randomly to two conditions that differed in the response deadline (either 1,000ms or 500ms).

Models. We illustrate the application of Bayes factors in case of the comparison of competing theoretical accounts which are instantiated by different multinomial processing tree (MPT) models, a specific class of cognitive models (Batchelder & Riefer, 1999). MPT models are often used in research on memory, reasoning, decision making, or social cognition to disentangle different latent processes that are assumed to contribute jointly to observable behavior (Erdfelder et al., 2009). Here, we test two non-nested models for the classical process-dissociation procedure (Jacoby, 1991) which are commonly used in social cognition to disentangle automatic and controlled processes (Bishara & Payne, 2009; Buchner et al., 1995). In particular, two substantive MPT models represent different theoretical conceptions on whether automatic processes act conditional on a failure of controlled processes or other way round. The third model is the saturated (unconstrained) model which allows researchers to test whether the two substantive models provide a satisfactory account of the data in absolute terms.

The process-dissociation model with guessing (PD) in Figure 3 assumes that objects are correctly identified with probability C due to controlled processing of information. If the controlled process fails with probability $(1 - C)$, automatic stereotype activation determines responses with probability A. In this case, an individual responds "weapon" when primed with a black face, but "tool" when primed with a white face. If automatic stereotype activation fails with probability $(1 - A)$, the person guesses "tool" or "weapon" with probabilities B and $(1 - B)$, respectively. Figure 3 also shows a second, alternative MPT model called "Stroop with guessing" (Stroop) in which the order of

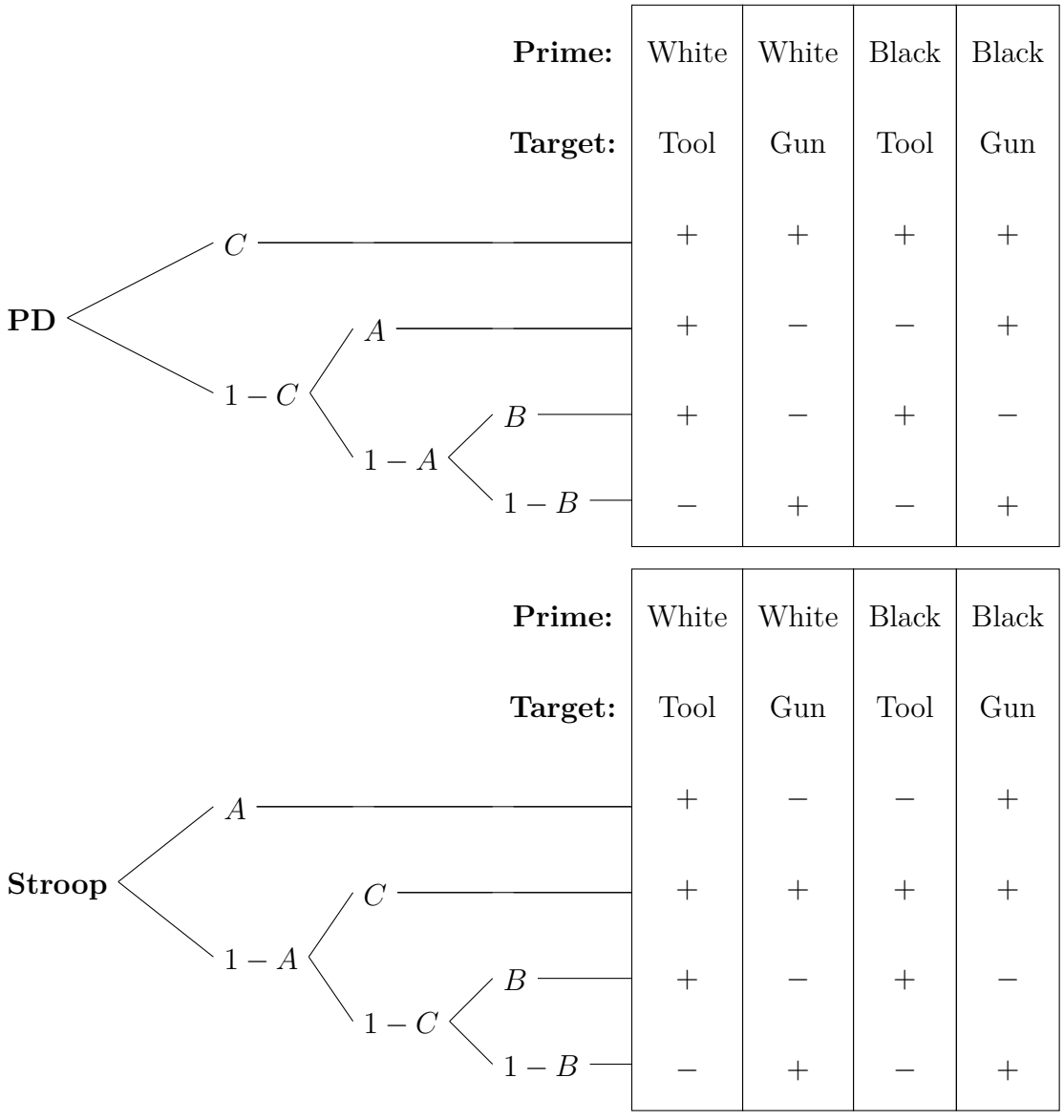
controlled and automatic processes is switched. Here, the automatic process may first succeed with probability A. Only conditional on failure with probability $(1 - A)$, the controlled process determines responses with probability C.

For the present analysis, we extended the standard models shown in Figure 3 to account for responses to neutral primes as included in the study by Rivers (2017). For neutral primes, automatic activation of stereotypes (parameter A) is irrelevant whereas controlled and guessing processes (C and B, respectively) are still involved in identifying an object as a tool or a weapon. It follows that the predicted accuracy in the neutral condition is $C + (1 - C)B$ for tools and $C + (1 - C)(1 - B)$ for weapons. For the PD and Stroop model, we assumed independent uniform distributions for the parameters A and C and an informed prior for B which assigns more probability mass to guessing probabilities around 50% (i.e., a symmetric beta distribution with shape parameters $\alpha = \beta = 3$). For the saturated model, we assumed independent uniform priors for the six response probabilities.

Bayes factor, Prior and Posterior Model Probabilities. We computed Bayes factors and posterior model probabilities based on the aggregated response frequencies in each response-deadline condition using the `R` package `TreeBUGS` (Heck et al., 2018). The two substantive models fitted the data well in both experimental conditions as indicated by the Bayes factor which clearly favored both the PD model and the Stroop model over the saturated, unconstrained model (all BF > 530). For a response deadline of 500ms, the Bayes factor indicated moderate evidence for the Stroop over the PD model, $\text{BF}_{\text{Stroop,PD}} \approx 3.4$, whereas for a response deadline of 1,000ms, the Bayes factor indicated moderate evidence for the PD over the Stroop model, $\text{BF}_{\text{PD,Stroop}} \approx 8.1$. We obtained similar values for the Bayes factor when assuming a uniform instead of an informative prior distribution on the guessing parameter B. These results suggest that automatic processes dominate controlled processes under a short response deadline whereas controlled processes dominate automatic processes under a long response deadline. However, the Bayesian evaluation also shows that additional data are required to obtain more convincing evidence for the comparison between the PD and the Stroop

Figure 3

The Process-Dissociation Model with Guessing (PD) and the Stroop Model with Guessing (Stroop) for the Weapon Identification Task (Bishara & Payne, 2009).



Note. Each branch of the probability tree either leads to a correct (+) or incorrect (−) response. The parameters refer to the probabilities that one of the hypothesized processes succeeds (C = controlled; A = automatic; B = guessing "Tool").

model, for instance, by including additional experimental conditions that increase the diagnosticity of the data (Heck & Erdfelder, 2019).

Key References, Software, and Other Applications.

Lee and Wagenmakers (2013) and Farrell and Lewandowsky (2018) provide general introductions to Bayesian cognitive modeling whereas Vandekerckhove et al. (2015) give a primer on Bayes factors for cognitive modeling with a focus on model flexibility. In the OSF repository corresponding to the present paper, we provide a short overview of recent applications of the Bayes factor in cognitive modeling. Many cognitive models are tailored to specific theories and applications and can be fitted with `JAGS` (Plummer, 2003) or `Stan` (Stan Development Team, 2020) by drawing samples from the posterior distribution via Markov chain Monte Carlo (MCMC). However, Bayes factors are often not easy to compute from MCMC samples except for some nested models, in which case one can often apply the Savage-Dickey density ratio (Wagenmakers et al., 2010). A more general method often applied for mixture modeling and classification of individuals is the product-space approach (Lodewyckx et al., 2011). To compare only a few cognitive models, it is often more efficient to compute the marginal likelihood of each model directly by means of the `R` package `bridgesampling` (Gronau, Singmann, & Wagenmakers, 2020). Computing Bayes factors for specific families of cognitive models is facilitated by `R` packages such as `TreeBUGS` for MPT models (Heck et al., 2018) or `DMC` for evidence accumulation models (Heathcote et al., 2018).

Discussion

The present paper reviewed the wide range of applications of the Bayes factor in psychological research. Each section outlined a different type of research question, explained the corresponding models and Bayesian analysis, and illustrated how to apply the Bayes factor in practice by means of user-friendly software.

Testing Psychological Theories with the Bayes Factor

The various substantive applications showed that there is nothing like *the* Bayes factor in terms of a numeric value that has a universal interpretation independent of the specific statistical models being tested (in contrast to the p-value which is always

asymptotically uniformly distributed under the null hypothesis). Instead, the Bayes factor provides a method for quantifying the relative evidence for two competing hypotheses that are both instantiated by specific statistical models with prior distributions on the parameters. This general approach can be used to address many specific, theoretically relevant research questions, as illustrated in the different sections: whether a randomized experiment has an effect or not (null hypothesis), whether an effect is inside or outside a range of negligible effect sizes (interval hypothesis), whether a set of means follows a specific order (informative hypothesis), whether a set of studies jointly corroborate a theoretical claim (evidence synthesis), which variables are most relevant for prediction (variable selection), and which model provides the best account of latent processes (cognitive modeling). This wide range of applications also shows that researchers in psychology are regularly interested in quantifying the evidence for qualitatively different hypotheses (Dienes, 2021; Haaf et al., 2019).

To compute a Bayes factor, theoretical positions must be translated into statistical models with suitable priors on the parameters. The prior distribution formalizes one's expectations about which values of the parameters in each model are more or less plausible. This is required to make quantitative predictions for future data by means of the prior predictive distribution (cf. second row of Figure 1). Essentially, each combination of model and prior serves a precise and specific instantiation of a theoretical position (Vanpaemel, 2020). The Bayes factor then provides a quantitative measure of the relative evidence for different positions by comparing the predictive accuracy of two model-prior combinations (Morey et al., 2016). The better the predictions of a model for the observed data, the more it is supported by the Bayes factor (Jeffreys, 1939). Thereby, the Bayes factor naturally lends itself to a confirmatory test of competing theories that are a priori specified. The Bayes factor may also be used for exploratory purposes, for instance, when selecting variables in multiple regression or when testing all coefficients of a large correlation matrix. In such cases, the statistical models and priors for comparison must still be specified *before* inspecting the data which —in the absence of any theoretical expectations— requires the reliance on default priors.

Irrespective which types of models and priors are being compared, the Bayes factor allows one to distinguish whether there is evidence for one of the models (i.e., $BF_{01} \gg 1$ or $BF_{01} \ll 1$) or whether there is absence of evidence ($BF_{01} \approx 1$) meaning that the data are not informative for the comparison of interest (Keysers et al., 2020). In the context of null hypothesis testing, this implies that the Bayes factor can be used to quantify the evidence in favor of the null hypothesis compared to a specific alternative (Wagenmakers, 2007). In doing so, one should keep in mind that this comparison depends on the range of effect sizes assumed under the alternative hypothesis (i.e., on the prior distribution; Rouder et al., 2016).[12] Moreover, to obtain a large Bayes factor in favor of the absence of an effect, studies still require sufficiently large sample sizes.

Opportunities and Pitfalls of the Bayes Factor

The application of the Bayes factor in psychological research offers many opportunities but is also prone to possible pitfalls practitioners should be aware of. First, the specification of prior distributions offers the opportunity to specify precise theoretical assumptions and to incorporate expert knowledge. In the context of (null and interval) hypothesis testing, it is important to specify a prior distribution of plausible effect sizes under the alternative hypothesis (Dienes, 2021; Tendeiro & Kiers, 2019). When using default prior distributions, one can usually specify the scale of the expected effect size (Rouder et al., 2009; but see Hoijtink et al., 2019, for an alternative approach). However, the necessity to specify prior distributions has often been seen as a pitfall. In fact, the use of very vague, implausible priors may lead to erratic results (e.g., Jeffreys, 1935; Lindley, 1957), implying that researchers need to think about the specification of adequate models and (subjective or default) prior distributions for the substantive question of interest (Rouder et al., 2016).

Second, the Bayes factor allows researchers to quantify the relative evidence for different theoretical positions. This is achieved by assessing how well the different hypotheses predict the specific data at hand. However, it is a pitfall to interpret the

[12] In practice, Bayes factors with default priors often lead to similar conclusions when using a range of plausible choices for the scale of the prior (van Ravenzwaaij & Wagenmakers, in press).

Bayes factor as the relative *plausibility* of two hypotheses, that is, as posterior model odds (Tendeiro & Kiers, 2019). If researchers want to assess the strength of belief in different hypotheses, it is necessary to specify prior model probabilities and to interpret the corresponding posterior model probabilities or odds (e.g., Rouder & Morey, 2011). Moreover, if one wants to make decisions, then posterior beliefs need to be combined with a utility function.

Third, the Bayes factor provides the opportunity for a more intuitive interpretation of scientific evidence. To facilitate communication, observed Bayes factors can be described by labels such as "anecdotal" ($1 < \text{BF}_{10} < 3$), "moderate" ($3 < \text{BF}_{10} < 10$), or "strong" ($10 < \text{BF}_{10} < 30$) evidence (Wagenmakers et al., 2018). However, fixed thresholds for different levels of evidence should be used with great care. The Bayes factor is a continuous measure of evidence, implying that $\text{BF}_{10} = 2.9$ and $\text{BF}_{10} = 3.1$ do not provide qualitatively different levels of evidence (Tendeiro & Kiers, 2019; van Doorn et al., 2021). Moreover, whether the amount of evidence provided by the data is convincing depends on the prior plausibility of the different hypotheses in a specific context (cf. previous point). At worst, fixed thresholds for Bayes factors could encourage authors, reviewers, and editors to judge the relevance of empirical results based on an (implicit or explicit) "minimum level of evidence" (e.g., $\text{BF}_{10} > 3$). In turn, this could foster questionable research practices and publication bias.

Forth, a pragmatic benefit of using the Bayes factor concerns the possibility of optional stopping in data collection when sufficient evidence has been obtained (Rouder, 2014; Schönbrodt et al., 2017; for an example, see the vignette by Schönbrodt available on OSF). Since the Bayes factor adheres to the likelihood principle (Berger & Wolpert, 1988), the specific sampling plan has no impact on the interpretation of the Bayes factor as a measure of relative evidence for two models. Hence, researchers can assess the current value of the Bayes factor at any time to decide whether or not to continue sampling. This implies that data collection can be rendered more efficient in a range of different scenarios: One may collect a minimum amount of trials within a person, a minimum amount of participants within a study, or a minimum amount of studies within

a meta-analysis. However, it is a pitfall to start data collection without considering the expected sample size required for answering a specific research question. Just because the Bayes factor *can* indicate convincing evidence based on only a few participants, the chances of doing so could be very low. As a remedy, tools have been developed to judge the probability of obtaining sufficient evidence and to estimate the expected sample size when using optional stopping (Fu et al., 2021; Stefan et al., 2019; see also the vignette by Fu available on OSF). Moreover, the Bayes factor can be used to improve the efficiency of data collection via adaptive design optimization, that is, by selecting the most diagnostic stimuli for model comparison while the experiment is running (Myung et al., 2013).

Fifth, the Bayes factor offers the opportunity to test complex predictions about more than one parameter (Hoijtink et al., 2019). This is relevant for many psychological theories which often predict specific patterns of equality and order constraints on a set of group means, regression coefficients, or correlations (see Section "Informative Hypotheses"). Such complex hypotheses on more than one parameter cannot be evaluated by estimating all parameters and comparing the corresponding 95% credible intervals (Hoijtink, 2012).[13] However, it is a pitfall to assume that the Bayes factor renders posterior estimates of parameters and effect sizes irrelevant. Quite on the contrary, hypothesis testing and parameter estimation can be combined to answer different research questions (e.g., van Doorn et al., 2021). Whereas hypothesis testing addresses the question "which theory provides the most accurate predictions?", parameter estimation addresses the question "what are the most plausible parameter values?" The former requires one to compare two or more distinct statistical models whereas the latter requires one to assess the posterior distribution of a model including all parameters. Hence, it is a perfectly valid strategy to report both a Bayes factor for testing different theoretical positions and the corresponding parameter estimates with a measure of uncertainty (e.g., a credible interval). In a larger research program, testing and estimation may be combined by using the posterior distribution of an original study

[13] For an overview of how to evaluate a simple (1-dimensional) effect size using Bayesian parameter estimation instead of the Bayes factor, see Dienes (2021).

to construct an informed prior distribution for testing the existence of the effect in a replication (see Section "Bayesian Evidence Synthesis").

Overall, the Bayes factor offers many opportunities for psychological research. Nevertheless, it is a pitfall to assume that the Bayes factor can compensate flaws in the design of a study such as unreliable measures, weak manipulations, or low construct validity. In such cases, the data may simply not be informative for answering the substantive research question. This does not imply that the Bayes factor will necessarily reflect this lack of information by having a value close to one. The Bayes factor merely quantifies the *statistical* evidence provided by the data observed in a specific study without taking into account whether this study serves as an "informative" (i.e., *substantively* valid) test of the theory.

Conclusion

The present review highlighted the wide range of applications of the Bayes factor in psychological research. The illustrations in the different sections and the additional examples on the OSF repository (`https://osf.io/k9c5q/`) show that it has become very easy to use the Bayes factor in practice due to recent technical developments and innovations (e.g., standard models, default prior distributions, and computational tools), user-friendly software (e.g., `JASP` and `R` packages such as `BayesFactor`, `bain`, `baymedr`, or `BFpack`), and introductory papers and tutorials (e.g., Hoijtink et al., 2019; van den Bergh et al., 2020; Wagenmakers et al., 2018).

Still, future work is required to develop Bayesian equivalents for less common classical tests. In making the Bayes factor available for new types of research questions, it is necessary to translate substantive theories to statistical models with corresponding prior distributions. To render the Bayes factor a meaningful measure of the relative evidence, suitable prior distributions for effect sizes and other parameters should reflect theoretical expectations (Dienes, 2019; Lee & Vanpaemel, 2018). This can be achieved either by specifying context-specific, subjective priors based on expertise and prior knowledge (Stefan et al., in press), or by relying on default priors that satisfy general technical requirements (e.g., information consistency; Ly et al., 2016b) while still allowing

researchers to incorporate theoretical expectations to a certain degree (e.g., by defining the scale of the effect size; Rouder et al., 2009).

At first sight, the application of the Bayes factor may seem to be more difficult than the (more-or-less mindless) ritual of null hypothesis significant testing with p-values. After all, researchers need to think about the statistical hypotheses that best reflect the substantive research questions. However, investing some time and effort into the specification of theoretically meaningful models and priors pays off twice: It advances the precise specification of substantive theories and thus enables a principled comparison of competing theories by means of the Bayes factor.

References

Andersson, G., Hesser, H., Veilord, A., Svedling, L., Andersson, F., Sleman, O., Mauritzson, L., Sarkohl, A., Claesson, E., Zetterqvist, V., Lamminen, M., Eriksson, T., & Carlbring, P. (2013). Randomised controlled non-inferiority trial with 3-year follow-up of internet-delivered versus face-to-face group cognitive behavioural therapy for depression. *Journal of Affective Disorders, 151,* 986-994. https://doi.org/10.1016/j.jad.2013.08.022

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment, 10,* 331-344. https://doi.org/10.1037/1040-3590.10.4.331

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6,* 57-86. https://doi.org/10.3758/BF03210812

Bayarri, M. J., Berger, J. O., Forte, A., & García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics, 40,* 1550–1577. https://dx.doi.org/10.1214/12-AOS1013

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

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

Bishara, A. J., & Payne, B. K. (2009). Multinomial process tree models of control and automaticity in weapon misidentification. *Journal of Experimental Social Psychology, 45,* 524–534. https://doi.org/10.1016/j.jesp.2008.11.002

Boehm, U., Steingroever, H., & Wagenmakers, E.-J. (2018). Using Bayesian regression to test hypotheses about relationships between parameters and covariates in

cognitive models. *Behavior Research Methods, 50,* 1248-1269. https://doi.org/10.3758/s13428-017-0940-4

Bott, F. M., Heck, D. W., & Meiser, T. (2020). Parameter validation in hierarchical MPT models by functional dissociation with continuous covariates: An application to contingency inference. *Journal of Mathematical Psychology, 98,* 102388. 10.1016/j.jmp.2020.102388

Branje, S. & Meeus, W. H. J. (2018). Research on adolescent development and relationships (young cohort). *Data Archiving and Networked Services.* https://doi.org/10.17026/dans-zrb-v5wp

Buchner, A., Erdfelder, E., & Vaterrodt-Plünnecke, B. (1995). Toward unbiased measurement of conscious and unconscious memory processes within the process dissociation framework. *Journal of Experimental Psychology: General, 124,* 137-160. https://doi.org/10.1037/0096-3445.124.2.137

Chadwick, D. & Vigabatrin European Monotherapy Study Group. (1999). Safety and efficacy of vigabatrin and carbamazepine in newly diagnosed epilepsy: A multicentre randomised double-blind study. *The Lancet, 354,* 13-19. https://doi.org/10.1016/S0140-6736(98)10531-7

Clyde, M. A. (2018). BAS: Bayesian adaptive sampling for Bayesian model averaging 1.5.3. Comprehensive R Archive Network. https://CRAN.R-project.org/package=BAS

Clyde, M. A., Ghosh, J., & Littman, M. L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics, 20,* 80-101. https://doi.org/10.1198/jcgs.2010.09049

Cohen, J. (1994). The earth is round, p<.05. *American Psychologist, 49,* 997-1003. https://dx.doi.org/10.1037/0003-066X.49.12.997

Consonni, G., Fouskakis, D., Liseo, B., & Ntzoufras, I. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis, 13,* 627-679. https://doi.org/10.1214/18-BA1103

Dienes, Z. (2019). How do I know what my theory predicts? *Advances in Methods and Practices in Psychological Science, 2, 364–377.* https://doi.org/10.1177/2515245919876960

Dienes, Z. (2021). How to use and report Bayesian hypothesis tests. *Psychology of Consciousness: Theory, Research, and Practice, 8*, 9–26. https://doi.org/10.1037/cns0000258

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

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Assfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie / Journal of Psychology, 217,* 108-124. https://doi.org/10.1027/0044-3409.217.3.108

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior.* New York, NY, Cambridge University Press.

Fu, Q., Hoijtink, H., & Moerbeek, M. (2021). Sample-size determination for the Bayesian t test and Welch's test using the approximate adjusted fractional Bayes factor. *Behavior Research Methods, 53,* 139-152. https://doi.org/10.3758/s13428-020-01408-1

Glatzer, W., & Gulyas, J. (2014). Cantril self-anchoring striving scale. In A. C. Michalos (Ed.), Encyclopedia of quality of life and well-being research (pp. 509–511).

Gronau, Q. F., Heck, D. W., Berkhout, S. W., Haaf, J. M., & Wagenmakers, E.-J. (2021). A primer on Bayesian model-averaged meta-analysis. *Advances in Methods*

and Practices in Psychological Science, 4, 1-19. https://doi.org/10.1177%2F25152459211031256

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2020). Informed Bayesian t-tests. *The American Statistician, 74*, 137–143. https://doi.org/10.1080/00031305.2018.1562983

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

Gu, X., Hoijtink, H., Mulder, J., & van Lissa, C.J. (2019). bain: Bayes factors for informative hypotheses. R package version 0.2.1. https://CRAN.R-project.org/package=bain

Haaf, J. M., Ly, A., & Wagenmakers, E.-J. (2019). Retire significance, but still test hypotheses. *Nature, 567,* 461–461. https://doi.org/10.1038/d41586-019-00972-7

Heathcote, A., Brown, S. D., & Wagenmakers, E.-J. (2015). An introduction to good practices in cognitive modeling. In: *An Introduction to Model-Based Cognitive Neuroscience*, (pp. 25-48). Springer, New York, NY.

Heathcote, A., Lin, Y.-S., Reynolds, A., Strickland, L., Gretton, M., & Matzke, D. (2018). Dynamic models of choice. *Behavior Research Methods, 51,* 961-985. https://doi.org/https://doi.org/10.3758/s13428-018-1067-y

Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods, 50,* 264-284. https://doi.org/10.3758/s13428-017-0869-7

Heck, D.W. & Davis-Stober, C.P. (2019). Multinomial models with linear inequality constraints: Overview and improvment of computational methods for Bayesian inference. *Journal of Mathematical Psychology, 91,* 70-87. https://doi.org/10.1016/j.jmp.2019.03.004

Heck, D. W., & Erdfelder, E. (2019). Maximizing the expected information gain of cognitive modeling via design optimization. *Computational Brain & Behavior, 2,* 202-209. https://doi.org/10.1007/s42113-019-00035-0

Heck, D. W., Gronau, Q. F., & Wagenmakers, E.-J. (2019). metaBMA: Bayesian model averaging for random and fixed effects meta-analysis. https://CRAN.R-project.org/package=metaBMA

Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2020). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science, 3*, 200–215. https://doi.org/10.1177/2515245919898657

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science 14,* 382-401. https://doi.org/10.1214/ss/1009212519

Hoijtink, H. (2012). *Informative hypotheses: Theory and practice for behavioral and social scientists.* Chapman & Hall/CRC.

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods, 24,* 539-556. https://doi.org/10.1037/met0000201

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30,* 513–541. https://doi.org/10.1016/0749-596X(91)90025-F

JASP Team (2020). JASP (Version 0.13.1)[Computer software]. https://jasp-stats.org/

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophy Society, 31*, 203–222. https://doi.org/10.1017/S030500410001330X Proceedings of the Cambridge Philosophy Society, 31, 203–222.

Jeffreys, H. (1939). *Theory of probability (1st ed.).* Oxford: The Clarendon Press.

Kaul, S., & Diamond, S. (2006). Good enough: A primer on the analysis and interpretation of noninferiority trials. *Annals of Internal Medicine, 145,* 62-69. https://doi.org/10.7326/0003-4819-145-1-200607040-00011

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

Kellen, D. (2019). A model hierarchy for psychological science. *Computational Brain & Behavior, 2,* 160-165. https://doi.org/10.1007/s42113-019-00037-y

Kellen, D., & Klauer, K. C. (2014). Discrete-state and continuous models of recognition memory: Testing core properties under minimal assumptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 1795-1804. https://doi.org/10.1037/xlm0000016

Keysers, C., Gazzola, V., & Wagenmakers, E.-J. (2020). Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience, 23*, 788–799. https://doi.org/10.1038/s41593-020-0660-4

Kiers, H., & Tendeiro, J. (2019). *With Bayesian estimation one can get all that Bayes factors offer, and more.* PsyArXiv. https://doi.org/10.31234/osf.io/zbpmy

Klaassen, F., Zedelius, C. M., Veling, H., Aarts, H., Hoijtink, H. (2018). All for one or some for all? Evaluating informative hypotheses using multiple N = 1 studies. *Behavior Research Methods, 50*, 2276 – 2291. https://doi.org/10.3758/s13428-017-0992-5

Klugkist, I., Laudy, O., & Hoijtink, H. (2010). Bayesian evaluation of inequality and equality constrained hypotheses for contingency tables. *Psychological Methods, 15,* 281-299. https://dx.doi.org/10.1037/a0020137

Kooijman, M.N., Kruithof, C.J., van Duijn, C.M., . . . , & Jaddoe, V. W. V. (2017). The Generation R study: Design and cohort update 2017. *European Journal of Epidemiology, 31,* 1243–1264. https://doi.org/10.1007/s10654-016-0224-9

Kuiper, R. M., Buskens, V., Raub, W., Hoijtink, H. (2013). Combining statistical evidence from several studies: A method using Bayesian updating and an example from research on trust problems in social and economic exchange. *Sociological Methods & Research, 42*, 60–81. https://doi.org/10.1177/0049124112464867

Kuiper, R. M., Hoijtink, H., & Silvapulle, M. (2011). An Akaike-type information criterion for model selection under inequality constraints. *Biometrika, 98*, 495–501. https://doi.org/10.1093/biomet/asr002

Lee, M. D. (2016). Bayesian outcome-based strategy classification. *Behavior Research Methods, 48,* 29-41. https://doi.org/10.3758/s13428-014-0557-9

Lee, M. D., & Vanpaemel, W. (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review, 25,* 114-127. https://doi.org/10.3758/s13423-017-1238-3

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

Leplaa, H. J., Rietbergen, C., & Hoijtink, H. (2020). Bayesian evaluation of replication studies. PsyArXiv. https://doi.org/10.31234/osf.io/49tbz

Li, Y., & Clyde, M. A. (2018). Mixtures of g-priors in generalized linear models. *Journal of the American Statistical Association, 113* , 1828-1845. https://doi.org/10.1080/01621459.2018.1469992

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association, 103*, 410-423. https://doi.org/10.1198/016214507000001337

Lin, T.-K., You, K.-X., Hsu, C.-T., Li, Y.-D., Lin, C.-L., Weng, C.-Y., & Koo, M. (2019). Negative affectivity and social inhibition are associated with increased cardiac readmission in patients with heart failure: A preliminary observation study. *PLOS ONE, 14* (4), e0215726. https://doi.org/10.1371/journal.pone.0215726

Linde, M., & van Ravenzwaaij, D. (2019). baymedr: An R package for the calculation of Bayes factors for equivalence, non-inferiority, and superiority designs. arXiv: https://arxiv.org/abs/1910.11616

Linde, M. & van Ravenzwaaij, D. (2021). baymedr: Computation of Bayes factors for common designs. R package version 0.1.1. https://CRAN.R-project.org/package=baymedr

Lindley, D. V. (1957). A statistical paradox. *Biometrika, 44*, 187–192. https://doi.org/10.2307/2333251

Lodewyckx, T., Kim, W., Lee, M. D., Tuerlinckx, F., Kuppens, P., & Wagenmakers, E.-J. (2011). A tutorial on Bayes factor estimation with the product space method. *Journal of Mathematical Psychology, 55,* 331-347. https://doi.org/10.1016/j.jmp.2011.06.001

Lucas, J. W. (2003). Status processes and the institutionalization of women as leaders. *American Sociological Review*, 68 , 464-480. https://www.jstor.org/stable/1519733

Ly, A., Stefan, A., van Doorn, J., Dablander, F., van den Bergh, D., Sarafoglou, A., Kucharský, Š., Derks, K., Gronau, Q. F., Gupta, A.R.K.N., Boehm, U., van Kesteren, E.-J., Hinne, M., Matzke, D., Marsman., M., & Wagenmakers, E.-J. (2020). The Bayesian methodology of Sir Harold Jeffreys as a practical alternative to the p-value hypothesis test. *Computational Brain & Behavior, 3,* 153–161. https://doi.org/10.1007/s42113-019-00070-x

Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016a). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology.

Journal of Mathematical Psychology, 72, 19-32. https://dx.doi.org/10.1016/j.jmp.2015.06.004

Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016b). An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology, 72,* 43-55. https://dx.doi.org/10.1016/j.jmp.2016.01.003

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

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods, 16,* 406-419. https://doi.org/10.1037/a0024377

Morey, R. D., & Rouder (2018). BayesFactor: Computation of Bayes factors for common designs. R package version 0.9.12-4.2. https://CRAN.R-project.org/package=BayesFactor

Mulder, J., Gu, X., Olsson-Collentine, A., Tomarken, A., Böing-Messing, F., Hoijtink, H., Meijerink, M., Williams, D. R., Menke, J., Fox, J.-P., Rosseel, Y., Wagenmakers, E.-J., van Lissa, C. (2019). BFpack: Flexible Bayes factor testing of scientific theories in R. http://arxiv.org/abs/1911.07728

Myung, J. I., Cavagnaro, D. R., & Pitt, M. A. (2013). A tutorial on adaptive design optimization. *Journal of Mathematical Psychology, 57*, 53–67. https://doi.org/10.1016/j.jmp.2013.05.005

Myung, I. J., Forster, M. R., & Browne, M. W. (2000). Guest editors' introduction: Special issue on model selection. *Journal of Mathematical Psychology, 44*, 1-2. https://doi.org/10.1006/jmps.1999.1273

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

O'Hagan A. (1995). Fractional Bayes factors for model comparison (with discussion). *Journal of the Royal Statistical Society. Series B, 57,* 99–138. http://www.jstor.org/stable/2346088

Oldehinkel A. J., Rosmalen J. G. M., Buitelaar J. K., Hoek H. W., Ormel J., Raven, D., ..., Hartman C. A. (2015). Cohort profile update. The TRacking Adolescents' Individual Lives Survey (TRAILS). *International Journal of Epidemiology, 44*, 76-76n. https://doi.org/10.1093/ije/dyu225

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

Piaggio, G., Elbourne, D. R., Pocock, S. J., Evans, S. J. W., & Altman, D. G. (2012). Reporting of noninferiority and equivalence randomized trials. *Journal of the American Medical Association, 308,* 2594–2604. https://doi.org/10.1001/jama.2012.87802

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling, In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, Vienna, Austria.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rivers, A. M. (2017). The weapons identification task: Recommendations for adequately powered research. *PLOS ONE, 12,* e0177857. https://doi.org/10.1371/journal.pone.0177857

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107,* 358-367. https://doi.org/10.1037/0033-295X.107.2.358

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

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review, 18*, 682-689. https://doi.org/10.3758/s13423-011-0088-7

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research, 47,* 877-903. https://doi.org/10.1080/00273171.2012.734737

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E.-J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science, 8,* 520–547. https://doi.org/10.1111/tops.12214

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

Royal, R. (1997). *Statistical evidence: A likelihood paradigm.* New York: Chapman and Hall/CRC.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. Psychological Methods, 22(2), 322–339. https://doi.org/10.1037/met0000061

Scott, J. G., & Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of statistical planning and inference, 136,* 2144-2162. https://doi.org/10.1016/j.jspi.2005.08.031

Senn, S. (2008). *Statistical issues in drug development (2nd ed.).* Chichester, UK: Wiley.

Sinharay, S., & Stern, H. S. (2002). On the sensitivity of Bayes factors to the prior distributions. *The American Statistician, 56*, 196–201. https://doi.org/10.1198/000313002137

Stan Development Team. (2020). *Stan user's guide.* Version 2.22. https://mc-stan.org/

Stefan, A. M., Evans, N. J., & Wagenmakers, E.-J. (in press). Practical challenges and methodological flexibility in prior elicitation. *Psychological Methods.* https://doi.org/10.1037/met0000354

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,* 1042–1058. https://doi.org/10.3758/s13428-018-01189-8

Suppes, P. (1966). Models of data. In: E. Nagel, P. Suppes, & A. Tarski (Eds.), *Studies in Logic and the Foundations of Mathematics* (pp. 252-261). Elsevier. https://doi.org/10.1016/S0049-237X(09)70592-0

Tendeiro, J. N., & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods, 24,* 774–795. https://doi.org/10.1037/met0000221

van den Bergh, D., Haaf, J. M., Ly, A., Rouder, J. N., & Wagenmakers, E.-J. (2019). A cautionary note on estimating effect size [Manuscript submitted for publication]. PsyArXiv. https://psyarxiv.com/h6pr8

van den Bergh, D., van Doorn, J., Marsman, M., Draws, T., van Kesteren, E.-J., Derks, K., Dablander, F., Gronau, Q.F., Kucharskyý, Š., Gupta, A.R.K.N., Sarafoglou, A., Voelkel, J.G., Stefan, A., Ly, A., Hinne, M., Matzke, D., & Wagenmakers, E.-J. (2020). A tutorial on conducting and interpreting a Bayesian ANOVA in JASP. *L'Année psychologique, 120,* 73–69. https://doi.org/10.3917/anpsy1.201.0073

van den Bergh, D., Clyde, M. A., Gupta, A. R. K. N., de Jong, T., Gronau, Q. F., Marsman, M., Ly, A., & Wagenmakers, E.-J. (in press). A tutorial on Bayesian multi-model linear regression with BAS and JASP. *Behavior Research Methods.* https://doi.org/10.3758/s13428-021-01552-2

van de Werf, F., Adgey, J., Ardissino, D., Armstrong, P. W., Aylward, P., Barbash, G., ..., White, H. (1999). Single-bolus tenecteplase compared with front-loaded alteplase in acute myocardial infarction: The ASSENT-2 double-blind randomised trial. *The Lancet, 354,* 716-722. https://doi.org/10.1016/S0140-6736(99)07403-6

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., Hinne, M., Kucharský, Š., Ly, A., Marsman, M., Matzke, D., Gupta, A. R. K. N., Sarafoglou, A., Stefan, A., Voelkel, J. G., & Wagenmakers, E.-J. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review, 28*, 813–826. https://doi.org/10.3758/s13423-020-01798-5

van Lissa, C.J., Gu, X., Mulder, J., Rosseel, Y., van Zundert, C., & Hoijtink, H. (2021). Teacher's Corner: Evaluating informative hypotheses using the Bayes factor in structural equation models. *Structural Equation Modeling, 28,* 292-301. https://doi.org/10.1080/10705511.2020.1745644

van Ravenzwaaij, D., Monden, R., Tendeiro, J. N., & Ioannidis, J. P. A. (2019). Bayes factors for superiority, non-inferiority, and equivalence designs. *BMC Medical Research Methodology, 19,* 71. https://doi.org/10.1186/s12874-019-0699-7

van Ravenzwaaij, D., & Ioannidis, J. P. A. (2019). True and false positive rates for different criteria of evaluating statistical evidence from clinical trials. *BMC Medical Research Methodology, 19,* 218. https://doi.org/10.1186/s12874-019-0865-y

van Ravenzwaaij, D., & Wagenmakers, E.-J. (in press). Advantages masquerading as 'issues' in Bayesian hypothesis testing: A commentary on Tendeiro and Kiers (2019). *Psychological Methods.* https://doi.org/10.31234/osf.io/nf7rp

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In: *Oxford Handbook of Computational and Mathematical Psychology* (pp. 300-319). New York, NY, Oxford University Press.

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

Vanpaemel, W. (2020). Strong theory testing using the prior predictive and the data prior. *Psychological Review, 127,* 136-145. https://doi.org/10.1037/rev0000167

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

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,* 158-189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25,* 35–57. https://doi.org/10.3758/s13423-017-1343-3

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

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods, 4,* 212-213. https://dx.doi.org/10.1037/1082-989X.4.2.212

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology, 44*, 92–107. https://dx.doi.org/10.1006/jmps.1999.1278

Williams, L.E. & Bargh, J.A. (2008). Keeping one's distance: The influence of spatial distance cues on affect and evaluation. *Psychological Sciene, 19,* 302-308. https://dx.doi.org/10.1111/j.1467-9280.2008.02084.x

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. Smith (Eds.), Bayesian statistics: Proceedings of the first international meeting held in Valencia (Vol. 1, pp. 585-603). Springer.

Zondervan-Zwijnenburg, M.A.J., Veldkamp, S.A.M., Neumann, A., Barzeva, S.A., Nelemans, S.A., van Beijsterveldt, C.E.M., Branje, S.J.T., Hillegers, M.H.J., Meeus, W.H.J., Tiemeier, H., Hoijtink, H., Oldehinkel, A.J. & Boomsma, D.I. (2020). Parental age and offspring childhood mental health: A multi-cohort, population-based investigation. *Child Development, 91,* 964-982. https://doi.org/10.1111/cdev.13267