

# The JASP Guidelines for Conducting and Reporting a Bayesian Analysis

Johnny van Doorn<sup>\*1</sup>, Don van den Bergh<sup>1</sup>, Udo Böhm<sup>1</sup>, Fabian  
Dablander<sup>1</sup>, Koen Derks<sup>2</sup>, Tim Draws<sup>1</sup>, Nathan J. Evans<sup>1</sup>,  
Quentin F. Gronau<sup>1</sup>, Max Hinne<sup>1</sup>, Šimon Kucharský<sup>1</sup>, Alexander  
Ly<sup>1,3</sup>, Maarten Marsman<sup>1</sup>, Dora Matzke<sup>1</sup>, Akash R. Komarlu  
Narendra Gupta<sup>1</sup>, Alexandra Sarafoglou<sup>1</sup>, Angelika Stefan<sup>1</sup>, Jan  
G. Voelkel<sup>4</sup>, and Eric-Jan Wagenmakers<sup>1</sup>

<sup>1</sup>University of Amsterdam

<sup>2</sup>Nyenrode Business University

<sup>3</sup>Centrum Wiskunde & Informatica

<sup>4</sup>Stanford University

## Abstract

Despite the increasing popularity of Bayesian inference in empirical  
research, few practical guidelines provide detailed recommendations for

---

<sup>\*</sup>Correspondence concerning this article may be addressed to Johnny van Doorn, University of Amsterdam, Department of Psychological Methods, Valckeniersstraat 59, 1018 XA Amsterdam, the Netherlands. E-mail may be sent to JohnnyDoorn@gmail.com. This work was supported in part by a Vici grant from the Netherlands Organization of Scientific Research (NWO) awarded to EJW (016.Vici.170.083). DM is supported by a Veni grant (451-15-010) from the NWO. MM is supported by a Veni grant (451-17-017) from the NWO. Centrum Wiskunde & Informatica (CWI) is the national research institute for mathematics and computer science in the Netherlands.

how to apply Bayesian procedures and interpret the results. Here we offer specific guidelines for the different stages of Bayesian statistical inference in a research setting: planning, executing, interpreting, and reporting. The guidelines for each stage are illustrated with an example. Although the guidelines are geared toward analyses performed with the open-source statistical software JASP, most guidelines extend to Bayesian inference in general.

*Keywords:* Bayesian inference, scientific reporting, statistical software.

In recent years, Bayesian inference has become increasingly popular, both in statistical science and in applied fields such as psychology, biology, and econometrics (e.g., Vandekerckhove et al., 2018; Andrews & Baguley, 2013). For the pragmatic researcher, the adoption of the Bayesian framework brings several advantages over the standard framework of frequentist null-hypothesis significance testing (NHST), including the ability to (1) quantify uncertainty about effect sizes in a straightforward manner (Wagenmakers, Marsman, et al., 2018); (2) compare the predictive adequacy of two competing statistical models (Dienes & McLatchie, 2018); (3) obtain evidence in favor of the null hypothesis (Dienes, 2014); (4) integrate prior knowledge and observed data (Lee & Vanpaemel, 2018; Gronau et al., 2018); and (5) monitor the evidence as the data accumulate (Rouder, 2014). However, the relative novelty of conducting Bayesian analyses in applied fields means that there are no detailed reporting standards, and this may in term frustrate the broader adoption and proper interpretation of the Bayesian framework.

Some recent guidelines include Bayesian inference, but these are either minimalist (The BaSiS group, 2001; Appelbaum et al., 2018), focus only on relatively complex statistical tests (Depaoli & van de Schoot, 2017), are too specific to a certain field (Spiegelhalter et al., 2000; Sung et al., 2005), or do not cover the

full inferential process (Jarosz & Wiley, 2014). The current article aims to provide a general overview of the different stages of the Bayesian inference process. Specifically, we focus on guidelines for analyses conducted in JASP (JASP Team, 2018). JASP is an open-source statistical software program with a graphical user interface that features both Bayesian and frequentist versions of common tests such as the binomial test, the  $t$ -test, analysis of variance (ANOVA), the correlation test, and regression analysis (e.g., Marsman & Wagenmakers, 2017; Wagenmakers, Love, et al., 2018).

We discuss four stages of analysis: planning, executing, interpreting, and reporting. In order to provide a concrete illustration of the guidelines for each of the four stages, a data example by Frisby & Clatworthy (1975), taken from the JASP data library, will be discussed in each section. The data set concerns the time it took two groups of participants to see a figure hidden in a stereogram – one group received advance visual information about the scene (i.e., the VV condition), whereas the other group did not (i.e., the NV condition).<sup>1</sup> Three additional examples (mixed ANOVA, correlation analysis, and a  $t$ -test with an informed prior) are provided in an online appendix at <https://osf.io/nw49j/>.

## Stage 1: Planning the Analysis

The quality of the research starts with the planning. Good research starts with good planning. The planning stage consists of specifying the goal of the analysis, the statistical model, and the sampling plan. Ideally, the decisions made at this stage are solidified in the form of preregistration (e.g., Wagenmakers et al., 2012; De Groot, 1956/2014) or a Registered Report (e.g., Chambers, 2013). From a Bayesian perspective, preregistration ensures that the data are not used twice:

---

<sup>1</sup>The variables in the data are participant number, the time (in seconds) the participant needed to see the hidden figure (i.e., fuse time), experimental condition (VV = with visual information, NV = without visual information), and the log-transformed fuse time.

once to revive a comatose and unexpected theory (i.e., updating the prior odds), and then again to test that theory (computing the Bayes factor).

**Specifying the goal of the analysis.** We recommend that researchers carefully consider the goal, that is, the research question that they wish to answer. Usually, the goal is either hypothesis testing (for which we recommend the Bayes factor, see Box 1) or parameter estimation (for which we recommend to either plot the entire posterior distribution or summarize it by a credible interval, see Box 2). If the goal of the study is to determine the presence or absence of an effect, a hypothesis test is conducted. If the goal of the study is to determine the size of the effect, if it exists, estimation is used. These procedures are not mutually exclusive and can be combined.

**Box 1. Hypothesis testing.** The principled approach to Bayesian hypothesis testing is by means of the Bayes factor (e.g., Wrinch & Jeffreys, 1921; Etz & Wagenmakers, 2017; Jeffreys, 1939; Ly et al., 2016). The Bayes factor quantifies the relative predictive performance of two rival hypotheses. The Bayes factor can be seen as the updating factor with which prior beliefs about the plausibility of each hypothesis are updated by data to posterior beliefs about the plausibility of each hypothesis (see Equation 1). The first term in Equation 1, the prior odds, indicates the relative plausibility of either hypothesis, before seeing the data. The second term, the Bayes factor, indicates the evidence in the data for each hypothesis. The third term, the posterior odds, indicates the relative plausibility of either hypothesis, after seeing the data.

$$\underbrace{\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}}_{\text{Prior odds}} \times \underbrace{\frac{p(D | \mathcal{H}_1)}{p(D | \mathcal{H}_0)}}_{\text{Bayes factor}_{10}} = \underbrace{\frac{p(\mathcal{H}_1 | D)}{p(\mathcal{H}_0 | D)}}_{\text{Posterior odds}} \quad (1)$$

The subscript in the Bayes factor notation indicates which hypotheses are compared.  $\text{BF}_{10}$  indicates the Bayes factor in favor of  $\mathcal{H}_1$  over  $\mathcal{H}_0$ , whereas  $\text{BF}_{01}$  indicates the Bayes factor in favor of  $\mathcal{H}_0$  over  $\mathcal{H}_1$ . Specifically,  $\text{BF}_{10} = 1/\text{BF}_{01}$ . The Bayes factor can range from 0 to  $\infty$ , and a Bayes factor equal to 1 signifies that both hypotheses are supported equally by the data. The larger  $\text{BF}_{10}$  becomes, the more support is gained for  $\mathcal{H}_1$ . Conversely, the larger  $\text{BF}_{01}$  becomes, the more support is gained for  $\mathcal{H}_0$ . This principle is further illustrated in Figure 4.

**Box 2. Parameter estimation.** For Bayesian parameter estimation, the posterior distribution of the model parameters and its  $x\%$  credible interval is key. The posterior distribution reflects the relative plausibility of the parameter values after prior knowledge has been combined with the data. Specifically, we start with a prior distribution on the model parameters that reflects how plausible each parameter value is before seeing the data. This distribution is updated with the data to form the posterior distribution. Parameter values that predicted the data well receive a boost in credibility, whereas parameter values that predicted the data poorly suffer a decline (Wagenmakers et al., 2016). Equation 2 illustrates this principle. The first term indicates the prior beliefs about the parameter  $\theta$ . The second term is the updating factor: for each value of  $\theta$ , the quality of its prediction is compared to the average quality of the predictions over all values of  $\theta$ . The

third term indicates the posterior beliefs about  $\theta$ .

$$\underbrace{p(\theta)}_{\text{Prior belief about } \theta} \times \underbrace{\frac{p(D|\theta)}{p(D)}}_{\substack{\text{Predictive adequacy} \\ \text{of specific } \theta \\ \text{Average predictive} \\ \text{adequacy across all } \theta\text{'s}}} = \underbrace{p(\theta|D)}_{\text{Posterior belief about } \theta} . \quad (2)$$

The posterior distribution can be plotted, or summarized by an  $x\%$  credible interval. The credible interval is taken such that  $x\%$  of the posterior mass lies in that interval. Two popular ways of creating a credible interval are the highest density credible interval, which is the narrowest interval containing the specified mass, and the central credible interval, which is created by cutting off  $\frac{100-x}{2}\%$  from each of the tails of the posterior distribution.

**Specifying the statistical model.** The goal of the analysis informs the statistical model. For instance, if the goal of the analysis is to determine whether two variables are correlated, one could use a bivariate normal model and then conduct inference on the correlation parameter Pearson's  $\rho$ . The next step is determining the two-sidedness of the procedure. For hypothesis testing, this means deciding whether the procedure is one-sided (i.e., predicting the specific direction of the effect) or two-sided (i.e., evidence can be found for either a positive or negative effect). For parameter estimation, we recommend always using the two-sided model instead of the one-sided model, in order to get a sensible estimate: when using a positive one-sided model, and the observed effect turns out to be negative, we still get a positive estimate for the effect in the population under the one-sided alternative model due to truncation of the parameter space.

Next in model specification, the type and shape of the prior distribution,

including its justification, must be determined a priori. For the most common statistical models (e.g., correlations,  $t$ -tests, and ANOVA), certain “default” prior distributions are available that can be used in cases where prior knowledge is absent, vague, or difficult to elicit (for more information, see Ly et al., 2016). These priors are featured in JASP as the default options for the Bayesian analyses. In cases where prior information is present, different prior options can be used. However, the more the selected prior deviates from the default options, the more justification for its selection is needed. For an example, see the informed  $t$ -test example in the online appendix at <https://osf.io/ybszx/>. Additionally, the robustness of the result to different values of the hyperparameter (i.e., the parameter that governs the shape of the prior distribution) can be explored and included in the report after the data have been collected. In JASP, this is available in the form of a Bayes factor robustness plot. This plot graphs the Bayes factor as a function of the specified hyperparameter, and visualizes the influence of the specification on the resulting Bayes factor. An example of such a plot is provided in the section on reporting the analysis.

**Specifying the sampling plan.** When determining the sample size, it can be desirable to consider a sequential design, which can increase the efficiency of gaining information from experiments. As Bayesian inference allows for monitoring the results as the data come in, it is possible to not specify the sample size a priori, but instead specify the desired degree of evidence in favor of either the null ( $\mathcal{H}_0$ ) or the alternative ( $\mathcal{H}_1$ ) hypothesis in terms of the Bayes factor. An example of this would be to stop data collection when the obtained result is either  $\text{BF}_{10} = 10$  or  $\text{BF}_{01} = 10$ . This approach can also be combined with a maximum sample size ( $N$ ), where data collection stops when either the maximum  $N$  or the desired Bayes factor is obtained (for practical examples, see Matzke et al., 2015; Wagenmakers et al., 2015). A further possibility for planning the study

is Bayes factor design analysis (Schönbrodt & Wagenmakers, 2018; Stefan et al., 2018), a framework that helps researchers estimate and define an informed upper limit for sample size for maximum efficiency and informativeness. When the study is observational and the data are available ‘en bloc’, the sampling plan becomes irrelevant in the planning stage.

## Stereogram Example

First we consider the research question posed by the original authors. The original goal was to determine whether participants who received advance visual information about the scene exhibit a shorter fuse time. In order to determine whether or not an effect is present, we conduct a hypothesis test. If the test reveals support in favor of the presence of the effect then we have grounds for a secondary analysis that concerns the estimation of the size of the effect.

Second, we specify the model. The goal of the study is to investigate the difference between two between-subjects conditions, which can be achieved with a two-sample  $t$ -test. For hypothesis testing, we compare the null hypothesis (i.e., advance visual information has no effect on fuse times) with a one-sided alternative hypothesis (i.e., advance visual information *shortens* the fuse times). The hypotheses are thus  $\mathcal{H}_0 : \delta = 0$  and  $\mathcal{H}_+ : \delta > 0$ , where  $\delta$  is the standardized effect size (i.e., Cohen’s  $d$ ),  $\mathcal{H}_0$  denotes the null hypothesis and  $\mathcal{H}_+$  denotes the one-sided alternative hypothesis (note the ‘+’ in the subscript). In addition to testing the hypotheses specified above, we would like to estimate the magnitude of the effect (under the assumption that the effect exists). For estimation, we use the two-sided  $t$ -test model, and then plot the posterior distribution of  $\delta$  and summarize the posterior distribution by the 95% central credible interval.

We complete the model specification by assigning prior distributions to the model parameters. Since we only have little prior knowledge about the subject,

we will choose a default prior option for the two-sample  $t$ -test, which is a Cauchy distribution with the hyperparameter  $r$  set to  $1/\sqrt{2}$  (the Cauchy distribution fulfills certain desiderata, see Liang et al., 2008; Rouder et al., 2009; Ly et al., 2016). We will then assess the robustness of the Bayes factor to this prior specification by means of a Bayes factor robustness plot.

Since the data are already available, we do not have to specify a sampling plan. The original data set has a sample size of 103, out of which 25 participants were eliminated due to failing an initial stereoacuity test, leaving 78 participants (43 in the NV condition and 35 in the VV condition).

## Stage 2: Executing the Analysis

Before executing the analysis and interpreting the outcome, it is important to confirm that the intended analyses are appropriate and the models are not grossly misspecified for the data at hand. In other words, it is strongly recommended to examine the validity of the model assumptions (e.g., normally distributed residuals or equal variances across groups). Such assumptions may be checked by plotting the data, inspecting summary statistics, or conducting formal assumption tests (but see Tijmstra (2018)).

A powerful demonstration of the dangers of failing to check the assumptions is provided by Anscombe’s quartet (Anscombe, 1973; see Figure 1). The quartet consists of four fictitious data sets of equal size that each have the same observed Pearson’s product moment correlation  $r$ , and therefore lead to the same inferential result both in a frequentist and a Bayesian framework. However, visual inspection of the scatterplots immediately reveals that three of the four data sets are not suitable for a linear correlation analysis, and the statistical inference for these three data sets is meaningless or even misleading. This example highlights the adage that conducting a Bayesian analysis does not safeguard

against general statistical malpractice – the Bayesian framework is as vulnerable to violations of assumptions as its frequentist counterpart. In cases where assumptions are violated, a non-parametric test can be used, and the parametric results should be interpreted with caution.

Once the quality of the data has been confirmed, the planned analyses can be carried out. JASP offers a graphical user interface for both frequentist and Bayesian analyses. JASP 0.9.2 features the following Bayesian analyses: the binomial test, the Chi-square test, the  $t$ -test (one-sample, paired sample, two-sample, and Wilcoxon rank sum tests), ANOVA, ANCOVA, repeated measures ANOVA, correlations (Pearson’s  $\rho$  and Kendall’s  $\tau$ ), linear regression, and log-linear regression. After loading the data into JASP, the desired analysis can be conducted by dragging and dropping variables into the appropriate boxes; tick marks can be used to select the desired output (for details, see <https://jasp-stats.org/how-to-use-jasp>).

The resulting output (i.e., figures and tables) can be annotated and saved as a `.jasp` file. Output can then be shared with peers, with or without the real data in the `.jasp` file; if the real data are added, reviewers can easily reproduce the analyses, conduct alternative analyses, or insert comments.

## Stereogram Example

In order to check for violations of the assumptions of the  $t$ -test, Figure 2 shows the distribution of the dependent variable, split by condition. Interpreting the histograms and boxplots suggests that the data do contain outliers, and that the variances in the groups are roughly equal (observed standard deviations of 0.814 and 0.818 in the NV and the VV group, respectively); however, the histogram also suggests that the data may not be approximately normally distributed. Hence it seems prudent to assess the robustness of the result by also conducting

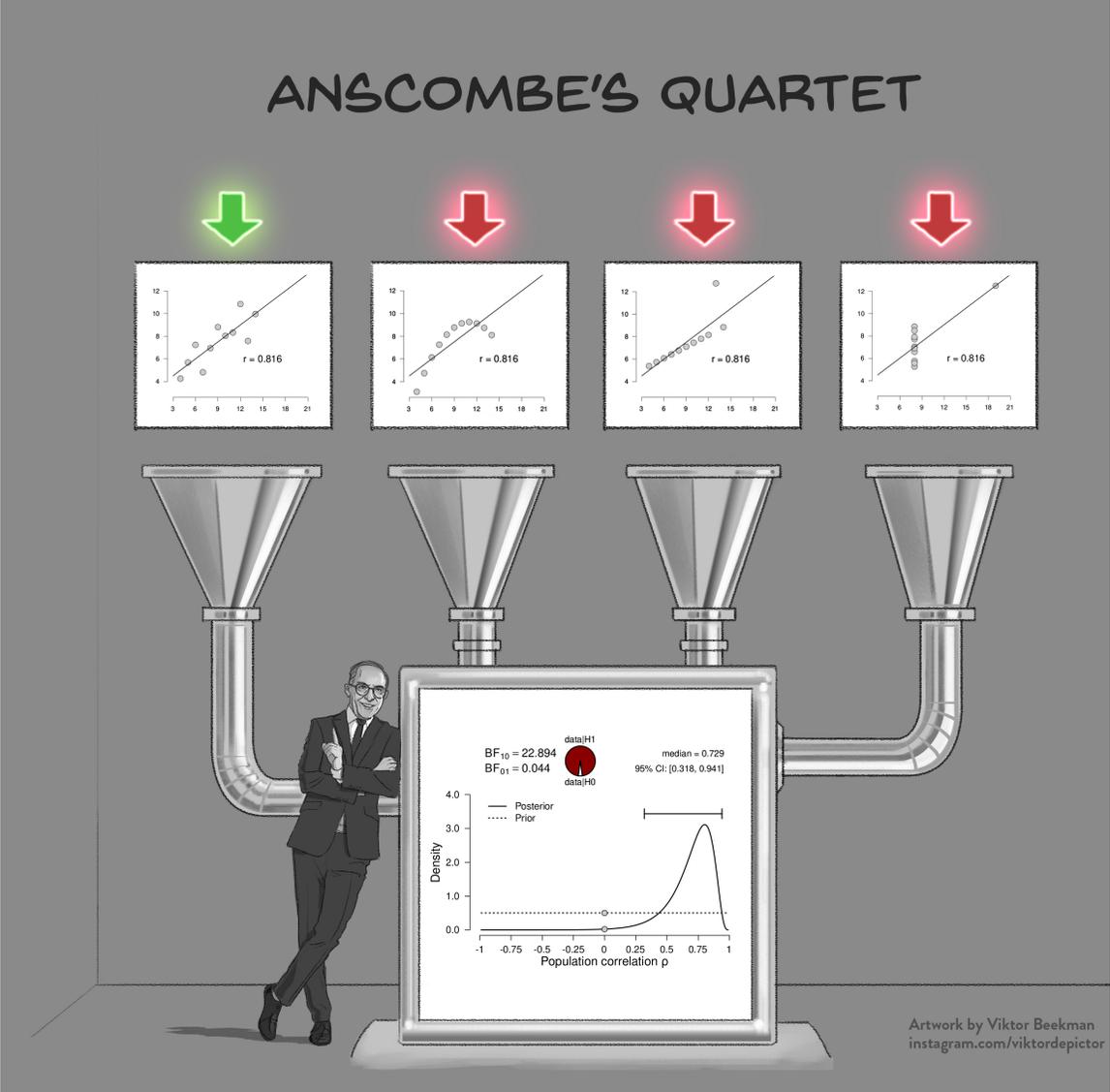


Figure 1: Model misspecification is also a problem for Bayesian analyses. The four scatterplots on top show Anscombe's quartet (Anscombe, 1973); the bottom panel shows the corresponding inference which is identical for all four scatter plots. Except for the leftmost scatterplot, all data violate the assumptions of the linear correlation analysis in important ways.

a non-parametric equivalent of the  $t$ -test (i.e., the Mann-Whitney test).

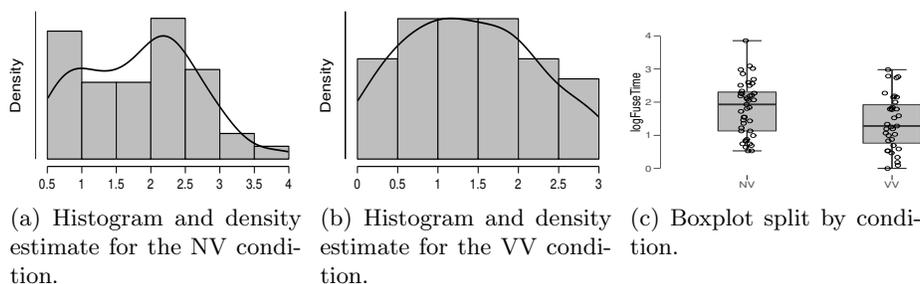


Figure 2: Descriptive plots allow a visual assessment of the assumptions of the  $t$ -test for the stereogram data. The left and middle panel show histograms of the dependent variable (i.e., log-transformed fuse time measured in seconds) split by experimental condition. The solid lines indicate a kernel density estimate. The right panel shows boxplots, including the jittered data points, for each of the experimental conditions. Figures from JASP.

Following the assumption check we proceed to execute the analyses in JASP. For hypothesis testing, we obtain a Bayes factor using the one-sided Bayesian two-sample  $t$ -test. Figure 3 shows the JASP user interface for this procedure. For parameter estimation, we obtain a posterior distribution and credible interval, using the two-sided Bayesian two-sample  $t$ -test. The relevant boxes for the various plots were ticked, and an annotated `.jasp` file was created with distribution plots, the one-sided results that were used for hypothesis testing (Bayes factor and robustness check), the two-sided results that were used for estimation (posterior distribution), and the one-sided results of the Bayesian Mann-Whitney test. The `.jasp` file can be found at <https://osf.io/nw49j/>. The next section outlines how these results are to be interpreted.

### Stage 3: Interpreting the Results

With the analysis outcome in hand, conclusions can be drawn. We first discuss the scenario of hypothesis testing, where the goal is to conclude whether an

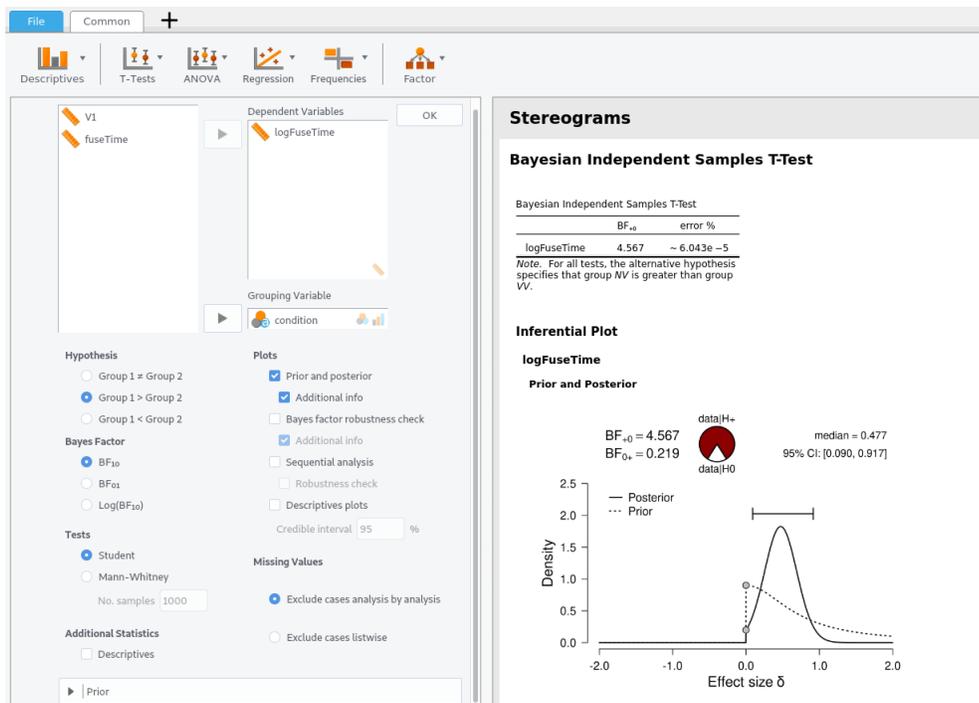


Figure 3: JASP menu for the Bayesian two-sample  $t$ -test. The left input panel offers the analysis options, including the specification of the alternative hypothesis and the selection of plots. The right output panel shows the corresponding analysis output. The prior and posterior plot is explained in more detail in Figure 5b. The input panel specifies the one-sided analysis for hypothesis testing; a two-sided analysis for estimation can be obtained by selecting “Group 1  $\neq$  Group 2” under “Hypothesis”.

effect is present or absent. Then, we discuss the scenario of parameter estimation, where the goal is to estimate the size of the population effect, assuming it is present. When both hypothesis testing and estimation procedures have been executed, there is no predetermined order for their interpretation. One may adhere to the adage “only estimate something when there is something to be estimated” (Wagenmakers, Marsman, et al., 2018) and first test whether an effect is present, and then estimate its size (assuming the test provided sufficiently strong evidence against the null), or to first estimate the magnitude of an effect, and then test if this magnitude is sufficiently big/small to accept/reject the alternative hypothesis.

If the goal of the analysis is hypothesis testing, we recommend using the Bayes factor. As described in Box 1, the Bayes factor can be seen as a weighing of the predictive quality of one hypothesis relative to that of another (Wagenmakers et al., 2016; see Box 1). Importantly, the Bayes factor is a *relative* metric of the hypotheses’ predictive quality. For instance, if  $BF_{10} = 5$ , this means that the data are 5 times as likely under  $\mathcal{H}_1$  than under  $\mathcal{H}_0$ . However, a Bayes factor in favor of  $\mathcal{H}_1$  does not mean that  $\mathcal{H}_1$  predicts the data well. As Figure 1 illustrates,  $\mathcal{H}_1$  provides a dreadful account of three out of four data sets due to violated assumptions, yet is still supported relative to  $\mathcal{H}_0$ .

Although there exists no hard bound for accepting or rejecting a hypothesis based on the Bayes factor, there have been some attempts to classify the strength of evidence that different Bayes factors represent, to make interpretation easier (e.g., Jeffreys, 1939; Kass & Raftery, 1995). One such classification scheme is shown in Figure 4. Several magnitudes of the Bayes factor are visualized as a probability wheel, where the proportion of red to white is determined by the degree of evidence in favor of  $\mathcal{H}_0$  and  $\mathcal{H}_1$ .<sup>2</sup> In line with Jeffreys, a Bayes factor

---

<sup>2</sup>Specifically, the proportion of red is the posterior probability of  $\mathcal{H}_1$  under a prior probability of 0.5; for a more detailed explanation and a cartoon see <https://tinyurl.com/ydhfndxa>

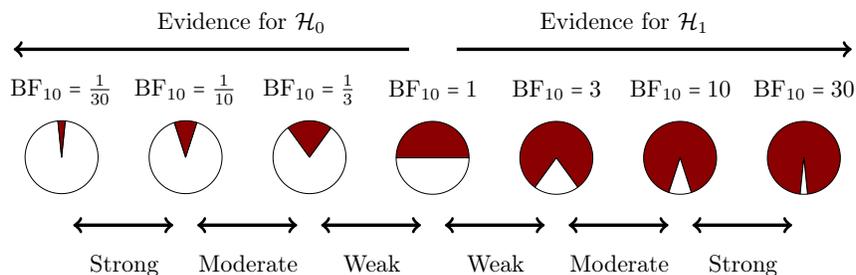


Figure 4: A graphical representation of a Bayes factor classification table. As the Bayes factor deviates from 1, which indicates equal support for  $\mathcal{H}_0$  and  $\mathcal{H}_1$ , more support is gained for either  $\mathcal{H}_0$  or  $\mathcal{H}_1$ . Bayes factors between 1 and 3 are considered to be weak (i.e., not worth more than a bare mention), Bayes factors between 3 and 10 are considered moderate, and Bayes factors greater than 10 are considered strong evidence. The Bayes factors are also represented as probability wheels, where the ratio of white (i.e., support for  $\mathcal{H}_0$ ) to red (i.e., support for  $\mathcal{H}_1$ ) surface is a function of the Bayes factor. The probability wheels further underscore the continuous scale of evidence that Bayes factors represent. We stress that these classifications are merely suggestive and should not be misused as an absolute rule for all-or-nothing conclusions.

between 1 and 3 is considered not worth more than a bare mention, a Bayes factor between 3 and 10 is considered moderate evidence, and a Bayes factor greater than 10 is considered strong evidence. However, these classifications should only be used as general rules of thumb to facilitate communication and interpretation of evidential strength. One of the merits of the Bayes factor is that it offers an assessment of evidence on a continuous scale.

When the goal of the analysis is parameter estimation, the posterior distribution is key (see Box 2). The posterior distribution is often summarized by a location parameter (point estimate) and uncertainty measure (interval estimate). For point estimation, the posterior median (reported by JASP), mean, or mode can be reported, although these do not contain any information about the uncertainty of the estimate. In order to capture the uncertainty of the estimate, an  $x\%$  credible interval can be reported. The credible interval, in short, has an  $x\%$  probability that the true parameter lies in this interval (an interpretation

that is often wrongly attributed to frequentist confidence intervals, see Morey et al., 2016). For example, if we obtain a 95% credible interval of  $[-1, 0.5]$  for effect size  $\delta$ , we can be 95% sure that the true value of  $\delta$  lies between  $-1$  and  $0.5$ , assuming that the alternative hypothesis we specify is true.

## Common Pitfalls in Interpreting Bayesian Results

Bayesian veterans tend to argue that Bayesian concepts are intuitive and easy to grasp. However, in our experience there exist persistent misinterpretations of Bayesian results. Here we list five:

- The Bayes factor does not equal the posterior odds; in fact, the posterior odds are equal to the Bayes factor multiplied by the prior odds (see also Equation 1). These prior odds reflect the relative plausibility of the rival hypotheses before seeing the data (e.g., 50/50 when both hypotheses are equally plausible, or 80/20 when one hypothesis is deemed to be 4 times more plausible than the other). For instance, a proponent and a skeptic may differ greatly in their assessment of the prior plausibility of a hypothesis; their prior odds differ, and, consequently, so will their posterior odds. However, as the Bayes factor is the updating factor from prior odds to posterior odds, proponent and skeptic ought to change their beliefs to the same degree (assuming they agree on the parameter prior distributions).
- Prior model probabilities (i.e., prior odds) and parameter prior distributions fulfill different roles. The former concerns prior beliefs about the hypotheses, for instance that both  $\mathcal{H}_0$  and  $\mathcal{H}_1$  are equally plausible a priori. The latter concerns prior beliefs about the model parameters within a model, for instance that all values of Pearson's  $\rho$  are equally likely a priori (i.e., a uniform prior distribution on the correlation parameter).
- The Bayes factor and credible interval have different purposes and can

yield different conclusions. Specifically, the credible interval is conditional on  $\mathcal{H}_1$  being true and quantifies the strength of an effect, assuming it is present; in contrast, the Bayes factor quantifies evidence for the presence or absence of an effect. A common misconception is only inspecting credible intervals for hypothesis testing. This is summed up by Berger (2006, p.383): “[...] Bayesians cannot test precise hypotheses using confidence intervals. In classical statistics one frequently sees testing done by forming a confidence region for the parameter, and then rejecting a null value of the parameter if it does not lie in the confidence region. This is simply wrong if done in a Bayesian formulation (and if the null value of the parameter is believable as a hypothesis).”

- The strength of evidence in the data is easy to overstate: a Bayes factor of 3 provides some support for one hypothesis over another, but should not warrant the confident all-or-none acceptance of that hypothesis.
- The results of an analysis always depend on the questions that were asked.<sup>3</sup> For instance, choosing a one-sided analysis over a two-sided analysis will often impact both the Bayes factor and credible interval. For an illustration of this, see Figure 5 for a comparison between one-sided and a two-sided results.

In order to avoid these and other pitfalls, we recommend that researchers who are doubtful about the correct interpretation of their Bayesian results solicit expert advice (for instance through the JASP forum at <http://forum.cogsci.nl>).

---

<sup>3</sup>This is known as Jeffreys’s platitude: “The most beneficial result that I can hope for as a consequence of this work is that more attention will be paid to the precise statement of the alternatives involved in the questions asked. It is sometimes considered a paradox that the answer depends not only on the observations but on the question; it should be a platitude” (Jeffreys, 1939, p.vi).

## Stereogram Example

For hypothesis testing, the results of the one-sided  $t$ -test are presented in Figure 5a. The resulting  $\text{BF}_{+0}$  is 4.567, indicating moderate evidence in favor of  $\mathcal{H}_+$ : the data are approximately 4.6 times more likely under  $\mathcal{H}_+$  than under  $\mathcal{H}_0$ . Due to potential non-normality of the dependent variable, we also conducted a Bayesian Mann-Whitney test (van Doorn et al., 2018). The resulting  $\text{BF}_{+0}$  is 6.038, which does not yield qualitatively different conclusions from the parametric test. The Bayesian Mann-Whitney test results are in the `.jasp` file at <https://osf.io/nw49j/>.

For parameter estimation, the results of the two-sided  $t$ -test are presented in Figure 5b. The 95% central credible interval for  $\delta$  is relatively wide, ranging from 0.046 to 0.904: this means that, under the assumption that the effect exists, we can be 95% certain that the true value of  $\delta$  lies between 0.046 to 0.904. In conclusion, there is moderate evidence for the presence of an effect, and large uncertainty about its size.

## Stage 4: Reporting the Results

For increased transparency, and to allow a skeptical assessment of the statistical claims, we advise to present an elaborate analysis report including relevant tables, figures, and background information. The extent to which this needs to be done in the manuscript itself depends on context. Ideally, an annotated `.jasp` file is created that presents the full results and analysis settings. The resulting file can then be uploaded to the Open Science Framework (OSF; <https://osf.io>), where it can be viewed by collaborators and peers, even without having JASP installed. Note that the `.jasp` file retains the settings that were used to create the reported output. Analyses not conducted in JASP

should mimic such transparency, for instance through uploading an R-script. In this section, we list several desiderata for reporting, both for hypothesis testing and parameter estimation.

In all cases, we recommend to provide a complete description of the prior specification (i.e., the type of distribution and its parameter values) and, especially for informed priors, to provide a justification for the choices that were made. When reporting a specific analysis, we advise to refer to the relevant background literature for details. In JASP, the relevant references for specific tests can be copied from the drop-down menus in the results panel.

When the goal of the analysis is hypothesis testing, it is key to outline which hypotheses are compared by clearly stating each hypothesis and including the subscript in the Bayes factor notation. Furthermore, we recommend to include, if available, the Bayes factor robustness check discussed in the section on planning (see Figure 6 for an example). This check provides an assessment of the robustness of the Bayes factor to different hyperparameter settings: if the qualitative conclusions do not change across a range of different plausible hyperparameter settings, this indicates that the analysis is robust to different values of the hyperparameter. If this plot is unavailable, the robustness of the Bayes factor can be checked manually by specifying different hyperparameter values (see the mixed ANOVA analysis in the online appendix at <https://osf.io/wae57/> for an example). Lastly, a sequential Bayes factor plot can be considered. This plot shows how the Bayes factor has developed as observations came in, by plotting the Bayes factor as a function of the sample size.

When the goal of the analysis is parameter estimation, it is key to report the summary statistics of the posterior distribution, such as a 95% credible interval and the median. Ideally, the results of the analysis are reported both graphically and numerically. This means that, when possible, a plot is presented

that includes the posterior distribution, prior distribution, Bayes factor, 95% credible interval, and posterior median. Such a figure is often provided by JASP, and will be discussed below.

Numeric results can be presented either in a table or in the main text. If relevant, we recommend to report the results from both estimation and hypothesis test. For some analyses, the results are based on a numerical algorithm (e.g., Markov chain Monte Carlo sampling), which yields an error percentage. If applicable and available, the error percentage ought to be reported too, to indicate the numeric robustness of the result. Lower values of the error percentage indicate greater numerical stability of the result.

## Stereogram Example

This is an example report of the stereograms *t*-test example:

Here we summarize the results of the Bayesian analysis for the stereogram viewing times data. For this analysis we used the Bayesian *t*-test framework proposed by Jeffreys (1961, see also Rouder et al. 2009). We analyzed the data with JASP (JASP Team, 2018). An annotated `.jasp` file, including distribution plots, data, and input options, is available at <https://osf.io/25ekj/>. First, we discuss the results for hypothesis testing. The null hypothesis states that there is no difference in fuse time between the groups and therefore  $\mathcal{H}_0 : \delta = 0$ . The one-sided alternative hypothesis states that only positive values of  $\delta$  are possible, and assigns more prior mass to values closer to 0 than extreme values. Specifically,  $\delta$  was assigned a Cauchy prior distribution with  $r = 1/\sqrt{2}$ , truncated to allow only positive effect size values. Figure 5a shows that the Bayes factor indicates evidence for  $\mathcal{H}_+$ ; specifically,  $\text{BF}_{+0} = 4.567$ , which means

that the data are approximately 4.5 times more likely to occur under  $\mathcal{H}_+$  than under  $\mathcal{H}_0$ . This result indicates moderate evidence in favor of  $\mathcal{H}_+$ . The error percentage is  $< 0.001\%$ , which indicates great stability of the numerical algorithm that was used to obtain the result. In order to assess the robustness of the Bayes factor to our prior specification, Figure 6 shows  $\text{BF}_{+0}$  as a function of the hyperparameter  $r$ . Across a wide range of hyperparameter specifications, the Bayes factor appears to be relatively stable, ranging from 3 to 5.

Second, we discuss the results for parameter estimation. Of interest is the posterior distribution of the standardized effect size  $\delta$  (i.e., Cohen’s  $d$ , the standardized difference in mean fuse times). For parameter estimation,  $\delta$  was assigned a Cauchy prior distribution with  $r = 1/\sqrt{2}$ . Figure 5b shows that the resulting posterior distribution peaks at  $\delta = 0.47$  (the posterior median) with a central 95% credible interval for  $\delta$  that ranges from 0.046 to 0.904. If the effect is assumed to exist, there remains substantial uncertainty about its size, with values close to 0 having the same posterior density as values close to 1.

## Limitations and Challenges

The Bayesian toolkit for the empirical social scientist is still a work in progress, and this means there are some limitations to overcome. First, for some frequentist analyses, the Bayesian counterpart has not yet been developed or implemented in JASP. Secondly, some analyses in JASP currently provide only a Bayes factor, and not a visual representation of the posterior distributions, for instance due to the multidimensional parameter space of the model. Thirdly, some analyses in JASP are only available with a relatively limited set of prior

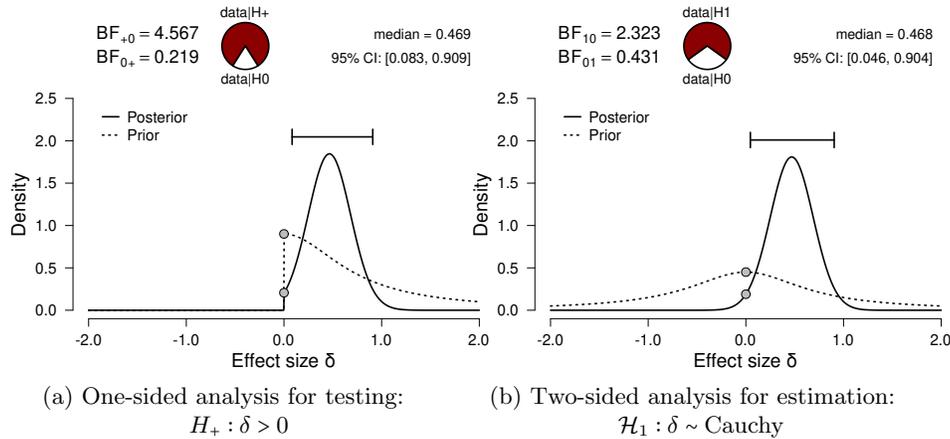


Figure 5: Bayesian two-sample  $t$ -test for the parameter  $\delta$ . The probability wheel at the top illustrates the ratio of the evidence in favor of the two hypotheses. The two gray dots indicate the prior and posterior density at the test value - the ratio of these is the Savage-Dickey density ratio (Dickey & Lientz, 1970; Wagenmakers et al., 2010). The median and the 95% central credible interval of the posterior distribution are shown in the top right corner. The left panel shows the one-sided model for hypothesis testing and the right panel shows the two-sided model for parameter estimation. Both figures from JASP.

distributions. However, these are not principled limitations and it is only a matter of time before they are overcome. When dealing with more complex models that go beyond the “staple” analyses such as  $t$ -tests, there exist a number of software packages that are aimed at coding custom models, such as JAGS (Plummer, 2003) or Stan (Carpenter et al., 2017). Another option for Bayesian inference is to code the analyses in a programming language such as R (R Core Team, 2018) or Python (van Rossum, 1995). This requires a certain degree of programming ability, but grants the user more flexibility. Popular packages for conducting Bayesian analyses in R are the BayesFactor package (Morey & Rouder, 2015) and the brms package (Bürkner, 2017), among others (see <https://cran.r-project.org/web/views/Bayesian.html> for a more exhaustive list). For Python, a popular package for Bayesian analyses is the PyMC3 package (Salvatier et al., 2016).

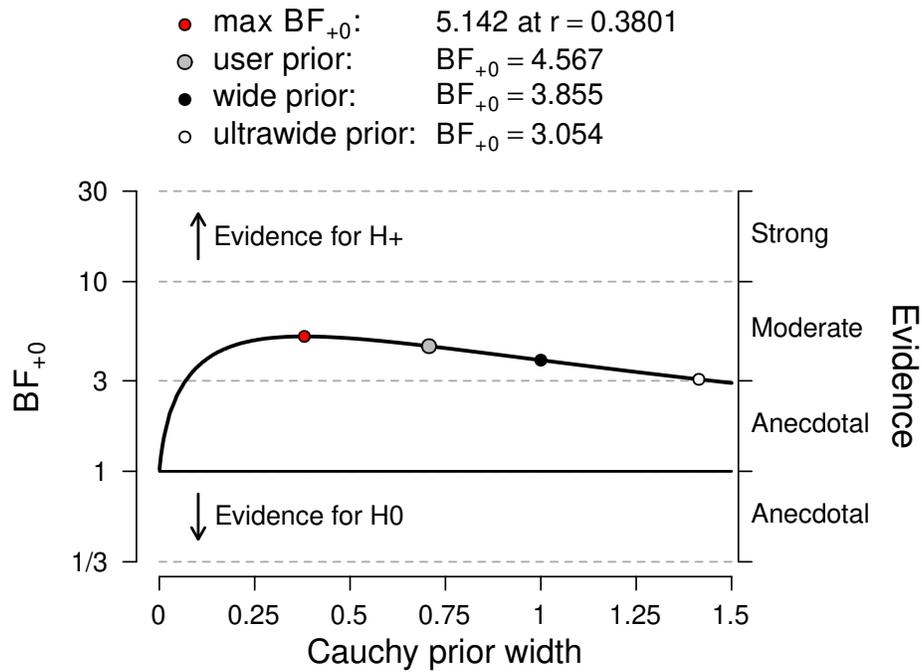


Figure 6: The Bayes factor robustness plot. The maximum  $BF_{+0}$  is attained when setting the hyperparameter  $r$  to 0.38. The plot indicates  $BF_{+0}$  for the user specified prior ( $r = 1/\sqrt{2}$ ), wide prior ( $r = 1$ ), and ultrawide prior ( $r = \sqrt{2}$ ). The evidence for the alternative hypothesis is relatively stable across a wide range of hyperparameter specifications. This suggests that the analysis is robust. However, the evidence in favor of  $\mathcal{H}_+$  is not particularly strong and will not convince a skeptic.

## Concluding Comments

We have attempted to provide concise recommendations for planning, executing, interpreting, and reporting Bayesian analyses. These recommendations are listed in Table 1. Our guidelines focused on the standard analyses that are currently featured in JASP. When going beyond these analyses, some of the discussed guidelines will be easier to implement than others. However, the general process of transparent, comprehensive, and careful statistical reporting extends to all Bayesian procedures and indeed to statistical analyses across the board.

Stage	Recommendation
Planning	<ul style="list-style-type: none"> <li>Write the methods section in advance of data collection</li> <li>Preregister the analysis plan for increased transparency</li> <li>Distinguish between exploratory and confirmatory research</li> <li>Specify the goal; estimation, testing, or both</li> <li>If the goal is testing, decide on one-sided or two-sided procedure</li> <li>Choose a statistical model</li> <li>Choose a prior distribution</li> <li>Specify the sampling plan</li> <li>Consider a Bayesian design analysis</li> </ul>
Executing	<ul style="list-style-type: none"> <li>Check the quality of the data (e.g., assumption checks)</li> <li>Annotate the JASP output</li> </ul>
Interpreting	<ul style="list-style-type: none"> <li>Beware of the common pitfalls</li> <li>Use the correct interpretation of Bayes factor and credible interval</li> <li>When in doubt, ask for advice (e.g., on the JASP forum)</li> </ul>
Reporting	<ul style="list-style-type: none"> <li>Mention the goal of the analysis</li> <li>Include a plot of the prior and posterior distribution, if available</li> <li>If testing, report the Bayes factor, including its subscripts</li> <li>If estimating, report the posterior median and <math>x\%</math> credible interval</li> <li>Include which prior settings were used</li> <li>Justify the prior settings (particularly for informed priors in a testing scenario)</li> <li>Discuss the robustness of the result</li> <li>If relevant, report the results from both estimation and hypothesis testing</li> <li>Refer to the statistical literature for details about the analyses used</li> <li>Consider a sequential analysis</li> <li>Make the <code>.jasp</code> file and data available online</li> </ul>

Table 1: A summary of the guidelines for the different stages of a Bayesian analysis, with a focus on analyses conducted in JASP.

## **Author Contributions**

JvD wrote the main manuscript and EJW contributed to manuscript revisions.

All authors reviewed the manuscript and provided feedback.

## References

- Andrews, M., & Baguley, T. (2013). Prior approval: The growth of Bayesian methods in psychology. *British Journal of Mathematical and Statistical Psychology*, *66*, 1–7.
- Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*, 17–21.
- Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. *American Psychologist*, *73*, 3–25.
- Berger, J. O. (2006). Bayes factors. In S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 1 (2nd ed.) (pp. 378–386). Hoboken, NJ: Wiley.
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28.
- Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*, 1–37.
- Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex*, *49*, 609–610.
- De Groot, A. D. (1956/2014). The meaning of “significance” for different types of research. Translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. van der Maas. *Acta Psychologica*, *148*, 188–194.

- Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-checklist. *Psychological Methods*, *22*, 240–261.
- Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*, 214–226.
- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*:781.
- Dienes, Z., & McLatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review*, *25*, 207–218.
- Etz, A., & Wagenmakers, E.-J. (2017). J. B. S. Haldane’s contribution to the Bayes factor hypothesis test. *Statistical Science*, *32*, 313–329.
- Frisby, J. P., & Clatworthy, J. L. (1975). Learning to see complex random-dot stereograms. *Perception*, *4*, 173–178.
- Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2018). Informed Bayesian t-tests. Manuscript submitted for publication. *arXiv preprint arXiv:1704.02479*.
- Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *Journal of Problem Solving*, *7*, 2-9.
- JASP Team. (2018). *JASP (Version 0.9.2)[Computer software]*. Retrieved from <https://jasp-stats.org/>
- Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford, UK: Oxford University Press.
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.
- Lee, M. D., & Vanpaemel, W. (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review*, *25*, 114–127.
- Liang, F., German, R. P., Clyde, A., & Berger, J. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*, 410–424.
- Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016). Harold Jeffreys’s default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32.
- Marsman, M., & Wagenmakers, E.-J. (2017). Bayesian benefits with JASP. *European Journal of Developmental Psychology*, *14*, 545–555.
- Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E.-J. (2015). The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. *Journal of Experimental Psychology: General*, *144*, e1–e15.
- Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*, 103–123.
- Morey, R. D., & Rouder, J. N. (2015). *BayesFactor 0.9.11-1*. Comprehensive R Archive Network. Retrieved from <http://cran.r-project.org/web/packages/BayesFactor/index.html>
- Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.),

*Proceedings of the 3rd international workshop on distributed statistical computing.* Vienna, Austria.

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <https://www.R-project.org/>

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

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.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, *2*, e55.

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

Spiegelhalter, D. J., Myles, J. P., Jones, D. R., & Abrams, K. R. (2000). Bayesian methods in health technology assessment: A review. *Health Technology Assessment*, *4*, 1–130.

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). A tutorial on Bayes factor design analysis with informed priors. *Manuscript submitted for publication*.

Sung, L., Hayden, J., Greenberg, M. L., Koren, G., Feldman, B. M., & Tomlinson, G. A. (2005). Seven items were identified for inclusion when reporting a Bayesian analysis of a clinical study. *Journal of Clinical Epidemiology*, *58*, 261–268.

- The BaSiS group. (2001). *Bayesian standards in science: Standards for reporting of Bayesian analyses in the scientific literature*. Internet. Retrieved from <http://lib.stat.cmu.edu/bayesworkshop/2001/BaSis.html>
- Tijmstra, J. (2018). Why checking model assumptions using null hypothesis significance tests does not suffice: A plea for plausibility. *Psychonomic Bulletin & Review*, *25*, 548–559.
- Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (Eds.). (2018). Beyond the new statistics: Bayesian inference for psychology [special issue]. *Psychonomic Bulletin & Review*, *25*.
- van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2018). Bayesian latent-normal inference for the rank sum test, the signed rank test, and Spearman’s rho. *arXiv preprint arXiv:1712.06941*.
- van Rossum, G. (1995). *Python tutorial* (Tech. Rep. No. CS-R9526). Amsterdam: Centrum voor Wiskunde en Informatica (CWI).
- Wagenmakers, E.-J., Beek, T., Rotteveel, M., Gierholz, A., Matzke, D., Steingroever, H., ... Pinto, Y. (2015). Turning the hands of time again: A purely confirmatory replication study and a Bayesian analysis. *Frontiers in Psychology: Cognition*, *6*:494.
- 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.
- Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., ... Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*, 58–76.

- Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ...  
Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical  
advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*,  
35–57.
- Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for  
the pragmatic researcher. *Current Directions in Psychological Science*, *25*,  
169–176.
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., &  
Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives  
on Psychological Science*, *7*, 627–633.
- Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific  
inquiry. *Philosophical Magazine*, *42*, 369–390.