Bayes Factor and Its Implementation in JASP

Authors: Hu Chuanpeng, Kong Xiangzhen, Eric-Jan Wagenmakers is a scholar with significant influence in the fields of statistics and psychology. He currently serves as a professor in the Department of Psychological Methods at the University of Amsterdam and is the head of the "Psychological Methods" research group in the department. Wagenmakers' research primarily focuses on Bayesian statistics, hypothesis testing, the development of statistical software, and the application of these methods to empirical research in psychology and cognitive science. He is particularly renowned for his critique of Null Hypothesis Significance Testing (NHST) and his advocacy for Bayesian methods, which has sparked widespread discussion and reflection in the academic community.

Wagenmakers' critique of NHST primarily centers on its misuse and misinterpretation in scientific research. He points out that p-values do not provide direct evidence regarding the probability of a hypothesis, yet researchers often erroneously interpret $p < 0.05$ as the probability of the hypothesis being true being less than 5%. This misinterpretation has contributed to the reproducibility crisis in scientific research, particularly in psychology. To address this issue, Wagenmakers advocates for the use of Bayesian statistical methods, particularly Bayes factors, as an alternative to NHST. Bayes factors can quantify the relative degree of support that data provide for competing hypotheses, thereby offering more intuitive and meaningful statistical inference.

In promoting Bayesian statistics, Wagenmakers goes beyond theoretical advocacy and actively participates in the development of statistical software to lower the barrier for researchers to apply Bayesian methods. His team has developed JASP (Jeffreys's Amazing Statistics Program), a free, open-source statistical software that provides a user-friendly graphical interface and supports both Bayesian statistical analysis and traditional frequentist statistical analysis. The design goal of JASP is to enable researchers to easily conduct Bayesian analyses without writing complex code. The software has gained widespread attention and application, becoming one of the preferred tools for many psychology researchers to perform statistical analysis.

Beyond software development and theoretical advocacy, Wagenmakers is also committed to demonstrating the advantages of Bayesian methods through empirical research. He has participated in numerous reproducibility research projects, such as the analysis of the "Reproducibility Project: Psychology." In these studies, he reanalyzed original data using Bayes factors and found

that many effects considered significant under the NHST framework actually received weak evidentiary support in Bayesian analysis. These studies further highlight the limitations of NHST and provide empirical support for the application of Bayesian methods in psychological research.

Wagenmakers' work extends beyond the field of psychology, and his influence has spread to other disciplines such as neuroscience, medicine, and social sciences. His papers and tutorials have helped many researchers understand and apply Bayesian statistics, promoting statistical practice reform in these fields. Although his views remain controversial in the academic community, there is no denying that Wagenmakers has made important contributions to the practice and development of modern statistics, particularly in promoting the application of Bayesian statistics in the social sciences., Alexander Ly, Peng Kaiping, Chuanpeng Hu

Date: 2018-05-08T19:47:49+00:00

Abstract

Statistical inference plays a crucial role in scientific research; however, the most commonly used classical statistical method in current research—null hypothesis significance testing (NHST)—is misused or abused by some researchers due to its difficulty of understanding. Some researchers have proposed using the Bayes factor as an alternative and/or supplementary statistical method. The Bayes factor is an important method in Bayesian statistics for model comparison and hypothesis testing, which can be interpreted as the degree of support for the null hypothesis H0 or the alternative hypothesis H1. Compared with NHST, it has the following advantages: it simultaneously considers H0 and H1 and can be used to support H0, it does not severely bias against H0, it can monitor changes in the strength of evidence, and it is not affected by sampling plans. Currently, Bayes factors can be conveniently implemented through the open statistical software JASP; this paper uses Bayesian t-tests as a demonstration. The use of Bayes factors is of great significance to psychology researchers, but attention must be paid to the rationality of prior distribution selection and to maintaining transparency and openness in the data analysis process.

Full Text

The Bayes Factor and Its Implementation in JASP: A Practical Primer

Chuan-Peng HU^1,2,; ***Xiang-Zhen KONG***^*3*; ***Eric-Jan WAGENMAKERS***^*4*; ***Alexander LY***^*4*; ***Kaiping PENG***^*1*,

^1 Department of Psychology, School of Social Sciences, Tsinghua University, Beijing, China, 100084
^3 Language and Genetics Department, Max Planck Institute for Psycholinguistics, 6500 AH Nijmegen, The Netherlands

^4 Department of Psychological Methods, University of Amsterdam, 1018 VZ Amsterdam, The Netherlands

Abstract: Statistical inference plays a critical role in scientific research, yet the most commonly used classical statistical method—null hypothesis significance testing (NHST)—is often misused or abused by researchers due to its difficulty in comprehension. Some researchers have proposed using the Bayes factor as an alternative and/or supplementary statistical method. The Bayes factor is an important method in Bayesian statistics for model comparison and hypothesis testing, which can be interpreted as the degree of support for the null hypothesis H0 or the alternative hypothesis H1. Compared with NHST, it has the following advantages: it considers both H0 and H1 simultaneously and can be used to support H0, it does not "severely" bias against H0, it can monitor changes in evidence strength, and it is not influenced by sampling plans. Currently, Bayes factors can be conveniently implemented through the open-source statistical software JASP, and this article provides a demonstration using Bayesian t-tests. The use of Bayes factors is of significant importance for psychology researchers, but attention must be paid to the reasonableness of prior distribution selection and to maintaining transparency and openness in the data analysis process.

Keywords: Bayes factor, Bayesian statistics, Frequentist, hypothesis testing, JASP

Since the 20th century, statistical inference has played an increasingly important role in scientific research, and the validity of scientific conclusions has become increasingly dependent on the correct application of statistical inference. Currently, the most widely used statistical inference method is null hypothesis significance testing (NHST). However, alongside the widespread use of NHST across various fields, researchers' misunderstanding and blind use of NHST and p-values have led to negative consequences. For example, p-values have been used to support unreasonable and irreproducible research findings (e.g., Bem, 2011), sparking debates about whether NHST is suitable for scientific research (Miller, 2011). Against this background, some researchers have recommended using Bayes factors as an alternative to NHST (Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011; Zhong, Dienes, Z., & Chen, 2017).

The Bayes factor is a method in Bayesian statistics used for model comparison and hypothesis testing. In hypothesis testing, it represents the ratio of the strength of support that current data provides for the null hypothesis versus the alternative hypothesis. As will be detailed in the next section, Bayes factors can quantitatively reflect the degree of support that current data provides for each hypothesis, making them potentially more suitable for hypothesis testing in scientific research. However, due to the relatively complex statistical principles and implementation of Bayes factors, they have not been widely applied across various disciplines.

In recent years, with substantial improvements in computational power, Bayesian statistics has achieved tremendous success in computer science and

other fields (Zhu, Chen, Hu, & Zhang, 2017). Bayesian statistical tools have developed rapidly, such as WinBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009), JAGS (Plummer, 2003), Stan (Carpenter et al., 2017), and the Python package PyMC3 (http://docs.pymc.io/index.html). The emergence of these software packages and tools has promoted the use of Bayesian methods across various research fields (Depaoli & van de Schoot, 2017; van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017). Among these tools, some have been developed specifically for calculating Bayes factors, such as the BayesFactor package in R (http://bayesfactorpcl.r-forge.r-project.org/). In psychology and related fields, many researchers have recently attempted to introduce Bayesian statistical methods (Dienes, 2008, 2011, 2014; Hoijtink, 2011; Klugkist, Laudy, & Hoijtink, 2005; Kruschke, 2014; Masson, 2011; Morey & Rouder, 2011; Mulder et al., 2009; Rouder, Morey, Speckman, & Province, 2012; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Vanpaemel, 2010; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). Against the backdrop of a "replication crisis" in psychology (Open Science Collaboration, 2015; Hu et al., 2016) and neuroimaging research (Chen, Lu, & Yan, 2018; Zuo & Xing, 2014), using reasonable statistical methods has become more urgent. However, for many psychology researchers, using R or other programming languages to calculate Bayes factors remains difficult. To address this obstacle, researchers have developed JASP (https://jasp-stats.org/, JASP team 2017) (JASP Team, 2017; Marsman & Wagenmakers, 2016a; Wagenmakers, Love, et al., 2017; Wagenmakers, Marsman, et al., 2017), a statistical tool with a graphical interface similar to commercial software SPSS, which simplifies the calculation of Bayes factors.

This article aims to introduce Bayes factors and their use to psychology researchers and those in related disciplines. First, we introduce the principles of Bayes factors and their advantages over p-values in traditional hypothesis testing. Then, using independent samples t-tests as an example, we explain how to use JASP to calculate Bayes factors and how to interpret and report the results. Based on this, we discuss the application value and limitations of Bayes factors.

1 Principles of the Bayes Factor

The Bayes factor is an application of Bayesian statistics to hypothesis testing. Therefore, to understand Bayes factors, one must first understand the principles of Bayesian statistics.

1.1 Introduction to Bayesian Statistics

Bayesian statistics and Frequentist statistics are the two main schools of statistics, and their core difference lies in their different interpretations of what probability represents. For Frequentists, probability is the expected value of frequency in infinite repetitions of sampling. In contrast, Bayesians believe that probability represents the degree of belief in an event, with values from 0 to

1 indicating how much one believes an event is true based on available information. Since different people may have different degrees of belief in the same event, Bayesian probability is subjective. However, Bayesian probability is not arbitrary: through reasonable methods of continuously acquiring and updating known information, subjectivity can eventually be eliminated, leading to consensus.

Because Frequentists view probability as the result of long-run behavior, understanding Frequentist probability typically requires imagining events that have not yet occurred. For example, within the NHST framework, the meaning of a p-value is the probability of obtaining the current result and more extreme results, assuming H0 is true. In other words, the p-value expresses: if H0 were true and we could repeat the current experiment infinitely many times under identical conditions, what proportion of these experiments would produce the current result pattern or more extreme patterns? Therefore, the meaning of p-values implicitly contains an important assumption: we can repeat the experiment infinitely many times. However, researchers often ignore this assumption of infinite repetitions and mistakenly believe that the p-value is the probability of making an error when rejecting the null hypothesis in a single test (Greenland et al., 2016). This misunderstanding of NHST actually has a Bayesian flavor, as it calculates the probability that a model is correct or incorrect based on current data.

Unlike Frequentist statistics, one of the most important characteristics of Bayesian statistics is that it considers the credibility of different possibilities for individuals (Kruschke, 2014). Through continuously obtained data, people can change their corresponding beliefs about different possibilities. This way of thinking is very similar to people's daily life experiences: when we continuously obtain evidence supporting a certain viewpoint, we become more convinced of that viewpoint.

Although Bayesian statistics has a different understanding of probability from Frequentist statistics, its calculation of probability strictly follows the basic principles of probability: the addition rule and multiplication rule. The core Bayesian rule in Bayesian statistics is also derived from these simple addition and multiplication principles. According to the multiplication rule of probability, the probability of random events A and B occurring simultaneously is:

$$p(A \cap B) = p(A|B) \times p(B) = p(B|A) \times p(A)$$

Formula 1 is the joint probability formula, representing the probability of both A and B occurring. Its meaning is: the joint probability of A and B (p(A B)) equals the product of the probability of A occurring given that B has occurred (p(A|B)) and the probability of B occurring (p(B)), and also equals the product of the probability of B occurring given that A has occurred (p(B|A)) and the probability of A occurring (p(A)). Here, p(A|B) and p(B|A) are both conditional probabilities, but they have different meanings.

Transforming Formula 1 yields the following formula:

$$p(A|B) = \frac{p(A \cap B)}{p(B)} = \frac{p(B|A) \times p(A)}{p(B)}$$

Formula 2 is Bayes' theorem. It represents that if we want to calculate the probability of A occurring given that B has occurred (p(A|B)), we can divide the probability of both A and B occurring (p(A B)) by the probability of B occurring (p(B)), which equals the product of the probability of B occurring given that A has occurred and the probability of A occurring, divided by the probability of B occurring. Formula 2 connects two conditional probabilities, making it possible to calculate different conditional probabilities.

Within the Bayesian statistical framework, Formula 2 can be viewed as an information update. Suppose we need to test the probability that a theoretical model is true based on data collected from an experiment. Using the commonly used null hypothesis H0 in psychological research as an example, Formula 2 can be rewritten as:

$$p(H_0|\text{data}) = \frac{p(\text{data}|H_0) \times p(H_0)}{p(\text{data})}$$

p(H0|data) represents the probability that theoretical model H0 is correct after data updating, i.e., the posterior probability; p(H0) represents the probability that theoretical model H0 is correct before data updating, i.e., the prior probability; and p(data|H0) is the probability of obtaining the current data under model H0, i.e., the marginal likelihood. This shows that in Bayesian statistics, the main function of data collection (experiment) is to help us update the credibility of theoretical models.

According to Formula 3, we can use data to update the probability of any model being true. In hypothesis testing, we can simultaneously update the credibility of the null hypothesis (theoretical model H0) and the alternative hypothesis (theoretical model H1) based on observed data (see Formulas 3 and 4, respectively) to obtain their updated posterior probabilities.

$$p(H_1|\text{data}) = \frac{p(\text{data}|H_1) \times p(H_1)}{p(\text{data})}$$

After obtaining the posterior probabilities of H0 and H1, we can compare them, as shown in Formula 5:

$$\frac{p(H_1|\text{data})}{p(H_0|\text{data})} = \frac{p(\text{data}|H_1)}{p(\text{data}|H_0)} \times \frac{p(H_1)}{p(H_0)}$$

Where the Bayes factor is:

$$\mathrm{BF}_{10} = \frac{p(\text{data}|H_1)}{p(\text{data}|H_0)}$$

In Formula 6, the subscript 1 in BF10 represents H1, and 0 represents H0. Therefore, BF10 represents the Bayes factor comparing H1 to H0, while BF01 represents the Bayes factor comparing H0 to H1. For example, BF10 = 19 means that the likelihood of obtaining the current data under the alternative hypothesis H1 is 19 times that under the null hypothesis H0. From this definitional formula, we can see that the Bayes factor reflects the change in updating the prior probability to the posterior probability based on current data.

As such, the Bayes factor answers different questions than NHST. NHST attempts to answer the question: "Assuming we know the relationship between two variables (e.g., no difference between two conditions), what is the probability of obtaining the current observed data pattern or more extreme patterns (p(more extreme > observed data | H0))?" In contrast, the Bayes factor attempts to answer: "Under which theoretical model is the current data more likely to occur?" In hypothesis testing, the Bayes factor has some advantages that NHST lacks (see Table 1), which will be detailed in the next subsection.

Table 1. Comparison of Bayesian Inference and Traditional NHST Inference in Hypothesis Testing

Note: 10 = Jeffreys (1935); 11 = Jeffreys (1961); 12 = Rouder, et al. (2009); 13 = Wagenmakers (2007); 14 = Edwards (1965); 15 = Berger and Delampady (1987); 16 = Sellke, Bayarri, and Berger (2001); 17 = Edwards, Lindman, and Savage (1963); 18 = Rouder (2014); 19 = Berger and Berry (1988); 20 = Lindley (1993).

Building on Jeffreys (1961), Wagenmakers et al. (2017) proposed a principled classification of Bayes factor magnitudes (see Table 2). However, this classification is only a rough reference and cannot be strictly applied; researchers need to judge the meaning of Bayes factors based on specific research contexts.

Table 2. Decision Criteria for Bayes Factors

Bayes Factor, BF10	Interpretation
> 100	Extremely strong evidence for H1
30 – 100	Very strong evidence for H1
10 – 30	Strong evidence for H1
3 – 10	Moderate evidence for H1
1 – 3	Weak evidence for H1
1/3 – 1	Weak evidence for H0
1/10 – 1/3	Moderate evidence for H0
1/30 – 1/10	Strong evidence for H0
1/100 – 1/30	Very strong evidence for H0

Bayes Factor, BF10	Interpretation
$< 1/100$	Extremely strong evidence for H0

1.2 Default Priors for the Alternative Hypothesis

Since prior probabilities play a crucial role in Bayes factors, selecting the prior distribution for the alternative hypothesis becomes particularly important. One reasonable approach is to set the prior distribution for the alternative hypothesis based on previous research findings on a topic (e.g., effect sizes obtained from meta-analyses). However, this approach is often unrealistic in many situations: first, the possible distribution of effect sizes varies depending on the paradigm; more importantly, since many studies are exploratory in nature, there are no previous research findings to guide the selection. Therefore, a more common practice is to use a comprehensive, standardized prior.

For example, in Bayesian t-tests, using the Cauchy distribution as the prior for the alternative hypothesis may be a reasonable choice (Jeffreys, 1961; Ly, Verhagen, & Wagenmakers, 2016a, 2016b; Rouder et al., 2009). Compared to the standard normal distribution, the Cauchy distribution has relatively smaller probability density near 0, thus allowing for more large effects than the standard normal distribution (see Figure 1 [Figure 1: see original paper]); compared to the uniform distribution (i.e., effect sizes are equally distributed across all values), the Cauchy distribution favors the null hypothesis more (Jeffreys, 1961; Rouder et al., 2009). Therefore, the prior distribution for the alternative hypothesis can be expressed as:

$$\delta \sim \text{Cauchy}(x_0 = 0, \gamma = 1)$$

Figure 1. Comparison of Cauchy Distribution and Normal Distribution

Jeffreys (1961) first proposed using the Cauchy distribution as a prior in Bayes factors to compare two-sample problems. Recent validation work by researchers has shown that the Cauchy distribution can be used as a prior for calculating Bayes factors commonly used in psychological research, such as t-tests (Rouder et al., 2009), ANOVA (Rouder et al., 2012), and correlation analysis (Ly, Marsman, & Wagenmakers, 2017; Ly et al., 2016b). This validation work has laid the foundation for the application of Bayes factors in psychology and related disciplines.

2 Advantages of Bayes Factors

As mentioned earlier, in hypothesis testing, Bayes factors not only align better with people's intuition but also possess some advantages that NHST lacks. These advantages can be summarized into five aspects (see Table 1). The following sections elaborate on these five aspects.

2.1 Simultaneous Consideration of H0 and H1

The calculation of Bayes factors simultaneously considers both H0 and H1 and updates the prior probabilities of H0 and H1 being true based on all available data, thereby comparing which theoretical model (H0 or H1) is more reasonable under the current data. This approach differs from NHST: under the NHST framework, calculating p-values only requires assuming H0 is true, while no assumptions are made about H1, making p-values independent of H1. The logic of NHST is that if the probability of observing the current data under the assumption that H0 is true is very small, then H0 is rejected and H1 is accepted. In this case, NHST ignores one possibility: under the current data, the probability that H1 is true may be comparable to or even smaller than the probability that H0 is true (Wagenmakers, Verhagen, et al., 2017). For example, in Bem (2011), H0 was that participants' responses were not influenced by future stimuli, while H1 was that future stimuli could affect participants' current responses, meaning participants could "precognize" stimuli that had not yet appeared. Although Bem (2011) obtained $p < 0.05$ using NHST logic, meaning the probability of obtaining the current data under H0 (p(data|H0)) was low, the author chose to reject H0 and accept H1, concluding that participants could predict future stimuli. However, researchers are more concerned with the probability that a model/hypothesis (e.g., H1) is true based on the current data (p(H1|data)), rather than the probability of obtaining the current data under the null hypothesis H0 (p(data|H0)). In Bem's (2011) study, prior knowledge tells us that the probability of H1 being true may be very low, and under the current data pattern, the likelihood of H1 being true (p(H1|data)) is likely much lower than the likelihood of H0 being true (p(H0|data)) (Rouder & Morey, 2011; Wagenmakers et al., 2011), but NHST completely ignores this point.

2.2 Can Be Used to Support H0

Similarly, because Bayes factors simultaneously quantify the strength of support that current data provides for both H0 and H1, they can be used to support H0 (Dienes, 2014). However, under the traditional NHST framework, hypothesis testing is conducted only under the assumption that H0 is true, and relying solely on a significance level (such as 0.05 or 0.005) cannot provide evidence for whether H0 is true. For instance, a hypothesis test result of $p = 0.2$ alone cannot be interpreted as evidence of no effect (unless combined with sample size, effect size, and statistical power for a comprehensive judgment).

In actual research, the ability to provide quantified evidence for H0 is of great significance (Gallistel, 2009; Rouder et al., 2009), as it can intuitively help researchers distinguish between two situations: evidence of absence (evidence that there is no effect) and absence of evidence (lack of evidence that there is an effect) (Dienes, 2014). Specifically, Bayes factor results have three possible states: (1) providing evidence supporting H1 (i.e., evidence of an effect); (2) providing evidence supporting H0 (i.e., evidence of no effect); or (3) supporting neither hypothesis (insufficient evidence to indicate whether there is or is not

an effect). For example, a Bayes factor BF01 = 15 indicates that the observed data is 15 times more likely to occur under H0 than under H1, suggesting that the current data more strongly support the null hypothesis of no effect. However, if BF01 = 1.5, this indicates that the observed data is only 1.5 times more likely to occur under H0 than under H1, suggesting that the current data provide roughly equivalent support for both hypotheses and there is insufficient evidence to support either H0 or H1 (see Table 2 for suggested interpretations of Bayes factor magnitudes).

It is worth noting that whether supporting H1 or H0, the evidence provided by Bayes factors is relative—that is, it supports one hypothesis relative to another. Therefore, there may exist a third model H2 that is closer to the true situation than both H1 and H0, with a higher posterior probability. It should be pointed out that some researchers have recently developed methods within the NHST framework that can accept the null hypothesis, such as equivalence testing. This method tests whether the effect size is not different from 0 by setting multiple H0s, thereby testing whether H0 can be accepted (Lakens, 2017). However, equivalence testing still uses p-values and cannot provide a direct measure of evidence (Schervish, 1996).

2.3 Not "Severely" Biased Against H0

Bayes factors simultaneously and separately quantify the strength of support that current data provides for H0 and H1. Compared with traditional NHST, their support for H0 and H1 is more balanced, and thus their tendency to reject H0 is relatively less strong.

Under traditional NHST assumptions, as long as researchers can collect enough data, they can always obtain $p < 0.05$ and reject H0. In contrast, Bayes factors tend to stabilize as data increase (see Section 3.2 for discussion on the convergence of Bayes factors). For the same data, p-values also appear to be stronger than Bayes factors in opposing H0. For example, one study analyzed the relationship between presidential candidates' height and election outcomes in U.S. presidential elections, finding $r = 0.39$, $p = 0.007$ after significance testing of the correlation coefficient (Stulp, Buunk, Verhulst, & Pollet, 2013). If using Bayesian factor analysis, BF10 = 6.33 would be obtained (Wagenmakers, Marsman, et al., 2017). Although both methods generally support the same conclusion (i.e., rejecting H0 and moderately supporting H1), the p-value seems to indicate strong evidence for rejecting H0, while the support obtained from the Bayes factor is more cautious. Wetzels et al. (2011) compared the results of 855 t-tests and found that although p-values and Bayes factors were generally consistent in the direction of their conclusions in most cases, Bayes factors were relatively more conservative: statistically significant results with p-values between 0.01 and 0.05 corresponded to Bayes factors indicating only very weak evidence. For a Bayesian interpretation of traditional p-values, see Johnson (2013) and Marsman & Wagenmakers (2016b).

2.4 Can Monitor Changes in Evidence Strength

When calculating Bayes factors, the degree of support for H0 and H1 can be updated based on data. Therefore, as new data emerge, the support for different hypotheses can be continuously updated. Within the Bayesian framework, the calculation and interpretation of Bayes factors do not require assuming infinite repetitions of experiments; instead, Bayes factors are updated according to the law of likelihood. Additionally, the order in which data appear does not affect the interpretation of Bayes factors (Rouder, 2014).

Under the Bayesian framework, there is no need to assume infinite repeated trials, and the interpretation of Bayes factors is not affected by when data collection is stopped (Rouder, 2014). In fact, if researchers adopt a sequential Bayes factor design and set reasonable thresholds for Bayes factors in advance (typically 10, representing strong evidence), they can update posterior probabilities as data accumulate during the experiment and stop data collection when appropriate (Schlaifer & Raiffa, 1961; Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017). This principle of not being affected by stopping rules is of great significance for practical research, enabling researchers to collect data reasonably and efficiently.

2.5 Not Affected by Sampling Plans

A sampling plan refers to researchers' pre-study planning of sample selection and data collection processes based on the assumptions of data analysis to ensure that data meet statistical assumptions. For example, random sampling and random assignment are commonly used practices in psychological experiments. Since NHST involves certain underlying assumptions, sampling plans (especially power analysis) are important for interpreting p-values (Halsey, Curran-Everett, Vowler, & Drummond, 2015).

However, the interpretation of Bayes factors is not affected by sampling plans because Bayes factor calculations use the likelihood principle (Berger & Wolpert, 1988), which makes no prior assumptions about data analysis. In other words, even if researchers are unclear about the data collection process, they can still calculate and interpret Bayes factors. This characteristic is very practical for analyzing data obtained in natural settings.

Again using the study of the relationship between presidential candidates' height and election outcomes as an example, researchers found $r = 0.39$, $p = 0.007$ (Stulp et al., 2013). Under the NHST framework, to properly interpret the p-value, we must assume that the experimenter had planned to conduct 46 elections before the presidential election and would stop collecting data after the 46th election, calculating the correlation coefficient based on this plan. Without meeting these assumptions, the meaning of $p = 0.007$ is difficult to interpret. However, it is obvious that these assumptions are not valid.

This example also involves issues related to stopping rules (i.e., under what

conditions to stop collecting data): in real life, U.S. presidential elections will continue, and data will continue to accumulate. How should these future data be analyzed? If NHST analysis is conducted every time a new data point is added, it will cause multiple comparison problems, increasing false positives†.

† For Frequentist analysis, multiple comparisons are non-independent, and correction methods reduce but cannot eliminate Type I errors.

Unlike NHST, Bayes factors can be continuously updated as new data emerge, enabling the analysis of real-world data outside the laboratory and allowing for meaningful interpretation of data. From this perspective, the advantage of Bayes factors in real-time evidence monitoring is related to their advantage of not being affected by sampling plans: both advantages exist because Bayes factors do not depend on researchers' intentions in collecting data. However, as we will mention later, although updating Bayes factors with data does not affect their interpretation, this approach of ignoring false positives cannot prevent the increase in false positives, and researchers still need to control false positives by setting reasonable thresholds in advance and/or selecting appropriate priors.

In summary, Bayes factors condition on observed data to quantitatively analyze the degree of support that current data provide for H0 and H1. By monitoring changes in evidence strength in real time, Bayes factors allow researchers to track evidence accumulation while collecting data. If predetermined stopping thresholds for Bayes factors are set (e.g., stop collecting data when $BF10 > 10$ or $BF10 < 1/10$), researchers can stop data collection when evidence is sufficiently strong. Moreover, even without information about data collection plans, Bayes factors can still obtain evidence from observed data to determine which hypothesis is more strongly supported.

3 Calculating Bayes Factors Using JASP

Due to the unique advantages of Bayes factors, researchers have long attempted to introduce them into psychological research (Edwards et al., 1963). However, the calculation of Bayes factors becomes more complex in practice with different data types and analysis types (for relevant formulas, see Morey & Rouder, 2011; Rouder et al., 2012; Rouder, Morey, Verhagen, Swagman, & Wagenmakers, 2017; Rouder et al., 2009). It is precisely for this reason that Bayes factors have been greatly limited in psychological research. Recently, researchers have utilized R's rich software packages to develop the visual statistical tool JASP (https://jasp-stats.org/), which uses a graphical interface similar to SPSS, making the calculation of Bayes factors more easily achievable. This section introduces the JASP software and its use‡.

‡ Some content in this subsection is adapted from Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., et al. (2017). Bayesian Inference for Psychology. Part II: Example Applications with JASP. Psychonomic Bulletin & Review.

3.1 Introduction to JASP Software

JASP is a free, open-source statistical software that uses R packages for data processing but does not require R installation. JASP's long-term goal is to enable everyone to access state-of-the-art statistical techniques, especially Bayes factors, through free statistical software.

JASP was developed against the backdrop of the reproducibility crisis in psychological research, with the following development principles: First, open-source and free, because openness should be an essential element of scientific research; second, inclusiveness, including both Bayesian analysis and NHST methods, with the latter also providing output of effect sizes and their confidence intervals (Cumming, 2014); third, simplicity, meaning that JASP's basic software includes only the most commonly used analyses, while more advanced statistical methods can be supplemented through plugin modules; fourth, a user-friendly graphical interface, for example, the output section updates in real-time as users select variables for input, and tables use APA format. At the same time, JASP uses progressive output, meaning that the default result output is the most concise, and more detailed output can be defined by researchers themselves. Additionally, to facilitate the public sharing of analysis processes, JASP saves input data and output results in the same file with a .jasp extension, with each analysis result associated with the corresponding analysis and variable data. This integrated file of results and data is compatible with the Open Science Framework (OSF, https://osf.io/), thereby achieving open data and results.

3.2 Implementation of Bayes Factor Analysis in JASP and Interpretation of Results

Currently, JASP can implement Bayes factor analysis for various experimental designs, including one-sample t-tests, independent samples t-tests, paired samples t-tests, ANOVA, repeated measures ANOVA, ANCOVA, and correlation analysis. For each analysis, both Frequentist and Bayesian methods are provided. JASP's Bayes factor analysis uses default prior distributions, but these can be modified. Next, this article uses the replication experiment data from Wagenmakers et al. (2015, https://osf.io/uszvx/) on Topolinski and Sparenberg (2012) as an example to demonstrate how to conduct independent samples t-tests using JASP. For other commonly used Bayes factor analyses, see Wagenmakers et al. (2017).

In the second experiment of Topolinski and Sparenberg (2012), one group of participants turned a kitchen roll clockwise, while another group turned it counterclockwise. Subsequently, participants completed a questionnaire assessing openness to experience. Their data showed that participants who turned clockwise reported higher openness to experience than those who turned counterclockwise (Topolinski & Sparenberg, 2012) (but see Francis, 2013). Wagenmakers et al. (2015) used preregistration to replicate this study, determining the criterion for stopping data collection before the experiment: data collection would stop

when the Bayes factor supporting a hypothesis reached 10, or when 50 participants per condition were reached. Additionally, the preregistration used the default prior for one-sided t-tests, i.e., a Cauchy distribution with $\gamma = 1$. The prior for one-sided t-tests is a Cauchy distribution with only positive effects, meaning the alternative hypothesis is H+: Cauchy (0, 1).

Some researchers argue that the default prior distribution Cauchy (0, 1) is unrealistic because this distribution gives too large a proportion to large effect sizes (effect sizes greater than 1 account for more than 50% of the distribution). Conversely, others find it unrealistic because this distribution gives too much weight to effect sizes near 0, making an effect size of 0 the most likely value. One way to avoid these problems is to reduce the parameter r of the Cauchy distribution. In the BayesFactor package, the default value is √2 0.707. JASP also uses this prior for one-sided t-tests. Reducing r means H1 and H0 become more similar, their predictions for observed data become more similar, making it harder to obtain strong evidence supporting H0.

Using JASP, we can conduct Bayesian independent samples t-tests on this dataset. First, open the data in JASP (File → Examples → "Kitchen Rolls", or download from https://osf.io/9r423/ and click File → Open), then select "Bayesian Independent Samples T-test" in the T-tests panel. This will display a dialog box as shown in the middle panel of Figure 1. We have set "mean NEO" as the dependent variable and "Rotation" as the grouping variable. As shown in the middle of Figure 2 [Figure 2: see original paper], set the width of the Cauchy prior to JASP's default value $\gamma = 0.707$, and check the options "Prior and posterior" and its sub-option "Additional info", which yields the result shown on the right side of Figure 2: compared to clockwise rotation, counterclockwise rotation shows slightly higher openness to experience, a result in the opposite direction of what Topolinski and Sparenberg (2012) hypothesized. In the lower right part of Figure 2, the solid line is the posterior distribution, and the dashed line is the prior distribution. It can be seen that most of the posterior probability is negative, with a median of -0.13 and a 95% credible interval from -0.5 to 0.23. BF01 = 3.71, indicating that the observed data is 3.71 times more likely under the H0 hypothesis than under the H1 hypothesis (we chose BF01 because BF01 = 3.71 is easier to interpret than the equivalent BF10 = 0.27).

Figure 2. Screenshot of conducting Bayesian independent samples t-test in JASP. The left side shows the data; the middle shows analysis options; the right shows results.

Through this preliminary demonstration, we can understand how to conduct Bayesian independent samples t-tests. Next, we demonstrate how to conduct Bayesian one-sided independent samples t-tests on this dataset according to the preregistered method. Since descriptive statistics output shows that clockwise is group 1 and counterclockwise is group 2, we check "group 1 > group 2" in the "Hypothesis" panel, as shown in the middle of Figure 3 [Figure 3: see original paper].

Figure 3. Illustration of conducting Bayesian one-sided independent samples t-test on Wagenmakers et al. (2015) data in JASP. Left side shows data, middle shows operation process, right shows result output. Details are described in the text.

The results of the one-sided test are shown on the right side of Figure 3. As expected, if the observed effect is opposite to the hypothesis, this approach of integrating prior knowledge into the analysis increases the relative evidence supporting H0 (also see Matzke et al., 2015), with the Bayes factor BF01 increasing from 3.71 to 7.74, meaning the observed data is 7.74 times more likely under H0 than under H+.

It is worth noting that the posterior distribution under H+ is concentrated at 0 but not without negative values (see right side of Figure 3), which is consistent with the order restriction in H+. This differs from traditional Frequentist one-sided confidence intervals, where the traditional one-tailed confidence interval is [-0.23, $+\infty$) §. Although the traditional Frequentist interval is mathematically well-defined (i.e., it includes all values that would not be rejected by a one-tailed $\alpha = 0.05$ significance test), most researchers find this interval neither understandable nor informative (Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016).

§ This interval can be obtained using the t.test function in R language.

Figure 4 [Figure 4: see original paper] Bayesian robustness check in JASP.

In addition to calculating Bayes factors, JASP can also conduct robustness checks to quantify the impact of the Cauchy prior distribution parameter r on Bayes factors. As shown in Figure 4, check the option "Bayes factor robustness check", which will produce the upper right graph in Figure 4. From this graph, we can see that when the Cauchy prior r is 0, H0 and H+ are identical (BF0+ = 1), and BF0+ increases as r increases. At JASP's default value r = 0.707, the Bayes factor BF0+ = 7.73; for Jeffreys' default r = 1, the Bayes factor BF0+ = 10.75. Therefore, across a range of prior values for r, the current data show moderate to strong evidence supporting H0.

Furthermore, we can check "Sequential analysis" and its sub-option "Robustness check" in the middle part of Figure 4 to conduct sequential analysis. The results are shown in the lower right graph of Figure 4. Sequential analysis displays how Bayes factors change with sampling, meaning researchers can monitor and visualize evidence accumulation as new data are collected.

From the graph, we can see that Wagenmakers et al. (2015) did not actually calculate BF0+ using the preregistered prior of $\gamma = 1$ and stop data collection immediately when BF0+ > 10 or BF+0 > 10: after 55 participants, the dashed line exceeded BF0+ > 10, but data collection continued. In practice, checking Bayes factors every few days helps researchers understand whether Bayes factors exceed predetermined criteria at some point and decide whether to stop data

collection accordingly.

One advantage of sequential analysis is that it visualizes the convergence process of Bayes factors under different prior conditions, i.e., when the differences in Bayes factors on a log scale begin to stabilize (e.g., Bahadur & Bickel, 2009; Gronau & Wagenmakers, 2017). In the current example, when the number of participants reached 35, Bayes factors under different priors began to converge. To understand why the differences in log Bayes factors stop changing after some initial observations, we can assume data y includes two parts y1 and y2. According to the conditional probability formula, BF0+(y) = BF0+(y1) × BF0+(y2|y1). This formula shows that Bayes factors are not simply multiplied blindly across different data; the second factor—BF0+(y2|y1)—actually reflects how data y2 updates the Bayes factor after the prior distribution has been updated based on data y1 (Jeffreys, 1961, p. 333). Converting this formula to log form yields log(BF0+(y)) = log(BF0+(y1)) + log(BF0+(y2|y1)). Assuming data y1 contains sufficient information, regardless of how r varies, roughly the same posterior distribution is obtained through y1 (in most cases, this happens quickly). This posterior distribution obtained through y1 then becomes the prior distribution for data y2, i.e., the prior for log(BF0+(y2|y1)). In this case, the value of log(BF0+(y2|y1)) is roughly similar (similar prior distribution, same data). Therefore, different r values produce different posterior distributions from data y1, but when data y1 is sufficiently large to produce roughly similar posterior distributions, the amount of updating by y2 to the model is also similar, making log(BF0+(y2|y1)) similar across different r values, resulting in convergence.

3.3 How to Report Bayes Factor Results

Bayesian statistics are not yet common in current psychological research. Although most journal editors and reviewers appreciate the use of more reasonable statistical methods, researchers using Bayes factors need to provide relevant background information due to unfamiliarity with Bayesian methods. Therefore, in addition to reporting Bayes factor results, the following points should also be reported (Kruschke, 2014). First, the motivation and reasons for choosing Bayes factors—that is, why Bayes factors are used instead of NHST in a particular report. As mentioned earlier, one can explain that Bayes factors provide richer information or that the data characteristics do not meet the assumptions of NHST (e.g., data collected in natural settings where data collection motivations and experimental hypotheses cannot be determined). Second, describe the basic logic of model comparison in Bayes factors—that is, briefly explain the idea of model comparison in Bayes factors, assuming readers are not very familiar with the method. Third, describe the prior distributions used in Bayes factor analysis and the reasons for choosing them; prior distributions should provide some information for data analysis. Fourth, interpret the Bayes factor by connecting it to the theories or hypotheses in the research.

Bayes factors do not use statistical significance but instead describe the degree

of support that data provide for hypotheses. For example, in Wagenmakers et al. (2015), the Bayes factor result under Jeffreys' default prior was described as follows:

"The Bayes factor is BF01 = 10.76, indicating that the observed data is 10.76 times more likely under the null hypothesis (which assumes no effect) than under the alternative hypothesis (which assumes an effect). According to the classification criteria proposed by Jeffreys (1961), this is strong evidence supporting the null hypothesis, i.e., there is no difference in NEO scores between people who turn clock hands clockwise and counterclockwise."

Additionally, when using Bayes factors for analysis, exploratory results such as robustness distributions and sequential analysis results can be reported, which will further enrich the results and provide more comprehensive information to other researchers.

4 Summary and Outlook

In recent years, the issue of reproducibility in scientific research has attracted much attention (Baker, 2016; Begley & Ellis, 2012; Munafò et al., 2017), especially in psychology (Ebersole et al., 2016; Klein et al., 2014; Open Science Collaboration, 2015) and neuroimaging (Poldrack et al., 2017; Zuo & Xing, 2014). Over-reliance on NHST is one of the causes (Lindsay, 2015; Hu et al., 2016). Therefore, researchers hope that Bayes factors, as a hypothesis testing method, can change the current situation of over-reliance on NHST in psychological research. Of course, other solutions have also been proposed, such as lowering the significance threshold to 0.005 (Benjamin et al., 2017) or using likelihood ratios for model comparison (Etz, in press). However, it is worth noting that there are various reasons for replication failures in psychological research, and changing statistical methods alone cannot make psychological research reproducible. Issues such as non-open data and non-transparent research processes (Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014; Lindsay, 2015; Nosek et al., 2015), failure to distinguish between exploratory and confirmatory analyses (Kerr, 1998; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012), and a publication-based reward system (Nosek, Spies, & Motyl, 2012) may all contribute to low research reproducibility. Therefore, to some extent, maintaining openness and transparency in data analysis processes and results is a key solution (e.g., Poldrack & Gorgolewski, 2017; Zuo et al., 2014).

Nevertheless, as a method different from traditional NHST, Bayes factors help researchers use multiple methods to analyze the same study, thereby obtaining accurate statistical inferences and conclusions closer to the truth. It should be noted that when using multiple methods for analysis, all analysis processes and results need to be reported, rather than selectively reporting only the results most favorable to one's conclusions.

4.1 Limitations of Bayes Factors

Bayes factors are an application of Bayesian statistics to hypothesis testing, and the debate between Bayesian and Frequentist statistics has long existed (Miller, 2011). In fact, researchers have pointed out that Bayes factors may also have many problems, and fully understanding these opposing views will be more conducive to the reasonable use of Bayes factors in research.

The strongest criticism of Bayes factors comes from the setting of their prior probabilities, with concerns that prior probabilities are too subjective, too conservative, and thus unlikely to produce strong evidence (Wagenmakers, Marsman, et al., 2017). Some researchers also believe that default priors are unfavorable to small effects. For example, Bem, Utts, and Johnson (2011) argued that when Wagenmakers et al. (2011) reanalyzed Bem's (2011) data, their failure to obtain consistent conclusions was due to the use of inappropriate prior probabilities. This criticism essentially represents a misuse of Bayes factors, i.e., failing to transform prior knowledge into appropriate prior probabilities (Hoijtink, van Kooten, & Hulsker, 2016). Interestingly, as long as researchers maintain transparency and openness about the prior probabilities they use, other researchers can conduct cross-validation, thereby achieving full exploration.

Second, some researchers believe that Bayes factors do not consider the problem of false positives. Under the NHST framework, researchers emphasize controlling Type I and Type II errors. For example, psychological research generally controls Type I errors within 5%, thus setting the significance level at 0.05. It is precisely because of the need to control Type I errors that there are many methods in the NHST framework to adjust thresholds so that Type I error rates are not too high, such as multiple comparison correction methods. Bayesian statistics mainly aims to continuously measure the strength of evidence and does not consider controlling false positives (i.e., Type I errors). Therefore, when researchers make decisions based on Bayes factors (whether an effect exists), they may commit Type I errors (Kruschke & Liddell, 2017a). In actual Bayes factor analysis, the problem of multiple comparisons can be addressed through priors (Jeffreys, 1938; Scott & Berger, 2006, 2010). For example, directly stating how large the researcher expects the false positive rate to be (Stephens & Balding, 2009).

Some researchers also point out that estimation-based statistics are always superior to hypothesis testing because estimation itself incorporates uncertainty. For example, Cumming (2014) recommends using effect sizes and their confidence intervals to replace p-values. However, considering that both parameter estimation and hypothesis testing have their most applicable problems in research, Bayes factors cannot be directly compared with estimation-based Frequentist statistics. Nevertheless, there are also estimation-based methods in Bayesian statistics (Kruschke & Liddell, 2017b).

Finally, hypothesis testing using Bayes factors is essentially about the continuous accumulation of evidence rather than obtaining a dichotomous conclusion.

Therefore, the results of a single experiment can be considered tentative, and researchers can continue to collect data or conduct replication studies (Ly, Etz, Marsman, & Wagenmakers, 2017).

4.2 Application Prospects of Bayes Factors

As a hypothesis testing method based on Bayesian statistics, Bayes factors have some advantages over NHST, enabling researchers to directly test whether data support the null hypothesis and no longer be affected by sampling intentions and stopping criteria, thus allowing more flexible data analysis. These advantages may help psychologists make better decisions during the research process, and the adoption of Bayes factors can also promote researchers to more deeply understand the applicable scope and prerequisites of Bayesian methods (Depaoli & van de Schoot, 2017).

The development of JASP makes the calculation and interpretation of Bayes factors more convenient, allowing researchers to conduct Bayes factor analysis using JASP even without strong programming skills. This may help promote the wider use of Bayes factors among researchers. Moreover, JASP itself is rapidly developing, with its functional depth and breadth continuously expanding, and new methods and standards will be continuously integrated into the software, potentially helping researchers conduct more scientific research.

Acknowledgments: We thank Zhang Mi from the Department of Psychology, School of Social Sciences, Tsinghua University for her help in the early stages of writing this article, and we thank two anonymous reviewers for their valuable comments on this article.

References:

Hu, C.-P., Wang, F., Guo, J.-C.-S., Song, M.-D., Sui, J., & Peng, K.-P. (2016). The reproducibility crisis in psychological research: From crisis to opportunity. *Advances in Psychological Science*, 24(9), 1504–1518.

Luo, D.-S. (2017). Evaluating two sources of the reproducibility crisis in psychology. *Studies of Psychology and Behavior*, 15(5), 577–586.

Zhong, J.-J., Dienes, Z., & Chen, Z.-Y. (2017). The necessity, application ideas, and fields of introducing Bayesian statistical inference in psychological research. *Journal of Psychological Science*, 40(6), 1477–1482.

Bahadur, R. R., & Bickel, P. J. (2009). An optimality property of Bayes' test statistics. *Lecture Notes-Monograph Series*, 57, 18–30.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 553, 452–454.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391), 531–533.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100(3), 407–425.

Bem, D. J., Utts, J., & Johnson, W. O. (2011). Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology*, 101(4), 716–719.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., et al. (2017). Redefine statistical significance. *Nature Human Behaviour*, 2(1), 6–10.

Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, 76(2), 159–165.

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, 2(3), 317–335.

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle* (2nd ed.). Hayward (CA): Institute of Mathematical Statistics.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. J. (2014). Instead of "playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. *AIMS Neuroscience*, 1(2373-7972), 4–17.

Chen, X., Lu, B., & Yan, C.-G. (2018). Reproducibility of R-fMRI metrics on the impact of different strategies for multiple comparison correction and sample sizes. *Human Brain Mapping*, 39(1), 62–73.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25(1), 7–29.

Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods*, 22(2), 240–261.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. London, UK: Palgrave Macmillan.

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

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

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., et al. (2016). Many Labs 3: Evaluating participant pool

quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67, 68–82.

Edwards, W. (1965). Tactical note on the relation between scientific and statistical hypotheses. *Psychological Bulletin*, 63(6), 400–402.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70(3), 193–242.

Etz, A. (in press). Introduction to the concept of likelihood and its applications. *Advances in Methods and Practices in Psychological Science*.

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology*, 57(5), 153–169.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, 116(2), 439–453.

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

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., et al. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, 31(4), 337–350.

Gronau, Q. F., & Wagenmakers, E.-J. (2017). Bayesian evidence accumulation in experimental mathematics: A case study of four irrational numbers. *Experimental Mathematics*, 1–10.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature Methods*, 12(3), 179–185.

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists*. Boca Raton, FL: Chapman & Hall/CRC.

Hoijtink, H., van Kooten, P., & Hulsker, K. (2016). Why Bayesian psychologists should change the way they use the Bayes factor. *Multivariate Behavioral Research*, 51(1), 2–10.

JASP Team. (2017). JASP (Version 0.8.2) [Computer software].

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(2), 203–222.

Jeffreys, H. (1938). Significance tests when several degrees of freedom arise simultaneously. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 165(921), 161–198.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, 110(48), 19313–19317.

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

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., et al. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, 45(3), 142–152.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, 10(4), 477–493.

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A Tutorial with R, JAGS, and Stan* (2nd ed.). San Diego, CA: Academic Press/Elsevier.

Kruschke, J. K., & Liddell, T. M. (2017a). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, 1–23.

Kruschke, J. K., & Liddell, T. M. (2017b). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*.

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

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, 15(1), 22–25.

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, 26(12), 1827–1832.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 28(25), 3049–3067.

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2017). Replication Bayes factors from evidence updating. *PsyArXiv*. Retrieved from https://osf.io/preprints/psyarxiv/u8m2s/

Ly, A., Marsman, M., & Wagenmakers, E.-J. (2017). Analytic posteriors for Pearson's correlation coefficient. *Statistica Neerlandica*.

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016a). An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology*, 72, 43–56.

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016b). 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. (2016a). Bayesian benefits with JASP. *European Journal of Developmental Psychology*, 14(5), 545–555.

Marsman, M., & Wagenmakers, E.-J. (2016b). Three Insights from a Bayesian Interpretation of the One-Sided P Value. *Educational and Psychological Measurement*, 77(3), 529–539.

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, 43(3), 679–690.

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(1), e1–e15.

Miller, G. (2011). ESP paper rekindles discussion about statistics. *Science*, 331(6015), 272–273.

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(1), 103–123.

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

Mulder, J., Klugkist, I., van de Schoot, R., Meeus, W. H. J., Selfhout, M., & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology*, 53(6), 530–546.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., et al. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(1), 0021.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., et al. (2015). Promoting an open research culture. *Science*, 348(6242), 1422–1425.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6), 615–631.

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

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Paper presented at the Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003).

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafo, M. R., et al. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, 18(2), 115–126.

Poldrack, R. A., & Gorgolewski, K. J. (2017). OpenfMRI: Open sharing of task fMRI data. *NeuroImage*, 144, Part B, 259–261.

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

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, 18(4), 682–689.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, 56(5), 356–374.

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E.-J. (2017). Bayesian analysis of factorial designs. *Psychological Methods*, 22(2), 304–321.

Rouder, J. N., Speckman, P., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2), 225–237.

Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century.* New York, NY: W. H. Freeman and Company.

Schervish, M. J. (1996). P values: What they are and what they are not. *The American Statistician*, 50(3), 203–206.

Schlaifer, R., & Raiffa, H. (1961). *Applied statistical decision theory.*

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.

Scott, J. G., & Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136(7), 2144–2162.

Scott, J. G., & Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5), 2587–2619.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of Values for testing precise null hypotheses. *The American Statistician*, 55(1), 62–71.

Stephens, M., & Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10), 681–690.

Stulp, G., Buunk, A. P., Verhulst, S., & Pollet, T. V. (2013). Tall claims? Sense and nonsense about the importance of height of US presidents. *The Leadership Quarterly*, 24(1), 159–171.

Topolinski, S., & Sparenberg, P. (2012). Turning the hands of time. *Social Psychological and Personality Science*, 3(3), 308–314.

van de Schoot, R., Winter, S., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian papers in psychology: The last 25 years. *Psychological Methods*, 22(2), 217–239.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, 54(6), 491–498.

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

Wagenmakers, E.-J., Beek, T. F., Rotteveel, M., Gierholz, A., Matzke, D., Steingroever, H., et al. (2015). Turning the hands of time again: a purely confirmatory replication study and a Bayesian analysis. *Frontiers in Psychology*, 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(3), 158–189.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., et al. (2017). Bayesian Inference for Psychology. Part II: Example Applications with JASP. *Psychonomic Bulletin & Review*.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., et al. (2017). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*.

Wagenmakers, E.-J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., et al. (2017). The need for Bayesian hypothesis testing in psychological science. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological Science Under Scrutiny* (pp. 123–138): John Wiley & Sons.

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: comment on Bem (2011). *Journal of Personality and Social Psychology*, 100(3), 426–432.

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(6), 632–638.

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

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, 6(3), 291–298.

Zhu, J., Chen, J., Hu, W., & Zhang, B. (2017). Big Learning with Bayesian methods. *National Science Review*, 4(4), 627–651.

Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance*. Ann Arbor: University of Michigan Press.

Zuo, X.-N., Anderson, J. S., Bellec, P., Birn, R. M., Biswal, B. B., Blautzik, J., et al. (2014). An open science resource for establishing reliability and reproducibility in functional connectomics. *Nature Scientific Data*, 1, 140049.

Zuo, X.-N., & Xing, X.-X. (2014). Test-retest reliabilities of resting-state FMRI measurements in human brain functional connectomics: A systems neuroscience perspective. *Neuroscience & Biobehavioral Reviews*, 45, 100–118.

Note: Figure translations are in progress. See original paper for figures.