A Bayes Factor for Replications of ANOVA Results

Christopher Harms[a],[*]

September 9, 2018

ABSTRACT

With an increasing number of replication studies performed in psychological science, the question of how to evaluate the outcome of a replication attempt deserves careful consideration. Bayesian approaches allow to incorporate uncertainty and prior information into the analysis of the replication attempt by their design. The Replication Bayes factor, introduced by Verhagen and Wagenmakers (2014), provides quantitative, relative evidence in favor or against a successful replication. In previous work by Verhagen and Wagenmakers (2014) it was limited to the case of t-tests. In this paper, the Replication Bayes factor is extended to F-tests in multi-group, fixed-effect ANOVA designs. Simulations and examples are presented to facilitate the understanding and to demonstrate the usefulness of this approach. Finally, the Replication Bayes factor is compared to other Bayesian and frequentist approaches and discussed in the context of replication attempts. R code to calculate Replication Bayes factors and to reproduce the examples in the paper is available at https://osf.io/jv39h/.

[*] *Corresponding author. Email address:* christopher.harms@uni-bonn.de
[a] *Department of Psychology, University of Bonn, Bonn, Germany*

1 INTRODUCTION

The "replication crisis" (Pashler and Wagenmakers, 2012; Asendorpf et al., 2013) has been a focus of discussion in psychological science in recent years. There is ongoing debate about how to improve methodological rigor and statistical analysis in order to improve reliability and replicability of published findings. When considering a given replication study, a central question is the final evaluation, i.e. the question whether a previous finding has successfully been replicated.

As with any scientific question in empirical disciplines, researchers use statistical tools in order to find an answer to this question. An intuitive way to evaluate replication results is to compare statistical significance in both an original study and a replication study ("vote counting"): If both studies are significant and show effects in the same direction, the replication is deemed "successful". If, on the other hand, the original study reported a statistically significant test which is non-significant in the replication study, the replication might be considered "failed". This interpretation, while intuitive and common practice in past research (Maxwell et al., 2015), is flawed in general and wrong in the case of non-significant replications. A non-significant result cannot be interpreted as evidence for the absence of an effect – especially since the difference between a significant result in an original study and a non-significant result in a replication study might not be significant in itself (Gelman and Stern, 2006). Moreover, statistical significance and p-values do not contain information about uncertainty in estimates and their misinterpretations have repeatedly been covered over the past decades (Wasserstein and Lazar, 2016; Ionides et al., 2017; Nickerson, 2000; Bakan, 1966). A related question is whether the strict dichotomy between "successful" and "failed" replications is sensible in practice.

To overcome problems with simply comparing p-values based on their significance, other methods for evaluating replication studies have been proposed, partly also alleviating the strict dichotomy: confidence intervals (CI) for effect size estimates have been advocated by some authors for both the reporting of statistical summaries (Cumming and Finch, 2001; Cumming, 2012) and for the evaluation of replications (Cumming, 2008; Gilbert et al., 2016, but see also Anderson et al., 2016). A researcher could, for example, check if the effect size estimate from a replication study falls within the 95% confidence interval of the effect size estimate from the original study – or, *vice versa*, check if the effect size estimate from the original study falls within the 95% confidence interval of the replication study. While this approach takes uncertainty into account (in contrast to p-values), confidence intervals can be difficult to interpret and give rise to misinterpretations (Belia et al., 2005; Anderson et al., 2016).

The width of a confidence interval is directly related to sample size and thus to the power curve of a statistical test. Since power is notoriously low

in psychological science (Button et al., 2013; Sedlmeier and Gigerenzer, 1989; Szucs and Ioannidis, 2017), confidence intervals tend to be generally wide for original studies from the published literature (Morey and Lakens, 2016). This can make a sensible comparison of confidence intervals difficult and lead to inconclusive or misleading results (depending on the particular decision rule). Simonsohn (2015) therefore proposed a method to take the power of the original study into account when evaluating a replication study. In his "small telescopes" approach, the confidence interval of a replication study is compared not to the effect size estimate from the original study, but to an effect the original study had 33% power to detect. The benchmark of 33% is arbitrary, but Simonsohn (2015) considers it a "small effect". The goal is to determine if a replication study, yielding a smaller effect size than in the original study, is still in line with a small effect that the original study had only little power to detect. This allows for a more nuanced interpretation of a replication outcome than simply comparing p-values or confidence intervals.

The approaches outlined so far rely on a frequentist interpretation of probability and data analysis. Frequentist statistics is primarily concerned with repeated sampling and long-run rates of events. It does not allow a researcher to answer questions such as "How much more should I believe in an effect?" or "Given the data, what are the most credible effect size estimates?". Bayesian statistics allows to address such questions in a coherent framework and uses uncertainty in estimates as an integral part of the analysis.

Related to the analysis of a replication study, various Bayesian approaches are possible. Marsman et al. (2017), for example, have evaluated the results from the "Reproducibility Project: Psychology" (Open Science Collaboration, 2015) using several methods, including Bayesian parameter estimation on the individual study level, Bayesian meta-analysis, and Bayes factors. For psychology in particular, Bayes factors (Kass and Raftery, 1995; Jeffreys, 1961) have repeatedly been proposed as an alternative or addition to significance testing in common study scenarios (see Morey et al., 2016, or Bayarri et al., 2015, for introductions to Bayesian hypothesis testing).

Verhagen and Wagenmakers (2014) compared several Bayes factors to be used in the context of replication studies and introduced a Bayes factor to specifically investigate the outcome of a replication study, which they termed *Replication Bayes factor*. It is constructed to be used only with reported test-statistics from an original study and a replication study. As some researchers disagree with the notion of "belief" in data-analysis, it is also a favorable property of the Replication Bayes factor that it makes only little assumptions about prior beliefs. The Replication Bayes factor allows to test the hypothesis that the effect in a replication study is in line with an original study against the hypothesis that there is no effect. The result is a continuous quantification of relative evidence (Morey et al., 2016), showing

by much the data are more in line with one hypothesis compared to the other.

The present article focuses on the Bayesian perspective and extends the Replication Bayes factor. Verhagen and Wagenmakers (2014) introduced the Replication Bayes factor solely for the case of one- and two-sample t-tests. In this paper, it will be extended to the case of F-tests in fixed-effect ANOVA designs, a common study design in cognitive, social, and other sub-fields of psychology. In order to do so, an outline of hypothesis testing using Bayes factors and how it can be applied to replication studies is presented. In the second section, it is shown how the Replication Bayes factor can be used in studies investigating difference between several groups, when fixed-effect ANOVAs are carried out. Simulations and example studies are then followed by a general discussion of the method.

In general, it has been recommended (Brandt et al., 2014; Anderson and Maxwell, 2016) that replicators should include different methods to evaluate the outcome of a replication study to account for the advantages, limitations, and varying statistical questions the different approaches present. This allows for a more nuanced interpretation of replication studies than a simple comparison of statistical significance. The Replication Bayes factor is hence presented as an addition to existing methods of analyzing the outcome of a replication study. The article aims to provide an accessible introduction in the setup of the Replication Bayes factors, but assumes a general understanding of Bayesian statistics – in particular with the general concepts of priors, marginal likelihoods, and Bayesian belief updating. Readers unfamiliar with the Bayesian approach might find the excellent textbooks by McElreath (2016) and Kruschke (2015) helpful. A reading list of relevant introductory papers is given by Etz et al. (2018). Rouder et al. (2012) and Morey et al. (2016) have elaborated specifically on the use of Bayes factors.

To reproduce the examples in this article, all scripts are available in an OSF repository at `https://osf.io/jv39h/` (Harms, 2018). To calculate the Replication Bayes factor for t-tests (Verhagen and Wagenmakers, 2014) and F-tests (the present article), an R-Package is available at `https://github.com/neurotroph/ReplicationBF`.

2   BAYESIAN HYPOTHESIS TESTING

While Bayes' theorem allows several interpretations, Bayes factors are best understood in the context of *Bayesian belief updating*. Bayes factors indicate how to rationally shift beliefs in two competing hypotheses (formalized as models M_0 and M_1) based on the observed data Y:

$$\underbrace{\frac{P(M_0|Y)}{P(M_1|Y)}}_{\text{Posterior Odds}} = \underbrace{\frac{\pi(M_0)}{\pi(M_1)}}_{\text{Prior Odds}} \times \underbrace{\frac{P(Y|M_0)}{P(Y|M_1)}}_{\text{Bayes factor}} \tag{1}$$

The Bayes factor $BF_{01} = \frac{P(Y|M_0)}{P(Y|M_1)}$ is the factor by which the prior odds are multiplied in order to get updated posterior odds (Jeffreys, 1961; Lindley, 1993; Kass and Raftery, 1995). It is a ratio of the *marginal likelihoods* or *model evidence* of the two models, which is given by

$$P(Y|M_i) = \int p(Y|\theta, M_i)\pi(\theta|M_i)\, d\theta \qquad (2)$$

where θ is the vector of model parameters, $\pi(\theta|M_i)$ is the prior distribution for the parameters in Model M_i and $p(Y|\theta, M_i)$ is the likelihood function of M_i. The marginal likelihood is the normalizing constant in Bayes' rule in order for the posterior to be a proper probability distribution. Hence, the integrand consists of the same parts as the nominator in Bayes' rule, namely likelihood and prior.

The challenge of computing Bayes factors lies in the possible complexity of the integral. As will be explained in more detail later, in many practical cases – such as in the case for the Replication Bayes factor – the marginal likelihood needs to be approximated, for example by using Monte Carlo methods.

Bayes factors provide relative evidence for one model when compared to another model and they can – in contrast to p-values in the Neyman-Pearson framework of null-hypothesis significance testing – be interpreted directly in a continuous, quantitative manner. To facilitate verbal interpretation, Jeffreys (1961) and Kass and Raftery (1995) provided guidelines for the description of Bayes factors: A BF_{10} is "not worth more than a bare mention" if it is between 1 (the data provide evidence for both models equally) and about 3; the evidence against model M_0 is "substantial" if $3 < BF_{10} \leqslant 10$, "strong" if $10 < BF_{10} \leqslant 100$ and "decisive" if $BF_{10} > 100$.

In recent years it has become increasingly popular to report Bayes factors as an addition or alternative to traditional null-hypothesis significance testing using p-values. Calculations and tools for common scenarios in psychological research exist (e.g. Rouder et al., 2012; Rouder and Morey, 2012; Rouder et al., 2009; Dienes, 2016). In the context of replications, Etz and Vandekerckhove (2016) have used Bayes factors to show that the original studies in the "Reproducibility Project: Psychology" (Open Science Collaboration, 2015) have provided only little relative evidence against the null hypothesis (when taking into account publication bias). Replication studies yielding substantial in favor or against the null hypothesis were mostly studies with larger sample sizes. The analysis is an example how Bayes factors could be used to provide more insights into the results of empirical studies.

To summarize, the Bayesian frameworks allows a researcher to incorporate existing, previous knowledge into the statistical analysis. This is particularly useful in the context of replication studies: An original experiment has provided information about an effect and if one wants to evaluate a replication study, the information from the original should be considered.

The Replication Bayes factor formalizes this and tests the hypothesis of a successful replication against the hypothesis of a null effect.

2.1 Bayes factors for Replications

The *Replication Bayes factor* introduced by Verhagen and Wagenmakers (2014) is a way to use the Bayesian hypothesis testing framework in the context of replications and to quantify the outcome of a replication attempt given the information from the original study. The general idea is to use the posterior from the original study as a prior for the analysis of the replication attempt – in line with the idea of updating beliefs as outlined above. Furthermore, it is desirable to only use data that is easily available from published studies. Thus, ideally, only reported test-statistics with degrees of freedom and sample size are used.

For a Bayesian hypothesis test, two models need to be set up and compared in the Bayes factor. For the Replication Bayes factor these two models or hypotheses about an effect size measure δ are:

1. $H_0 : \delta = 0$, that is the hypothesis that the true effect size is zero (the "skeptic's position" in the terms of Verhagen and Wagenmakers, 2014).

2. $H_r : \delta \approx \delta_{\text{orig}}$, i.e. the hypothesis that the original study is a good estimate of the true effect size (the "proponent's position").

The Replication Bayes factor is then the ratio of the marginal likelihoods of the two models considering the data from the replication study, denoted by Y_{rep} (Verhagen and Wagenmakers, 2014, p. 1461):

$$B_{r0} = \frac{p(Y_{\text{rep}}|H_r)}{p(Y_{\text{rep}}|H_0)} \tag{3}$$

That is, the Replication Bayes factor represents the relative evidence of the data in favor of a successful replication when compared to a null hypothesis of no effect. "Successful" here means, that the replication yields a similar effect size as the original study. This implies an assumption about the validity of the effect size estimate of the original study, which is also discussed later in this paper.

In order to calculate B_{r0} one needs to mathematically define the two models and find a useful representation of the data from the original study (Y_{orig}) and the replication study (Y_{rep}). While the data would ideally be the raw data and used to fit a probabilistic model describing the relationship between independent and dependent variables, for the Replication Bayes factor it is desired to rely on more accessible data. In the case of (published) replication studies often only summary and test statistics are reported. Thus, for the evaluation of replication studies it is desirable to be able to calculate the Replication Bayes factor based on the reported test statistics (with de-

grees of freedom and sample sizes) from the original study and the replication study alone.

2.2 Replication Bayes factor for t-tests

In their paper, Verhagen and Wagenmakers (2014) provided models and formulas for the computation of the Replication Bayes factor for t-tests. In the following section, the rationale and the setup of the model will be reiterated as a basis for the extension to F-tests.

The general model is based on the t-distribution used for the t-test and derives from the distribution of t-values under the alternative hypothesis. The marginal likelihood for the two models that are compared is given

$$p(Y_{rep}|H_i) = \int t_{df_{rep}, \delta\sqrt{N_{rep}}}(t_{rep})\pi(\delta|H_i) \, d\delta \tag{4}$$

with $t_{df,\Delta}(x)$ being the non-central t-distribution with df degrees of freedom and non-centrality parameter $\Delta = \delta\sqrt{N_{rep}}$.

The two models under consideration differ in their prior distributions $\pi(\delta|H_i)$: For the skeptic (H_0), the prior distribution is 1 at $\delta = 0$ and 0 everywhere else. Thus, the marginal likelihood simplifies to

$$p(Y_{rep}|H_0) = t_{df_{rep}}(t_{rep}) \tag{5}$$

Which is the central t-distribution (i.e. non-central t-distribution with non-centrality parameter $\Delta = 0$) evaluated at the point of the t-value observed in the replication study, t_{rep}.

For the proponent on the other hand, the prior distribution for the replication, $\pi(\delta|H_r)$, is the posterior distribution of the original study. If one starts out with a flat, uninformative prior before the original experiment, the resulting posterior distribution was described as Λ'-distribution by Lecoutre (1999). While in Bayesian statistics flat priors are often disregarded in favor of at least minimally regularizing priors (Gelman et al., 2013), in the present case the prior for the original study plays a minor rule: It is quickly overruled by the data even when considering only a single original study. For the Replication Bayes factor, only the posterior of the original study is relevant. The Λ'-distribution can be approximated closely through a normal distribution as Verhagen and Wagenmakers (2014) showed in their appendix. If the prior for the replication, $\pi(\delta|H_r)$, is expressed through the posterior of the original study, so that $\pi(\delta|H_r) = p(\delta|\delta_{orig}, H_r)$, the marginal likelihood for the proponent's model is finally given as

$$p(Y_{rep}|H_r) = \int t_{df_{rep}, \delta\sqrt{N_{rep}}}(t_{rep})p(\delta|\delta_{orig}, H_r) \, d\delta \tag{6}$$

For the integral no closed form is yet known. Hence, it needs to be approximated and different methods exist. Verhagen and Wagenmakers (2014)

used the Monte Carlo estimate (Gamerman and Lopes, 2006, chap. 7.2.1): random samples from the posterior-turned-prior distribution, $p(\delta|\delta_{orig}, H_r)$, are repeatedly drawn and the average of the marginal likelihood term is calculated. Other approaches to approximating the marginal likelihood for the Bayes factor are available and will be discussed below.

Putting Equations 5 and 6 into Equation 3 yields the formula for the Replication Bayes factor for t-tests:

$$B_{r0} = \frac{\int t_{df_{rep}, \delta\sqrt{N_{rep}}}(t_{rep}) p(\delta|\delta_{orig}, H_r) \, d\delta}{t_{df_{rep}}(t_{rep})} \tag{7}$$

In their paper, Verhagen and Wagenmakers (2014) used simulation studies and examples to demonstrate the usefulness of the Replication Bayes factor, and show how it compares to other Bayes factors that might also be used for replication studies.

3 REPLICATION BAYES FACTOR FOR F–TESTS

Many studies in psychological research do not compare two independent groups through t-tests, but investigate differences across multiple groups and interactions between factors. Thus, there seems to be a need to extend this approach to other tests such as the F-test in ANOVAs. In this section the steps necessary to apply the Replication Bayes factor to other tests are explained, and the Bayes factor is derived for the F-test in fixed-effect ANOVA designs.

As a general difference to the t-test, F-tests do not convey information about the direction or location of an effect; the hypothesis under investigation is an omnibus hypothesis if the effect degrees of freedom are $df_{effect} > 1$ (Rosenthal et al., 2000; Steiger, 2004). Therefore incorporating the F-statistic in the Replication Bayes factor does not allow researchers to take the direction of the effect into account. As will be shown in Example 3 and discussed later, researchers need to consider additional information when evaluating the outcome of replication studies with ANOVA designs. Nevertheless, researchers use and report F-tests both for omnibus and interaction hypotheses. And the F-statistic does contain information about the effect size. Thus the F-statistic can be used in the Replication Bayes factor if a statement about the size of an effect is desired.

In order to maintain the general nature of the Replication Bayes factor from Equation 3, an effect size measure needs to be chosen to parametrise the model. Cohen's f^2 (Cohen, 1988) has a simple relationship to the non-centrality parameter λ of the non-central F-distribution (Steiger, 2004):

$$\lambda = f^2 \cdot N$$

Since this relationship holds only in the case of fixed-effects models, the Replication Bayes factor for F-tests can also only be used validly in these cases – a limitation that will also be discussed later.

Setting up the model based on the F-distribution in a way similar to the t-test case (cf. Equation 4) leads to the marginal likelihoods

$$p(Y_{rep}|H_i) = \int F_{df_{effect},df_{error},\lambda}(F_{rep})\pi(f^2|H_i)\, df^2 \tag{8}$$

Where $F_{df_{effect},df_{error},\lambda}(x)$ is the noncentral F-distribution with degrees of freedom df_{effect} and df_{error} and noncentrality parameter λ.

For the skeptic's position, H_0, the marginal likelihood simplifies – analogous to the t-test case – to the central F-distribution evaluated at the observed F-value from the replication study, since the prior $\pi(f^2|H_0)$ is chosen so that it is 1 at $f^2 = 0$ and 0 everywhere else:

$$p(Y_{rep}|H_0) = F_{df_{effect},df_{error}}(F_{rep}) \tag{9}$$

For the H_r model on the other hand, the prior $\pi(f^2|H_r)$ should again be the posterior of the original study. Starting out with a uniform prior before the original study results in the following posterior distribution for the original study:

$$p(f^2|Y_{orig}) = \frac{F_{df_{effect,orig},df_{error,orig},f^2 \cdot N_{orig}}(F_{orig})}{\int F_{df_{effect,orig},df_{error,orig},f^2 \cdot N_{orig}}(F_{orig})\, df^2} \tag{10}$$

Lecoutre (1999) called this distribution Λ^2-distribution, similar to the Λ'-distribution for t-tests. In contrast, this distribution cannot be easily approximated by a normal distribution (see shape of distribution in Figure 1). Despite the use of an improper prior, the posterior distribution is valid as all parameters are actually observed in the original study, i.e., $df_{effect,orig}$, $df_{error,orig}$, $F_{orig} > 0$ and $N_{orig} \gg 1$.

The Replication Bayes factor for F-tests is then given by

$$B_{r0} = \frac{\int F_{df_{effect},df_{error},f^2 \cdot N}(F_{rep})p(f^2|Y_{orig})\, df^2}{F_{df_{effect},df_{error}}(F_{rep})} \tag{11}$$

The challenge to compute the Replication Bayes factor lies primarily in the calculation of the integral of the numerator. Marginal likelihoods in general are often difficult to compute or intractable and analytical solutions are rarely available in applied settings. Verhagen and Wagenmakers (2014) chose the Monte Carlo estimator (Gamerman and Lopes, 2006, p. 239). For the Monte Carlo estimate samples are randomly drawn from the prior distribution of the H_r model, i.e. from the original study's posterior distribution

$p(f^2|Y_{orig})$, the integrand is calculated and the marginal likelihood is approximated by taking the average.

$$B_{r0} \approx \frac{1}{M} \sum_{i}^{M} \frac{F_{df_{effect}, df_{error}, f^2_{(i)} \cdot N_{rep}}(F_{rep})}{F_{df_{effect}, df_{error}}(F_{rep})}, \qquad f^2_{(i)} \sim p(f^2|Y_{orig}) \qquad (12)$$

The Monte Carlo estimate, however, is inefficient and unstable, especially when prior and likelihood disagree (i.e. when the original and the replication study yield very different effect size estimates). There are other estimators for the marginal likelihood available. Gamerman and Lopes (2006, chap. 7) provide an overview on twelve different estimators. Bos (2002) compared seven estimators in simulations and showed, that the Monte Carlo estimate is highly unstable and often yields values very different from an analytically derived solution. *Numerical integration* can work well in low-dimensional settings, but does not scale to problems involving multiple parameters or wide ranges of possible parameter values.

For complex models, *bridge sampling* provides a stable and efficient way to estimate the marginal likelihood (Meng and Wing, 1996; Gronau et al., 2017). Since the present case involves only a single parameter and relatively simple likelihood function and posterior distribution, *importance sampling* is a faster, but also efficient and stable way to estimate the marginal likelihood.

The importance sampling estimate for the marginal likelihood (Gamerman and Lopes, 2006, chap. 7) is calculated by drawing M random samples from an importance density $g(f^2)$ and averaging an adjusted likelihood term:

$$p(Y_{rep}|H_r) \approx \frac{1}{M} \sum_{i}^{M} \frac{p(Y_{rep}|\tilde{f}^2_i)\pi(\tilde{f}^2_i)}{g(\tilde{f}^2_i)}$$

$$\approx \frac{1}{M} \sum_{i}^{M} \frac{p(Y_{rep}|\tilde{f}^2_i)p(\tilde{f}^2_i|Y_{orig})}{g(\tilde{f}^2_i)}, \qquad (13)$$

$$\tilde{f}^2_i \sim g(f^2)$$

Since the prior for H_r is the posterior distribution given in Equation 10, the marginal likelihood $p(Y_{orig}|H_r)$ for the original study also needs to be computed in order for $p(f^2|Y_{orig})$ to be a proper probability density and the estimator to yield correct results.

How should the importance density $g(f^2)$ be chosen? A simple and straight forward way is to use a half-normal distribution with mean and standard deviation determined by samples from the posterior. The half-normal distribution is easy for drawing random samples and can easily be calculated at any point. While it is not as close to the true posterior (see Figure 1) as in the t-test case, it is sufficient for the purpose of importance sampling since the tails of the half-normal distribution are fatter than the non-normalized posterior density (see requirements for importance densities listed in Gronau et al., 2017).
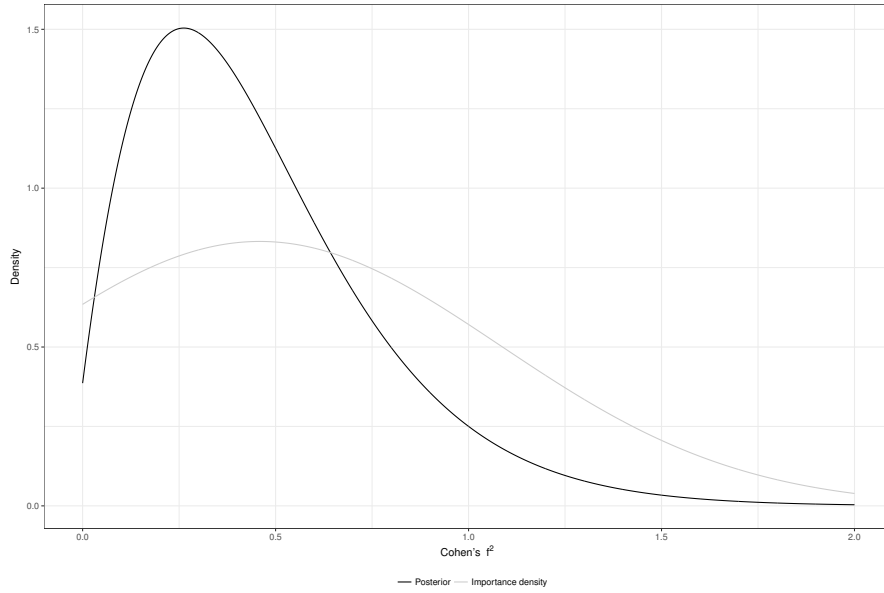
Figure 1: Plot of posterior distribution for a study with an observed $f^2 = 0.625$ in an one-way ANOVA with two groups and 10 participants each, i.e. 20 participants in total, yielding $F(1,8) = 5.0$. The grey line represents the half-normal importance density $g(f^2)$, which is constructed based on samples from the posterior distribution generated by Metropolis-Hastings, used for estimating the marginal likelihood.

To sample from the unknown and yet non-normalized posterior-distribution Markov Chain Monte Carlo (MCMC) techniques can be used, e.g. the Metropolis-Hastings algorithm (Chib and Greenberg, 1995). In Bayesian statistics approximating the posterior distribution is one of the core challenges and thus several implementations for different algorithms are available. Metropolis-Hastings yields samples from a target distribution by taking a random walk through the parameter space. An accessible introduction is given in McElreath (2016, chap. 8) and a more mathematical description can be found in Gelman et al. (2013, chap. 11.2). Software packages such as JAGS or Stan (Gelman et al., 2015) are commonly used to draw samples from the posterior distribution of Bayesian models.

The mean and standard deviation of the posterior samples are then used to construct a half-normal distribution as importance density $g(f^2)$. Subsequently, random samples from the importance density are used to estimate the marginal likelihood according to Equation 13, especially in regions where the posterior distribution has high probability mass, resulting in a better estimate compared to the Monte Carlo estimate.

For the case of the Replication Bayes factor it will be shown below in simulations, that the differences between the estimates are small except in the most extreme cases. Since it is nonetheless desirable to have accurate and robust estimates, in the remainder of the article the Importance Sampling estimate is used.

Dividing the resulting estimate for the marginal likelihood of H_r by the H_0 model evidence (Equation 11) allows the calculation of the Replication Bayes factor:

$$B_{r0} = \frac{\int F_{df_{effect}, df_{error}, f^2 \times N}(F_{rep}) p(f^2 | Y_{orig}) \, df^2}{F_{df_{effect}, df_{error}}(F_{rep})} \qquad (14)$$

The resulting Bayes factor can then be interpreted based on its quantitative value: $B_{r0} > 1$ is evidence in favor of the proponent's hypothesis, i.e. evidence in favor of a true effect of a similar size, while a $B_{r0} < 1$ is evidence against a true effect of similar size. The more the Bayes factor deviates from 1, the stronger the evidence. It might be helpful to use the commonly used boundaries of 3 and $\frac{1}{3}$ for sufficient evidence: $\frac{1}{3} < B_{r0} < 3$ is weak evidence for either hypothesis (Jeffreys, 1961, p. 432) and should lead the researcher to collect further data to strengthen the evidence (Schönbrodt et al., 2015; Edwards et al., 1963).

The provided R-package contains functions to calculate Replication Bayes factors for both t- and F-tests using the formulas provided here, relying on importance sampling to estimate marginal likelihoods.

4   SIMULATION STUDIES

In order to show properties of the Replication Bayes factors, three simulation studies are presented. First, the numerical results in different scenarios are shown to allow comparisons across different study designs and effect sizes. Second, the Monte Carlo estimate and the Importance Sampling estimate for the marginal likelihood are compared. While the differences are small except for the most extreme cases, it is argued that the more robust method is to be preferred. Last, the relationship between t- and F-values in studies of two independent groups is used to compare the Replication Bayes factor for t-tests with the proposed adaptation to the F-test.

4.1   Simulation 1: Behavior in Different Scenarios

To better understand the Replication Bayes factor, it is useful to explore different scenarios in which a researcher might calculate it. For Figure 2, different combinations for original and replication studies were considered. In particular, one can see that the Replication Bayes factor increases towards support for the proponent's position when the replication has large sample size and yields a large effect size estimate.

For situations in which the effect size estimate is very small in both the original and the replication study, the Replication Bayes factor is close to 1. In these cases, the proponent's and the skeptic's position are very similar
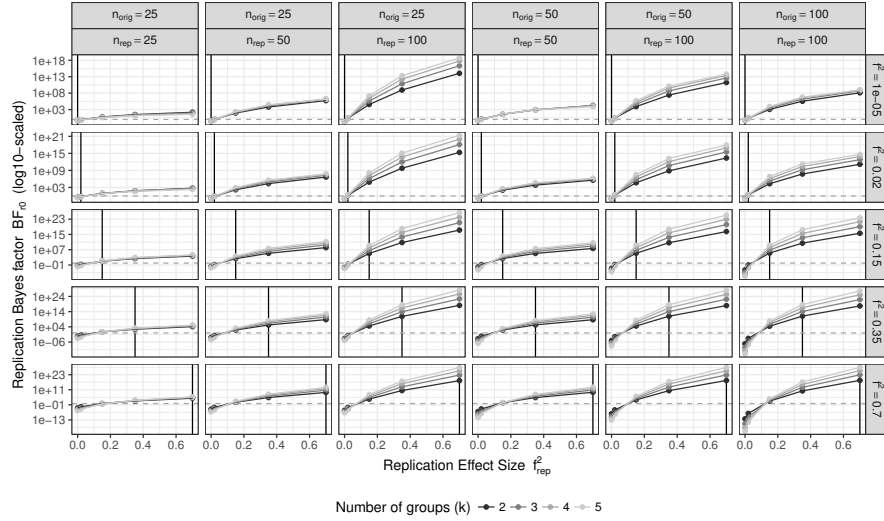
Figure 2: Value of the Replication Bayes factor for F-tests in various scenarios. Columns show sample sizes per group in original and replication study, rows are f^2 effect sizes in the original study. Horizontal axes in each plot show f^2 effect size in replication study and vertical axes are $\log_{10}$-scaled showing B_{r0}. Shades of grey denote number of groups. Each point is one calculated Replication Bayes factor in a given scenario. Vertical black lines indicate the effect size in the original study, i.e. points on this line are replication studies where the effect size equals the original study. Horizontal dashed line is $B_{r0} = 1$.

and even 100 participants per group are not enough to properly distinguish between the two models (see left most points in first row plots of Figure 2). If, in contrast, the original study reported a relatively large effect and the replication study yields a small effect, the Replication Bayes factor quantifies this correctly as evidence in favor of the skeptic's positions. Moreover, the considered setup shows that – holding group sizes equal – more groups allow stronger conclusions since total sample size is higher.

4.2 Simulation 2: Monte Carlo vs Importance Sampling Estimate

The aforementioned difference between the Monte Carlo estimate for the marginal likelihood and an importance sampling estimate is especially relevant if prior and likelihood of a model disagree substantially. For the Replication Bayes factor this is the case when original study and replication study yield very different effect size estimates.

As referenced above, Bos (2002) has shown in simulations that the Monte Carlo estimate is generally unstable and leads to biased estimates of the marginal likelihood. He compared the results both against the analytical solution and other estimators, which performed substantially better.

In order to evaluate how this affects the Replication Bayes factor, pairs of studies are investigated. For illustrative purposes, original studies with $n_{orig} = 15$ per group with an observed effect size of $d_{orig} \in \{1; 2; 5\}$ and

subsequent replications with sample sizes $n_{rep} \in \{50; 100\}$ per group and observed effect sizes $d_{rep} \in \{0; 0.3; 0.5\}$ are entered in the Replication Bayes factor for t-tests. The resulting Bayes factor is estimated using the Monte Carlo estimate and the Importance Sampling estimate.

As can be seen from the results in Figure 3, the Monte Carlo estimate is generally close to identical to the importance sampling estimate. It is only in extreme cases that there are differences. For example, when an original study yields $d = 5$ in a small sample of $n = 15$ per group and a replication shows only a tiny or even a null effect in a much larger sample. The effect size of $d = 5$ is, however, an implausibly large effect size for most empirical research fields. Even when considering the larger sampling error in small samples, effect sizes around $d = 1$ are already considered a "large effect" in social sciences (Cohen, 1988) – effects of $d > 3$ do not realistically appear in the literature in psychology.

While the differences can become large in orders of magnitude, in these cases the Replication Bayes factor is also large in general because of the disagreement between original and replication study. Thus, the conclusions are not changed by the difference between estimates. Furthermore, in the practical examples reported below the differences between the two estimates are numerically very small and would not yield different conclusions.

Nevertheless, it should be desirable to provide an accurate estimates of a statistical indicator to make an informed judgement, e.g. about the relative evidence a replication study can provide. The Importance Sampling estimate has been shown to be more stable and more accurate and is thus preferable to the Monte Carlo estimate (Bos, 2002; Gronau et al., 2017; Gamerman and Lopes, 2006).

4.3 Simulation 3: F-test for two groups

In the context of significance testing, it is a well known relationship that the F-tests from a one-way ANOVA with two groups yield the same p-values as a two-sided, two-sample t-test, when $F = t^2$ is used. Accordingly, it might be an interesting question, how the Replication Bayes factor for F-tests relates to the Replication Bayes factor for t-tests when used on the same data from two independent groups.

For pairs of original and replication studies with two groups of different sizes ($n_{orig} \in \{15; 50\}$, $n_{rep} \in \{15; 30; 50; 100\}$) and different effect sizes ($d_{orig} \in \{0.2; 0.4; 0.6; 0.8; 1; 2\}$, $d_{rep} \in \{^1/_{10^5}; 0.2; 0.4; 0.6; 0.8; 1; 2\}$), the Replication Bayes factor for t-test was calculated using the observed t-value and for the F-test using $F = t^2$.

The results are shown in Figure 4. As can be expected, the resulting Bayes factors are very close ($r = 0.999$ across all scenarios). What cannot easily be seen in the figure, however, is that the Replication Bayes factor for
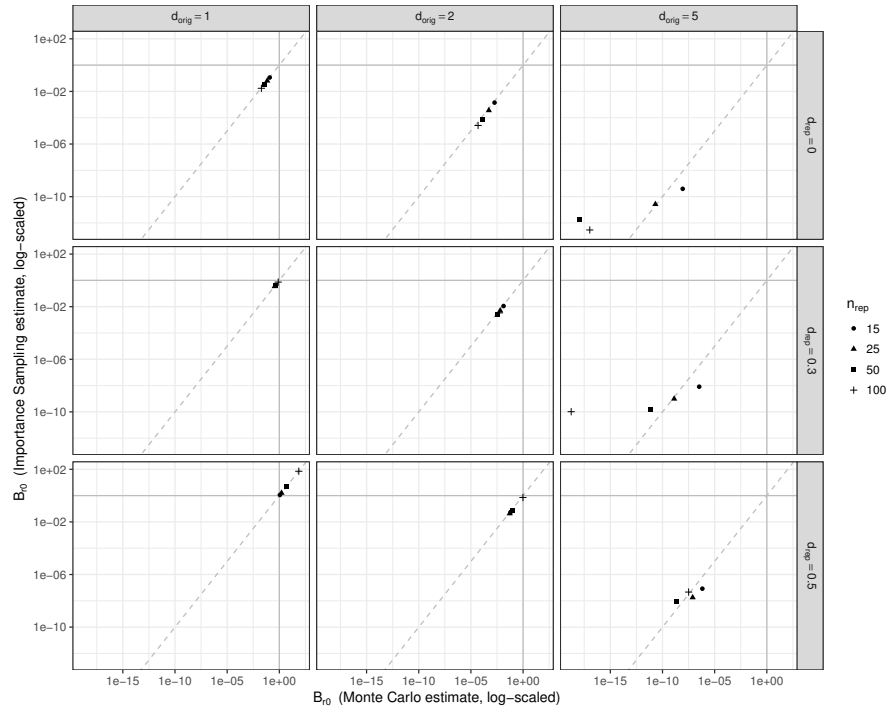
Figure 3: Comparison for Replication Bayes factors (t-test) with marginal likelihoods estimated by either Monte Carlo estimation (x-axis) or Importance Sampling estimation (y-axis). Points are individual Bayes factors. Grey dashed line indicates equality between both estimation methods. Only for extreme cases (i.e. $d_{orig} = 0.5$ and $d_{rep} > 2$) the estimations yield substantially different results, but conclusions would remain the same. Original studies always had $n_{orig} = 15$.

the F-test is about half the size of the Replication Bayes factor for the t-test ($\frac{B_{r0,t}}{B_{r0,F}} = 2.211$).

This result makes intuitively sense: The F-statistic does not contain information about the direction of the effect and thus cannot provide the same amount of relative evidence as a Bayes factor on the t-test.
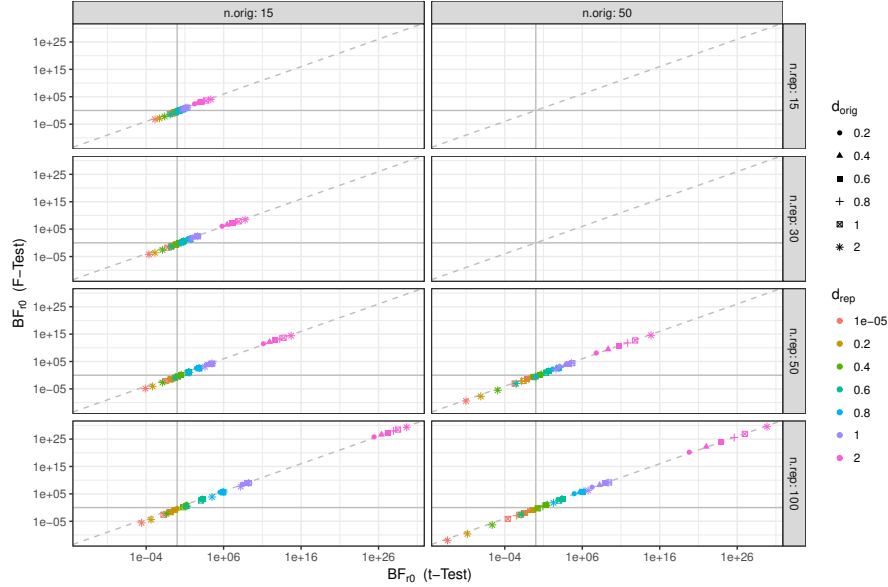

Figure 4: Comparison of Replication Bayes factors for t- and F-test on the same data-set with two groups. Study pairs with replication studies smaller than original studies were not simulated. Dashed grey line is equality.

5   EXAMPLES

In this section, the Replication Bayes factor for F-tests is applied in two example cases of replication studies from the "Reproducibility Project: Psychology" (Open Science Collaboration, 2015). The third example presented in this section aims to show how to investigate the pattern of effects to ensure valid conclusions based on a high value of the Replication Bayes factor.

5.1   Example 1

The first example is an original study conducted by Albarracín et al. (2008), which was replicated as part of the "Reproducibility Project: Psychology" (Open Science Collaboration, 2015) by Voracek and Sonnleitner (2016) and is available at https://osf.io/tarp4/.

Albarracín et al. (2008) investigated the effect "action" and "inaction goals" on subsequent "motor or cognitive output". In study 7 specifically, participants were primed with either words relating to "action", "inaction" or neutral words (control condition). Participants subsequently engaged in

either an active or inactive task before they were instructed to read a text and write down their thoughts about the text. The number of listed thoughts was used as a measure for "cognitive activity". The experiment thus had a 3 (*Prime*: action, inaction, control) × 2 (*Task:* active, inactive) between-subjects design.

Participants were predicted to write down more thoughts about the text when they are primed with an "action" goal compared to participants primed with an "inaction" goal. Futhermore, the possibility to exert activity in an active task, should moderate the effect: "satisfied action goals should yield less activity than would unsatisfied action goals". They found the two-way interaction *Prime* × *Task* to be significant ($F(2, 92) = 4.36$, $p = .02$, $\eta_p^2 = 0.087$ corresponding to $f^2 = 0.095$) in a sample of 98 student participants (group sizes were not reported).

The replication by Sonnleitner and Voracek did not find the same interaction to be significant in their sample of 105 participants, $F(2, 99) = 2.532$, $p = .085$, $\eta_p^2 = 0.049$, $f^2 = 0.051$.

Comparing the replication attempt to the original study using the Replication Bayes factor assuming equally sized groups yields $B_{r0} = 1.153$, which can be considered inconclusive as the data are nearly equally likely under both models. Based on the p-values, it was concluded that the study was not successfully replicated. Judging from the Bayes factor, however, it should be concluded that the replication study did not have enough participants to indicate stronger evidence in either direction. Considering the power of a frequentist test, recent recommendations suggest at least 2.5 times as many participants than in the original study one seeks to replicate (Simonsohn, 2015). The Replication Bayes factor thus shows absence of evidence rather than evidence of absence of a failed or successful replication.

For the main effect of *Task*, in contrast, which also was significant in the original study ($F(1, 92) = 7.57$, $p = .007$, $\eta_p^2 = 0.076$, $f^2 = 0.082$) but not in the replication ($F(1, 99) = .107$, $p = .745$, $\eta_p^2 \approx f^2 < .001$), the Replication Bayes factor yields strong evidence in favor of the skeptic's model ($B_{r0} = 0.057$). This is evidence in favor of an unsuccessful replication of the main effect, since the Bayes factor shows that the data are 17 times more likely under the model implying an effect size $f^2 = 0$ than under the model informed by the original study. It should be noted however, that this does not answer the question whether the replication is in line with an existing albeit smaller effect (which the replication also did not have sufficient power to detect, see Simonsohn, 2015).

5.2 Example 2

For the second example another replication from the "Reproducibility Project" is considered: The original study was conducted by Williams and Bargh (2008) and investigated cues of "spatial distance on affect and evaluation".

The replication was performed by Joy-Gaba et al. (2016). The replication data, materials and final report are available at `https://osf.io/vnsqg/`.

In study 4 of the original paper, Williams and Bargh (2008) have primed 84 participants in three different conditions. The number of participants per group was not reported. The authors hypothesized that different primes for spatial distance will effect evaluations of perceived "closeness" to siblings, parents and hometown. The dependent variable was an index of ratings to those three evaluations. Hence, the study was a between-subjects design with one factor (*Prime*: Closeness, Intermediate, Distance). They found a significant main effect of priming on the "index of emotional attachment to one's nuclear family and hometown" in a one-way ANOVA ($F(2, 81) = 4.97$, $p = .009$, $\eta_p^2 = .11$, $f^2 = .124$).

The replication by Joy-Gaba, Clay and Cleary did not find the same main effect in a sample of 125 participants ($F(2, 122) = .24$, $p = .79$, $\eta_p^2 = .003919$, $f^2 = .00393$). Based on the p-values they concluded, that the replication was not successful.

But how much more are the data in line with a null model when compared to the proponent's alternative? This is the answer the Replication Bayes factor can give, assuming equal group sizes: $B_{r0} = 0.031$. This means, the data is about 32 times more likely under the model stating that the true effect size is 0 than under the model using the original study's posterior.

5.3 Example 3

The final example aims to show a caveat when using the Replication Bayes factor for F-tests yielding strong support for the proponent's model. As has been addressed above and could be seen in the last simulation, the F-statistic (and, consequently, the f^2 effect size measure) does not convey information about the location or direction of an effect. This is a general problem when evaluating the outcomes from ANOVA-design studies based on the test-statistic alone. Researchers need additional judgments based on post-hoc t-tests or qualitative consideration of interaction plots.

In order to show how to inspect the results thoroughly, an imaginary study is conducted: In an original study, 15 participants each are randomly assigned to three different conditions (45 participants in total). The true population means of the three groups are $\mu_1 = 1.5$, $\mu_2 = 2.2$ and $\mu_3 = 2.9$ and standard deviation is 1 for all groups. Running a one-way ANOVA on a generated data-set yields a significant result, $F(2, 42) = 7.91$, $p = .001$, $f^2 = 0.377$. For the replication study, 30 participants are randomly sampled to the same three conditions each (thus 90 participants in total). For the generated replication data-set the ANOVA yields a significant result as well, $F(2, 87) = 7.60$, $p < .001$, $f^2 = 0.175$.

The test-statistics and effect size estimates show a significant effect in both studies. Did the replication happen to successfully replicate the orig-
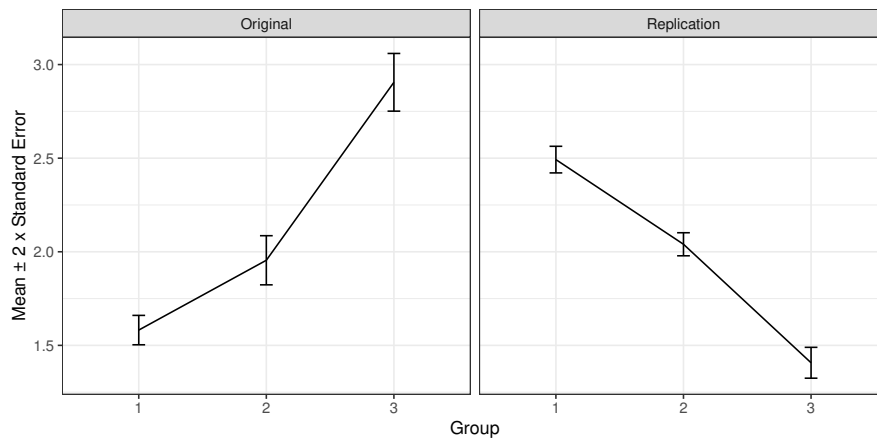
Figure 5: Summary plot for the imaginary study in Example 3. Original study has 15 participants per group with population means $\mu = (1.5; 2.2; 2.9)$. Replication has 30 participants per group with population means in reverse order.

inal finding? Before calculating the Replication Bayes factor, one should inspect the data in more detail. Considering the the plot in Figure 5, it is obvious that the pattern of effects is strikingly different. In fact, the replication shows the reverse pattern. Since the F-statistic is insensitive to this difference, the Replication Bayes factor does not provide us this information. Yet, it does correctly indicate that the data are more in favor of an effect in size – not direction – of the original study ($B_{r0} = 38.261$).

Considering post-hoc t-tests or planned contrasts helps to include location and direction of the effect in the analysis. For the present example, for example, the difference between groups 1 and 3 is significant in both the original study ($t(20.809) = -3.953$, $p = .001$) and the replication study ($t(56.765) = 3.6412$, $p < .001$). The difference in direction is visible in the sign of the t-statistic and thus the Replication Bayes factor for t-test can consider this information ($B_{r0} = 0.0015$ or the reciprocal $B_{0r} = 668.15$)

Consequently, researchers investigating F-tests should always pay attention to the particular nature of the effect under investigation. This is a recommendation not limited to the evaluation of replication studies, but is generally a crucial step when analyzing data from ANOVA-studies.

6   DISCUSSION

The Replication Bayes factor for F-tests in fixed-effect ANOVA designs outlined in this paper is an extension of the work by Verhagen and Wagenmakers (2014). It utilizes a Bayesian perspective on replications, namely using the results and uncertainty from the original study in the analysis of a replication attempt. The approach outlined in this paper adapts the Replication Bayes factor from t- to F-tests.

In evaluating replication studies it is reasonable to use the information available from the original study. When p-values of two studies are directly compared based on their significance, or when a confidence interval is used to examine whether the effect size in the original can be rejected, this is already essentially done. The Bayesian framework allows to do this in a more formal framework and incorporates uncertainty in estimates. The latter is ignored when only p-values are compared. This is one of the reasons why "vote counting" is not recommended when evaluating a replication study.

A general criticism of Bayesian hypothesis testing with Bayes factors is the role of the prior. While proponents of Bayesian statistics generally consider incorporating previous knowledge or expectation as a strength of Bayesian statistics, the perceived subjectivity in its selection is troublesome for some non-Bayesians. In contrast to other ways of including the results from the original study in Bayesian model (see below), the Replication Bayes factor introduces very little subjectivity to the analyses. The model for the Replication Bayes factor is derived directly from the assumptions of null-hypothesis significance testing.

While Bayes factors do not provide controls for error rates over many repeated samplings, they do allow for a quantification of relative model evidence. That is, a Bayes factor allows statements such as "the data are 5-times more likely under the alternative model than under the null model". Researchers reluctant to the Bayesian approach in general might nevertheless find the Replication Bayes factor in particular a useful addition without adding too much subjectivity. After all, the evaluation of replication studies should not rely on a single method (Marsman et al., 2017; Brandt et al., 2014; Anderson and Maxwell, 2016; Gilbert et al., 2016).

The steps taken and explained in this paper can also be used to apply the Replication Bayes factor logic to other tests such as χ^2-tests, z-tests, or correlations (see also Boekel et al., 2015, Appendix). One needs to use the appropriate test-statistic and sampling distribution under the alternative hypothesis and accordingly derive the resulting marginal likelihoods. The use of Importance Sampling for estimating the marginal likelihoods allows for more general cases than the Monte Carlo estimate; on the one hand it is more robust to large disagreements between original study and replication study (see Simulation 3) and, on the other hand, scales better if more parameters were introduced in the model.

6.1 Limitations

As explained above, Cohen's f^2 was chosen as the parameter of interest. This limits the application of the proposed Replication Bayes factor for F-tests to fixed-effects ANOVAs with approximately equal cell sizes. The relationship between the non-centrality parameter λ and f^2 is only valid in these cases.

In general, effect size measures in ANOVAs take into account the specific effects (i.e. each cell mean's deviation from the grand mean Steiger, 2004) and cell sizes. In many reported ANOVAs, however, only the total sample size along with the omnibus or interaction test is reported. Even if specific contrasts are of interest, they are not reported with sample sizes or descriptive statistics required to calculate the specific effects.

What happens when f^2 is used in unbalanced designs? If a specific effect is present in a group with larger sample size, f^2 will overestimate the overall effect. For studies with randomized allocation to the groups, the differences in cell sizes should be small, so the assumption of balancedness seems warranted. In non-random studies where unbalanced groups are to be expected, the Replication Bayes factor as outlined here should not be used. More elaborate alternatives as suggested below might be more appropriate.

The Replication Bayes factor is also limited by the information provided in the F-statistic: Since the F-statistic (and consequently f^2) does not convey information about the direction and location of the effect, the Replication Bayes factor cannot take the pattern of effects into account. The Replication Bayes factor still does give the relative evidence regarding the size of the effect. As shown in Example 3, one way is to follow-up evidence in favor of a replicated effect by investigating the pattern in the post post-hoc t-tests or contrasts. If no such tests are reported, one could use the Replication Bayes factor for the omnibus test as a quantitative indicator and use visual inspection whether the effect is indeed in the same direction. This is in parallel to the recommended practice for analyzing data with ANOVAs. Another way of dealing with this could be to extend the model used in the Replication Bayes factor to include the descriptive statistics of all groups instead of the test-statistic. This, however, would require additional assumptions and does not follow as directly from the significance tests as the method proposed here. The Bayesian alternatives mentioned below can be a starting point for further extensions of the model.

What is true for p-values (Wasserstein and Lazar, 2016) also holds true for Bayes factors: No single statistical parameter will take the responsibility from researchers to make a careful evaluation of the different statistical results at hand. Even if all statistical parameters are in agreement, differences in methodology and design might render the statistical comparison *ad absurdum*.

Last, effect size estimates are more difficult to compute for random- and mixed-effects ANOVAs. While effect size measures such as ω^2 can be calculated in these scenarios, the relationship to the non-central F-distribution is not as obvious. Further developments of the Replication Bayes factor might be useful in this direction.

6.2   Alternatives to the Replication Bayes factor

The Replication Bayes factor is only one way to evaluate a replication study in the Bayesian framework. As outlined above, its strength lies in making little assumptions and the ability to use only reported test-statistic for the evaluation.

One alternative way to use Bayes factors would be to model the information from the original study differently. Where the posterior of the original study is used here as a prior in the replication model, one could also use the effect size estimate and a reported confidence interval to construct a Normal distribution as prior. Statistical software such as JASP (JASP Team, 2018) allows to use different priors, including normal distributions, to calculate Bayes factors from test-statistics.

Another way to use the Bayesian framework would be not to rely on Bayes factors but use Bayesian estimation in hierarchical models and to incorporate both the original and the replication study to estimate the effect size across both studies. Etz and Vandekerckhove (2016) have used such models to evaluate the outcomes of the "Reproducibility Project: Psychology". Some of the methods outlined in this article can also be used for single study pairs.

Recently, Ly et al. (2017) have proposed a reconceptualization of the Replication Bayes factor. The goal is to avoid the computations for the posterior-turned-prior distribution and instead rely on "evidence updating". Different Bayes factors can be multiplied yielding the Replication Bayes factor $BF_{10}(d_{rep}|d_{orig})$ (Ly et al., 2017, p. 7):

$$BF_{10}(d_{rep}|d_{orig}) = \frac{BF_{10}(d_{orig}, d_{rep})}{BF_{10}(d_{orig})}$$

This requires the computation of a Bayes factor for the original study alone, $BF_{10}(d_{orig})$, and a Bayes factor of a combined data set from both studies, $BF_{10}(d_{orig}, d_{rep})$. In their pre-print, Ly et al. (2017) outline how to calculate this Bayes factor from "Evidence Updating" for t-tests and contingency tables.

The calculation of an F-value representing a combined data-set requires reported means, standard deviations and group sizes for both the original and replication study. These are not always reported and the Replication Bayes factor outlined here is designed to not require it. Therefore the extension of the original concept for the Replication Bayes factor (Verhagen and Wagenmakers, 2014) is useful when only limited data is available from a published article.

Last, the Replication Bayes factor uses the effect size estimate from the original study at face value. This is in many cases a questionable assumption, since publication bias and p-hacking are known to inflate effect size estimate in the reported literature. For sample size planning it is relevant to account for this, e.g. by planning a study at least about 2.5-times as large

as the original study (Simonsohn, 2015). For the analysis of a new replication study on an effect which might not yet have been estimated through meta-analyses the comparison with an originally reported effect size is useful and often the first step in an analysis. The Replication Bayes factor can then answer the question: How much more evidence in favor of an effect of this size does the data provide when compared to a null effect? A Bayes factor using a manually (i.e. more subjectively) chosen prior, as mentioned above, would be able to incorporate a corrected effect size. By what factor or procedure one should correct a reported effect size is an open question in meta-analytical research. Gelman (2014) has used the term "Edlin factor" for this: a measure by which published estimates should be correct by.

6.3 Conclusion

The Replication Bayes factor introduced by Verhagen and Wagenmakers (2014) and extended herein is one index to evaluate the results of a replication attempt. It is, obviously, not able to cover all questions and pitfalls in the analysis of a replication. Instead it is a way to formally and transparently integrate previously available information in the analysis within the Bayesian framework and allows to quantitatively assess the gained evidential value. It is further easy to apply to frequentist results as it uses reported test statistics from the original and replication study only. To cover replications comprehensively, however, researchers have to use different tools depending on the question asked. No single statistical index is sufficient to globally assess the quality of a study or a theory. This is true not only for p-values (Wasserstein and Lazar, 2016) but also for Bayes factors.

DATA ACCESSIBILITY

To reproduce the examples in this article, all scripts are available in an OSF repository at https://osf.io/jv39h/ (Harms, 2018). To calculate the Replication Bayes Factor for t-tests (Verhagen and Wagenmakers, 2014) and F-tests (the present article), an R-Package is available at https://github.com/neurotroph/ReplicationBF.

ACKNOWLEDGEMENTS

CONFLICTS OF INTEREST

None.

REFERENCES

Albarracín, D., Handley, I. M., Noguchi, K., McCulloch, K. C., Li, H., Leeper, J., Brown, R. D., Earl, A., and Hart, W. P. (2008). Increasing and decreasing motor and cognitive output: A model of general action and inaction goals. *Journal of Personality and Social Psychology*, 95(3):510–523.

Anderson, C. J., pan Bahnik, t., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., Cheung, F., Christopherson, C. D., Cordes, A., Cremata, E. J., Della Penna, N., Estel, V., Fedor, A., Fitneva, S. A., Frank, M. C., Grange, J. A., Hartshorne, J. K., Hasselman, F., Henninger, F., van der Hulst, M., Jonas, K. J., Lai, C. K., Levitan, C. A., Miller, J. K., Moore, K. S., Meixner, J. M., Munafo, M. R., Neijenhuijs, K. I., Nilsonne, G., Nosek, B. A., Plessow, F., Prenoveau, J. M., Ricker, A. A., Schmidt, K., Spies, J. R., Stieger, S., Strohminger, N., Sullivan, G. B., van Aert, R. C. M., van Assen, M. A. L. M., Vanpaemel, W., Vianello, M., Voracek, M., and Zuni, K. (2016). Response to Comment on "Estimating the reproducibility of psychological science". *Science*, 351(6277):1037–1037.

Anderson, S. F. and Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, 21(1):1–12.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., van Aken, M. A. G., Weber, H., and Wicherts, J. M. (2013). Recommendations for Increasing Replicability in Psychology. *European Journal of Personality*, 27(2):108–119.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66(6):423–437.

Bayarri, M. J., Benjamin, D. J., Berger, J. O., and Sellke, T. M. (2015). Rejection Odds and Rejection Ratios: A Proposal for Statistical Practice in Testing Hypotheses. *Journal of Mathematical Psychology*.

Belia, S., Fidler, F., Williams, J., and Cumming, G. (2005). Researchers Misunderstand Confidence Intervals and Standard Error Bars. *Psychological Methods*, 10(4):389–396.

Boekel, W., Wagenmakers, E.-J., Belay, L., Verhagen, J., Brown, S., and Forstmann, B. U. (2015). A purely confirmatory replication study of structural brain-behavior correlations. *Cortex*, 66:115–133.

Bos, C. S. (2002). A Comparison of Marginal Likelihood Computation Methods. In *Compstat*, pages 111–116. Physica-Verlag HD, Heidelberg.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. a., Perugini, M., Spies, J. R., and van 't Veer, A. (2014). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, 50(1):217–224.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49(4):327.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2nd ed. edition.

Cumming, G. (2008). Replication and p Intervals: p Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better. *Perspectives on Psychological Science*, 3(4):286–300.

Cumming, G. (2012). *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. Routledge, New York.

Cumming, G. and Finch, S. (2001). A Primer on the Understanding, Use, and Calculation of Confidence Intervals that are Based on Central and Noncentral Distributions. *Educational and Psychological Measurement*, 61(4):532–574.

Dienes, Z. (2016). How Bayes factors change scientific pratice. *Journal of Mathematical Psychology*, pages 1–29.

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

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

Etz, A. and Vandekerckhove, J. (2016). A Bayesian Perspective on the Reproducibility Project: Psychology. *PLOS ONE*, 11(2):e0149794.

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall/CRC, Boca Raton, FL.

Gelman, A. (2014). "edlin's rule" for routinely scaling down published estimates – statistical modeling, causal inference, and social science.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press, Boca Raton, FL, third edit edition.

Gelman, A., Lee, D., and Guo, J. (2015). Stan: A Probabilistic Programming Language for Bayesian Inference and Optimization. *Journal of Educational and Behavioral Statistics*, 40(5):530–543.

Gelman, A. and Stern, H. (2006). The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. *The American Statistician*, 60(4):328–331.

Gilbert, D. T., King, G., Pettigrew, S., and Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science*, 351(6277):1037–1037.

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

Harms, C. (2018). A bayes factor for replications of anova results.

Ionides, E. L., Giessing, A., Ritov, Y., and Page, S. E. (2017). Response to the ASA's Statement on p -Values: Context, Process, and Purpose. *The American Statistician*, 71(1):88–89.

JASP Team (2018). JASP (Version 0.8.6) [Computer software].

Jeffreys, H. (1961). *Theory of Probability*. Clarendon Press, Oxford, 3. ed. edition.

Joy-Gaba, J. A., Clay, R., and Cleary, H. (2016). Replication of Williams & Bargh (2008, PS, Study 4).

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kruschke, J. K. (2015). Introduction: Credibility, Models, and Parameters. In *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*, chapter 2, pages 15–32. Academic Press / Elsevier, Waltham, MA, 2nd ed. edition.

Lecoutre, B. (1999). Two useful distributions for Bayesian predictive procedures under normal models. *Journal of Statistical Planning and Inference*, 79(1):93–105.

Lindley, D. V. (1993). The Analysis of Experimental Data: The Appreciation of Tea and Wine. *Teaching Statistics*, 15(1):22–25.

Ly, A., Etz, A., Marsman, M., and Wagenmakers, E.-J. (2017). Replication Bayes Factors from Evidence Updating.

Marsman, M., Schönbrodt, F. D., Morey, R. D., Yao, Y., Gelman, A., and Wagenmakers, E.-J. (2017). A Bayesian bird's eye view of 'Replications of important results in social psychology'. *Royal Society Open Science*, 4(1).

Maxwell, S. E., Lau, M. Y., and Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, 70(6):487–498.

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press.

Meng, X. and Wing, H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 6:831–860.

Morey, R. D. and Lakens, D. (2016). Why most of psychology is statistically unfalsifiable.

Morey, R. D., Romeijn, J.-W., and Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5(2):241–301.

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

Pashler, H. and Wagenmakers, E. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science. *Perspectives on Psychological Science*, 7(6):528–530.

Rosenthal, R., Rosnow, R. L., and Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge University Press.

Rouder, J. N. and Morey, R. D. (2012). Default Bayes Factors for Model Selection in Regression. *Multivariate Behavioral Research*, 47(6):877–903.

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

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

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., and Perugini, M. (2015). Sequential Hypothesis Testing With Bayes Factors: Efficiently Testing Mean Differences. *Psychological Methods*.

Sedlmeier, P. and Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105(2):309–316.

Simonsohn, U. (2015). Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychological Science*, 26(5):559–569.

Steiger, J. H. (2004). Beyond the F Test: Effect Size Confidence Intervals and Tests of Close Fit in the Analysis of Variance and Contrast Analysis. *Psychological Methods*, 9(2):164–182.

Szucs, D. and Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, 15(3):e2000797.

Verhagen, J. and Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, 143(4):1457–1475.

Voracek, M. and Sonnleitner, C. (2016). Replication of D Albarracín, IM Handley, K Noguchi, KC McCulloch, H Li, J Leeper, RD Brown, A Earl, WP Hart (2008, JPSP 95(3), Exp. 7).

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's Statement on p -Values: Context, Process, and Purpose. *The American Statistician*, 70(2):129–133.

Williams, L. E. and Bargh, J. a. (2008). Keeping One's Distance: The Influence of Spatial Distance Cues on Affect and Evaluation. *Psychological Science*, 19(3):302–308.