The sceptical Bayes factor for the assessment of replication success

Samuel Pawel, Leonhard Held
Epidemiology, Biostatistics and Prevention Institute (EPBI)
Center for Reproducible Science (CRS)
University of Zurich, Switzerland
E-mail: samuel.pawel@uzh.ch

September 3, 2020

Abstract
There is an urgent need to develop new methodology for the design and analysis of replication studies. Recently, a reverse-Bayes method called the sceptical p-value has been proposed for this purpose; the inversion of Bayes’ theorem allows us to mathematically formalise the notion of scepticism, which in turn can be used to assess the agreement between the findings of an original study and its replication. However, despite its Bayesian nature, the method relies on tail probabilities as primary inference tools. Here, we present an extension that uses Bayes factors as an alternative means of quantifying evidence. This leads to a new measure for evaluating replication success, the sceptical Bayes factor: Conceptually, the sceptical Bayes factor provides a bound for the maximum level of evidence at which an advocate of the original finding can convince a sceptic who does not trust it, in light of the replication data. While the sceptical p-value can only quantify the conflict between the sceptical prior and the observed replication data, the sceptical Bayes factor also takes into account how likely the data are under the posterior distribution of the effect conditional on the original study, allowing for stronger statements about replication success. Moreover, the proposed method elegantly combines traditional notions of replication success; it ensures that both studies need to show evidence against the null, while at the same time penalising incompatibility of their effect estimates. Case studies from the Reproducibility Project: Cancer Biology and the Social Sciences Replication Project show the advantages of the method for the quantitative assessment of replicability.

Key words: Replication, reverse-Bayes, Bayesian hypothesis testing, Bayes factor, sceptical p-value

1 Introduction
As a consequence of the so-called replication crisis, the conduct of replication studies has increased considerably (Errington et al., 2014; Klein et al., 2014; Open Science Collaboration, 2015; Camerer et al., 2016, 2018; Cova et al., 2018). Despite the fact that most researchers agree on the importance of replication studies, there is currently no agreement on a statistical criterion for replication success. Instead, replication projects typically report the results of several different analysis approaches which can roughly be divided by two spectra: Frequentist or Bayesian and based on hypothesis testing or parameter estimation.

Among frequentist methods, assessing whether original and replication studies both achieve statistical significance (“vote-counting”) is most commonly used but has also been criticized for many reasons. For example, non-significant replication results are expected if the original finding is a false positive (e.g., with 95% probability if the significance level is 5%). They are
also expected with non-negligible probability if the underlying effect is present (Goodman, 1992; Senn, 2002). Conversely, when the effect estimate of the replication is much smaller than the estimate from the original study, statistical significance can still be achieved by increasing the sample size of the replication.

Proponents of parameter estimation rather assess the compatibility between the effect estimates from original and replication study. A popular method is to examine whether the replication effect estimate is within a 95% prediction interval based on the original effect estimate (Patil et al., 2016) which is equivalent to a meta-analytic Q-test of both estimates at the 5% significance level. Several extensions of this method exist (see e.g. Mathur and VanderWeele, 2017; Pawel and Held, 2020). However, these procedures typically have very low power, especially when only a single replication is conducted (Hedges and Schauer, 2019). If the studies are underpowered estimates that go in opposite directions may even be considered compatible which seems to contradict with common intuition about replication success. An estimation-based approach that goes one step further is to synthesise both effect estimates using meta-analysis. When conducting a replication study, researchers want to challenge the findings from the original study results with the results from the replication study. It is questionable whether combining the effect estimates from both studies and treating them as interchangeable is a reasonable way to answer this question, especially in the presence of publication bias and questionable research practices (Held, 2020b).

In the Bayesian framework only a few contributions have been made regarding the assessment of replication success. Bayarri and Mayoral (1999, 2002b,a) consider a hierarchical model of effect estimates in the replication context which can be used for planning and analysis of replication studies. They explored several analysis strategies to quantify replication success, e.g. Bayesian hypothesis testing using Bayes factors, but also the analysis of the width of a highest posterior density interval of the difference between the effect estimates. A Bayesian meta-analysis method to synthesise original and replication estimates that adjusts for publication bias was proposed by van Aert and van Assen (2017). Verhagen and Wagenmakers (2014); Ly et al. (2018); Harms (2019) introduced and refined the replication Bayes factor which quantifies the evidence for the null hypothesis of no effect against the alternative hypothesis that the effect is distributed according to the effects’ posterior distribution conditional on the data from the original study. The replication Bayes factor was subsequently used in the analysis of the Social Sciences Replication Project (Camerer et al., 2018).

A method that offers a Bayes–frequentist compromise, the sceptical p-value, was recently proposed by Held (2020b). In this approach, the prior distribution of the effect is determined such that conditional on the original study, the \((1 - \alpha)\) credible interval of the posterior distribution of the effect just includes zero. This prior corresponds to the objection of a sceptic who argues that the original finding is no longer significant when combined with a sufficiently sceptical prior. Replication success at level \(\alpha\) is now achieved if the tail probability of the replication estimate under its prior predictive distribution is smaller than \(\alpha\), rendering the sufficiently sceptical prior unrealistic. The smallest level \(\alpha\) at which replication success can be declared corresponds to the sceptical p-value, similar to the duality of ordinary p-values and confidence intervals. The method comes with appealing properties: The sceptical p-value is never smaller than the ordinary p-values from both studies ensuring that they both provide evidence against the null, while it also takes into account the size of their effect estimates, penalising the case when the replication estimate is smaller than the original estimate.

Despite the methods’ Bayesian nature, it relies on tail probabilities as primary inference tools. An attractive alternative is the Bayes factor since it allows for direct quantification of evidence for one hypothesis versus another, whereas the p-value can do so only indirectly. For this reason, we extended the reverse-Bayes procedure from Held (2020b) to use Bayes factors for the purpose of quantifying evidence. This extension was suggested by Consonni (2019) and Pericchi (2020) independently in response to Held (2020b).

The inclusion of Bayes factors lead to a new quantity which we call the sceptical Bayes factor. Unlike standard forward-Bayes methods, but similar to the sceptical p-value, the proposed
method elegantly combines traditional notions of replication success: It ensures that both studies need to show evidence against the null while at the same time penalising incompatibility of their effect estimates. However, while the sceptical $p$-value quantifies compatibility only indirectly through conflict with the sceptical prior, the sceptical Bayes factor evaluates directly how likely the replication data are to occur under an advocacy prior (the posterior of the effect conditional on the original study). This more direct assessment of compatibility allows for stronger statements about the degree of replication success.

This paper is structured as follows: Section 2 provides definition and properties of the sceptical Bayes factor along with a comparison to other measures of replication success. Case studies from the Reproducibility Project: Cancer Biology (Errington et al., 2014) (Section 3.1) and the Social Sciences Replication Project (Camerer et al., 2018) (Section 3.2) illustrate how the method works in practice. Section 4 provides a discussion of the strengths, limitations, and extensions of the method. Finally, the appendices give insight into some technical details.

2 Methods

2.1 Notation and assumptions

We denote the underlying effect by $\theta$ and effect estimates by $\hat{\theta}_o$ and $\hat{\theta}_r$ with their subscript indicating whether they come from the original or the replication study. Let the corresponding (known) standard errors be denoted by $\sigma_o$ and $\sigma_r$, and the sceptical prior variance by $\sigma_s^2 = g \cdot \sigma_o^2$ where $g = \sigma_s^2 / \sigma_o^2$ is the relative sceptical prior variance (relative to the variance of the original effect estimate $\sigma_o^2$). We define the variance ratio as $c = \sigma_o^2 / \sigma_r^2$ and denote the $z$-values of original and replication study by $z_o = \hat{\theta}_o / \sigma_o$ and $z_r = \hat{\theta}_r / \sigma_r$.

All concepts discussed in this paper are derived under the assumption that after a suitable transformation, effect estimates are normally distributed with known variances. The normal model in combination with conjugate priors makes it possible to obtain closed-form expressions in many cases which allows us to easily study limiting behaviour and facilitates interpretability. This framework is similar to standard meta-analysis and covers a wide range of practically relevant scenarios.

2.2 Bayes factors

In the Bayesian hypothesis testing framework, the Bayes factor (BF) is a commonly used quantity to compare the plausibility of two competing hypotheses, say $H_1$ and $H_2$, with respect to the observed data $x$ (Kass and Raftery, 1995). It is defined by

$$BF_{1:2}(x) = \frac{f(x | H_1)}{f(x | H_2)} = \frac{\int_{\Theta_1} f(x | \theta_1) f(\theta_1) \, d\theta_1}{\int_{\Theta_2} f(x | \theta_2) f(\theta_2) \, d\theta_2},$$

where $\theta_1 \in \Theta_1$ are the parameters for the model under $H_1$ and $\theta_2 \in \Theta_2$ are the parameters for the model under $H_2$, not necessarily with the same dimension. Note that if all parameters are fixed, the BF is the usual likelihood ratio (for nested models), while for models with random parameters it is the ratio of the marginal likelihoods. The BF can be interpreted in several ways, for example it represents the updating factor of the prior odds $Pr(H_1) / Pr(H_2)$ to the posterior odds $Pr(H_1 | x) / Pr(H_2 | x)$ after observing the data. Moreover, it is not necessary to regard one of the models as “true” and the other as “false”, as the BF can be interpreted as how likely the data were predicted by $H_1$ compared to $H_2$. Table 1 shows a widely used classification of Bayes factors (Kass and Raftery, 1995).

2.3 Reverse-Bayes analysis

The idea of reversing Bayes’ theorem was first proposed by Good (1950) but remained unexplored until Matthews (2001a) introduced the Analysis of Credibility, which in turn lead to new...
Table 1: Classification of Bayes factors proposed by Kass and Raftery (1995)

| log\(_{10}\) BF\(_{1:2}\) | BF\(_{1:2}\) | Evidence against \(H_2\) |
|----------------|----------|---------------------|
| 0 to 1/2       | 1 to 3.2 | not worth a bare mention |
| 1/2 to 1       | 3.2 to 10| substantial          |
| 1 to 2         | 10 to 100| strong               |
| > 2            | > 100    | decisive             |

developments in reverse-Bayes analysis (Matthews, 2001b; Greenland, 2006, 2011; Held, 2013; Colquhoun, 2017; Matthews, 2018; Held, 2019, 2020b). Common to these developments is that first a certain property of the posterior (or marginal likelihood) is fixed. Then Bayes’ theorem is used backwards to obtain the prior distribution which combined with the data leads to the fixed posterior or marginal likelihood.

In this paper, we consider a two-stage procedure that naturally suits to the replication setting: We first determine the sufficiently sceptical prior of the effect such that the original result is no longer convincing in terms of its Bayes factor. Using another Bayes factor, we then quantify replication success by comparing how likely the replication data are predicted by the sufficiently sceptical prior relative to an advocacy prior, which is the posterior of the effect conditional on the original data and an uninformative prior. The following two sections will explain this procedure in more detail.

2.4 Data from the original study

For the effect estimate \(\hat{\theta}_o | \theta \sim N(\theta, \sigma_o^2)\) from the original study consider BF\(_o\)(\(g\)), the BF comparing the point null hypothesis \(H_0: \theta = 0\) to the local alternative \(H_S: \theta \sim N(0, g \cdot \sigma_o^2)\) as a function of the relative sceptical prior variance \(g\) for fixed \(\hat{\theta}_o\). It follows from well known results (see e.g. Bernardo and Smith, 2000, Appendix A.2) that the marginal distribution of the original effect estimate under \(H_S\) is \(\hat{\theta}_o | H_S \sim N(0, \sigma_o^2 \cdot [1 + g])\), and thus the analytical form of BF\(_o\)(\(g\)) is

\[
BF_o(g) \equiv BF_{0S}(\hat{\theta}_o; g) = \sqrt{1 + g} \cdot \exp \left\{ -\frac{1}{2} \cdot \frac{g}{1 + g} \cdot z_o^2 \right\}. \tag{1}
\]

Figure 1: BF\(_o\)(\(g\)) as a function of relative sceptical prior variance \(g\) for different values of \(|z_o| = |\hat{\theta}_o|/\sigma_o\). Minima BF\(_o\) are indicated by dots (●). Dashed vertical lines indicate sufficiently sceptical relative prior variance \(g\), at level \(\gamma = 1/10\), if existent for corresponding \(|z_o|\).

The BF from equation (1) is shown in Figure 1 as a function of \(g\) and for different values
of \(|z_0|\). For fixed \(z_0\), it is well known that this BF is bounded from below by

\[
\text{BF}_0 = \begin{cases} 
|z_0| \cdot \exp(-z_0^2/2) \cdot \sqrt{e} & \text{for } |z_0| > 1 \\
1 & \text{for } |z_0| \leq 1
\end{cases}
\]

which is reached at \(g = \max(0, z_0^2 - 1)\) (Edwards et al., 1963). Further increasing the relative sceptical prior variance increases (1) indefinitely because of the Jeffreys-Lindley’s paradox, \(i.e.\) \(\text{BF}_0(g) \to \infty\) for \(g \to \infty\) (Bernardo and Smith, 2000, Section 6.1.4). Hence, for a relative sceptical prior variance \(g \in [0, g^\gamma]\), the resulting BF will be \(\text{BF}_0(g) \in [\text{BF}_0, 1]\).

Similar to the derivation of the sceptical \(p\)-value, we now apply reverse-Bayes analysis. To do so, we fix a level \(\gamma\) above which the original finding is no longer convincing to us. For example, \(\gamma\) could be 1/10; the level for strong evidence against \(H_0\) according to the classification from Kass and Raftery (1995). Suppose now there exists a \(g^\gamma \leq g\) such that \(g \in [0, g^\gamma]\) implies that \(\text{BF}_0(g) \in [\gamma, 1]\). The sufficiently sceptical prior is then given by \(\theta \sim N(0, g^\gamma \cdot \sigma^2_\theta)\) and it can be interpreted as the view of a sceptic who argues that given their prior belief about the effect \(\theta\), the observed \(\hat{\theta}_0\) cannot convince them about the presence of an effect at level \(\gamma\).

From Figure 1 we can see that the more “compelling” the original result (\(i.e.\) the larger \(|z_0|\)), the smaller the sufficiently sceptical relative prior variance \(g^\gamma\) needs to be in order to make the result no longer convincing at level \(\gamma\). If \(|z_0|\) is not sufficiently large, \(\text{BF}_0(g)\) will be always increasing in \(g\) (if \(|z_0| \leq 1\)) or \(\text{BF}_0(g)\) will reach a minimum above the chosen level \(\gamma\). In both cases the sufficiently sceptical relative prior variance \(g^\gamma\) is not defined since there is no need to challenge an already “unconvincing” result.

It can be shown (see Appendix A for details) that \(g^\gamma\) can be explicitly computed by

\[
g^\gamma = \begin{cases} 
- \frac{z_0^2}{q} - 1 & \text{if } - \frac{z_0^2}{q} \geq 1 \\
\text{undefined} & \text{else}
\end{cases}
\]

(2)

where \(W_{-1}(\cdot)\) is the branch of the Lambert \(W\) function (Corless et al., 1996) that satisfies \(W(y) \leq -1\) for \(y \in [-e^{-1}, 0]\). The Lambert \(W\) function is defined as the function satisfying \(W(y) \cdot \exp\{W(y)\} = y\) and is also known as “product logarithm” since it returns the number which plugged in the exponential function and then multiplied by itself produces \(y\). For real \(y\), \(W(y)\) is only defined for \(y \geq -e^{-1}\) and for \(y \in [-e^{-1}, 0]\) the function has two branches that are commonly denoted by \(W_0(\cdot)\), the branch with \(W(y) \geq -1\), and \(W_{-1}(\cdot)\), the branch with \(W(y) \leq -1\) (see Figure 2 for an illustration).

![Figure 2: Lambert W function for real argument y.](image-url)
that leads to BF \( \text{reflects the view of the sceptic} \) ability, e.g. \( \theta \) the posterior of as a function of the relative sceptical prior variance restrict the alternative such that only effects in the same direction as \( \hat{\theta} \) incorporate this, we consider BF \( \text{light of the new data, the sceptic is now challenged by an advocate of the original finding. To} \) function that \( x \) \( N(\theta, \sigma^2_\theta) \) \( H \) \( x \) \( \gamma \) \( \text{chosen level} \) \( \gamma \) \( \text{is in the set} \) \( B \) \( \text{the function that} \) \( \text{incorporates} \) \( \text{we consider} \) \( g \) \( \gamma \) \( \text{for} \) \( g \) \( \gamma \) \( \text{if the replication effect estimate goes in the other direction than} \) \( \theta \) \( \text{given the original estimate and a uniform prior (also the reference prior for} \) \( \gamma \) \( \gamma \) \( \text{for fixed} \) \( \theta \) \( \text{and} \) \( \gamma \) \( \text{to exist.} \) 2.5 Data from the replication study

In a second step, a replication study is conducted and a new effect estimate \( \hat{\theta}_r \) is obtained. In light of the new data, the sceptic is now challenged by an advocate of the original finding. To incorporate this, we consider BF(\( g \)), the BF for the effect estimate from the replication \( \hat{\theta}_r \mid \theta \sim N(\theta_0, \sigma^2_\theta) \) comparing the hypothesis \( H_S: \theta \sim N(0, g \cdot \sigma^2_\theta) \) to the alternative \( H_A: \theta \sim N(\theta_0, \sigma^2_\theta) \) as a function of the relative sceptical prior variance \( g \) for fixed \( \theta_0 \) and \( \theta_r \). \( H_S \) in this case reflects the view of the sceptic whereas the view of an advocate is represented by \( H_A \) since this is the posterior of \( \theta \) given the original estimate and a uniform prior (also the reference prior for this model).

This is not the only reasonable hypothesis that could represent the advocate; we could also restrict the alternative such that only effects in the same direction as \( \hat{\theta}_0 \) have non-zero probability, e.g. for positive \( \hat{\theta}_0 \) consider \( H'_{A}: \theta \sim N(\bar{\theta}_0, \sigma^2_\theta) \mathbb{1}_{[0,\infty)}(\theta) \), where \( \mathbb{1}_B(x) \) is the indicator function that \( x \) is in the set \( B \). If the replication effect estimate goes in the other direction than the original one, it is still possible to arrive at the conclusion that the data favour \( H_A \) over \( H_S \) (so-called “replication-paradox” (Ly et al., 2018)) with \( H_A \), whereas this impossible under \( H'_{A} \). However, assigning zero probability to effects that go in the other direction violates Cromwell’s rule and may hide the fact that the effect actually goes in the opposite direction which is not unlikely if the original estimate is small and estimated imprecisely. Whether or not \( H_A \) or \( H'_{A} \) should be chosen is a philosophical question and we will focus on \( H_A \) in this paper for technical simplicity.

Similarly as BF(\( g \)) from (1), BF(\( g \)) depends on \( z_0 \) and the relative sceptical prior variance \( g \), but also on \( z_r \) and the relative variance \( c = \sigma^2_\theta / \sigma^2_r \), i.e.

\[
\text{BF}(\hat{\theta}; g) \equiv \text{BF}_{S,A}(\hat{\theta}_r; g) = \sqrt{\frac{1 + c}{1 + cg}} \cdot \exp \left\{ -\frac{1}{2} \left( \frac{z_r^2}{1 + cg} - \frac{(z_r - z_0 \sqrt{c})^2}{1 + c} \right) \right\}. \tag{4}
\]

If the BF from (4) evaluated at the sufficiently sceptical relative prior variance \( g \gamma \) is not larger than the corresponding level \( \gamma \)

\[
\text{BF}(g \gamma) \leq \gamma,
\]

the replication data are at least \( 1 / \gamma \) times better predicted by the advocate than by the sceptic. Thus, the sceptic’s objection is rendered unrealistic and we declare replication success at level \( \gamma \).
For example, if we observe $z_o = 3$ (equivalent to a two-sided $p$-value of 0.003) and choose a level $\gamma = 1/10$, using equation (2), the sufficiently sceptical relative prior variance is $g_\gamma \approx 1.6$. If a replication is conducted with the same precision ($c = 1$) and we observe $z_r = 2.5$ (equivalent to a two-sided $p$-value of 0.012), using equation (4) this would lead to $\text{BF}_r(1.6) \approx 1/3.5$, which means that the replication was not successful at level $\gamma = 1/10$. However, if we had chosen a higher level, e.g., $\gamma = 1/3$, the replication would have been considered successful since then $g_\gamma \approx 0.4$ and $\text{BF}_r(0.4) \approx 1/7.4$.

### 2.6 The sceptical Bayes factor

Apart from specifying a level $\gamma$, this procedure offers an automated way to assess replication success. However, it is unclear which level $\gamma$ should be chosen for several reasons: There are well-established classifications of Bayes factors (see e.g. Held and Ott (2018) for an overview), yet they are all still arbitrary and it is unclear whether they can be applied to the reverse-Bayes approach considered here. Moreover, the sufficiently sceptical relative prior variance $g_\gamma$ does not exist for levels $\gamma$ that are below the lower bound $\text{BF}_o$. Finally, a quantitative definition of replication success would be preferred to an arbitrary dichotomisation.

It seems more sensible to eliminate this arbitrariness by finding the smallest level for which the replication is still considered to be successful, similar to minimum Bayes factors (Berger and Sellke, 1987; Held and Ott, 2018). We thus define the sceptical Bayes factor $\text{BF}_S$ as the smallest level where replication success can be declared. Since we want to avoid that this level is attained at an unreasonably large relative sceptical prior variance $g$ due to the Jeffreys-Lindley’s paradox, we further restrict the range of $g$ over which the minimisation takes place to be the interval between zero and $g = \max \{ z_o^2 - 1, 0 \}$. The sceptical Bayes factor is hence given by

$$\text{BF}_S \equiv \inf_{g_\gamma \in [0, g]} \{ \gamma : \text{BF}_r(g_\gamma) \leq \gamma \}. \quad (5)$$

Figure 3 shows $\text{BF}_r(\cdot)$ and $\text{BF}_o(\cdot)$ over a grid of $g$ values for several scenarios of $z_o$ and $z_r$ along with the corresponding $\text{BF}_S$. As can be seen in the upper three plots, there are situations

![Figure 3: BF_r(·) and BF_o(·) as a function of g. In all examples c = 1. Maxima for BF_r and minima for BF_o are indicated by dots (●), BF_S indicated by crosses (×) where defined.](image-url)
when either \( z_o, z_r \), or both are so small that replication success cannot be established for any \( g \in [0,g_\gamma] \) and hence \( BF_S \) does not exist. This means that the replication was unsuccessful since it is impossible for the advocate to convince the sceptic at any level of evidence. Moreover, by construction \( BF_S \) cannot be larger than one and smaller than \( BF_o \). This implies that if \( |z_o| \leq 1 \), replication success can only be achieved at \( \gamma = 1 \) (the middle and bottom plots on the left), while for \( |z_o| > 1 \) it can be achieved at most at the lower bound \( BF_o = |z_o| \cdot \exp(-z_o^2/2) \cdot \sqrt{e} \) (the middle and bottom plots in the centre). Finally, in a typical situation when \( z_o \) and \( z_r \) are both reasonably large, \( BF_S \) is located at the intersection \( BF_r(g_\gamma) = BF_o(g_\gamma) \) in \( g_\gamma \in [0,g] \) (the middle and bottom plots on the right).

There is no closed form expression for the computation of the sceptical Bayes factor in the general case (see Appendix B for details). However, it is worth noting that for the special case when the replication is conducted with the same precision as the original study, the \( g_* \), at which \( BF_r(\cdot) \) and \( BF_o(\cdot) \) intersect can be explicitly computed using again the Lambert W function:

\[
g_* = \begin{cases} 
\frac{-z_A^2}{W(-x)} - 1 & \text{if } \frac{-z_A^2}{W(-x)} \geq 1 \\
\text{undefined} & \text{else} 
\end{cases}
\]

where \( z_A^2 = (z_o^2 + z_r^2)/2 \) is the arithmetic mean of the squared z-statistics and \( d = \hat{\theta}_r/\hat{\theta}_o \) is the relative effect estimate. Since \( z_o^2, z_r^2 \geq 0 \) it must also hold that \( x \geq 0 \) and therefore \( W(-x) \) has two solutions whenever \( x < e^{-1} \) and at least one of \( z_o, z_r \) is non-zero. Moreover, we can see that the existence of the solutions (whether or not \( x < e^{-1} \)) depends on two things: First, the size of the squared z-statistics \( z_o^2 \) and \( z_r^2 \) which increase with increasing evidence for non-null effects in original and replication study, respectively. Second, on the squared relative effect estimate \( d^2 \) which is a measure for the compatibility of both effect estimates.

Assuming now that \( c = 1 \), that \( z_o \) and \( z_r \) are large enough such that the sceptical Bayes factor is located at the intersection of \( BF_o(\cdot) \) and \( BF_r(\cdot) \), and that the sufficiently sceptical relative prior variance \( g_\gamma \) is obtained as the solution from the \( W_0 \) branch in equation (6), \( BF_S \) is given by

\[
BF_S = BF_o(g_*) = \sqrt{-\frac{-z_A^2}{W(-x)} \cdot \exp \left\{ \frac{-z_o^2}{2} \right\}} \cdot \exp \left\{ -\frac{W_0(-x)}{d^2 + 1} \right\}.
\]

### 2.7 Connection to other measures of replication success

Of interest is also the relationship between \( BF_S \) and other measures of replication success. We note that the lower bound \( BF_o \), is simply a rescaling of the original \( p \)-value, \( p_o = 2 \cdot \{1 - \Phi(|z_o|)\} \), where \( \Phi(\cdot) \) is the cumulative distribution function of the standard normal distribution, and it will indicate more evidence against the null for decreasing \( p \)-values (\( BF_o \downarrow 0 \) as \( p_o \downarrow 0 \)). Thus to achieve replication success at a compelling level, the original result needs to be convincing on its own as it also determines the lower bound of the sceptical Bayes factor (since \( BF_S \in [BF_o,1] \)). Similarly, for compatible estimates (\( \hat{\theta}_o \approx \hat{\theta}_r \)) and fixed \( g \) and \( c \), \( BF_{SA}(\hat{\theta}_r) \) as given by (4) will decrease as the replication \( p \)-value \( p_r = 2 \cdot \{1 - \Phi(|z_r|)\} \) decreases. For fixed \( \hat{\theta}_r \) and with increasing incompatibility of the estimates, \( BF_{SA}(\hat{\theta}_r) \) will increase and provide evidence for the sceptic, i.e., \( BF_{SA}(\hat{\theta}_r) \to \infty \) for \( Q \to \infty \), with \( Q = (\hat{\theta}_o - \hat{\theta}_r)^2 / (\sigma_o^2 + \sigma_r^2) \). Hence, for \( BF_{SA}(\hat{\theta}_r; g_\gamma) \) to be smaller than the corresponding level \( \gamma \), the replication result needs not only to be convincing on its own, but also compatible with the original result such that replication success can be established at a compelling level. In this way, the proposed method elegantly combines traditional notions of replication success; it ensures that both studies need to show evidence against the null, similar to the “vote-counting” approach that requires both
original and replication study to be statistically significant, while at the same time penalising incompatibility of their effect estimates, similar to assessing compatibility with a Q-test or prediction intervals.

Verhagen and Wagenmakers (2014) proposed a related quantity to assess replication success, the “replication Bayes factor” BF$_R$. It is defined as the BF comparing the point null hypothesis $H_0: \theta = 0$, to the alternative that the effect is distributed according to the posterior distribution of $\theta$ after observing the original data. For the normal model considered in this paper and if a uniform initial prior was chosen, this alternative is also the hypothesis of the advocate considered earlier, i.e. $H_A: \theta \sim N(\theta_0, \sigma_0^2)$, and therefore BF$_R$ is given by

$$BF_R \equiv BF_{0A}(\hat{\theta}_r) = \sqrt{1 + c} \cdot \exp \left\{ -\frac{1}{2} \left( z_r^2 - \frac{(z_r - z_0\sqrt{c})^2}{1 + c} \right) \right\}.$$

It is clear that $BF_r(g) \rightarrow BF_R$ as $g \downarrow 0$ and $BF_r(g)$ can thus be considered an extension of the replication Bayes factor where the likelihood under the point null hypothesis is replaced by the marginal likelihood under the sufficiently sceptical prior. Furthermore, while BF$_R$ shares the properties with BF$_S$ that it decreases with decreasing replication $p$-value $p_r$ and that it increases with increasing conflict of the estimates $Q$, it does not depend directly on how much evidence the original study provides against the null. Thus if the effect estimates are compatible, BF$_R$ can also indicate a large degree of replication success for unconvincing original studies, whereas this is impossible for BF$_S$ which requires both studies to be convincing on their own. Additionally, if the replication result was so unconvincing such that BF$_S$ does not exist, BF$_R$ can also quantify to which extent the replication data favour the null.

Of particular interest is the relationship between the sceptical Bayes factor and the sceptical $p$-value $p_S$ (Held, 2020b), as $p_S$ is the result of a similar reverse-Bayes procedure. One also considers a sceptical prior for the effect $\theta \sim N(0, \tau^2)$, the sufficiently sceptical prior variance at level $\alpha$ is then defined as $\tau^2 = \tau^2_{\alpha}$ such that the $(1 - \alpha)$ credible interval for $\theta$ based on the posterior $\theta | \hat{\theta}_o, \tau^2_{\alpha}$ does not include zero. Replication success is declared if the tail probability of the replication effect estimate under its prior predictive distribution $\hat{\theta}_r | \tau^2_{\alpha} \sim N(0, \tau^2_{\alpha} + \sigma^2_r)$ is smaller than $\alpha$. The smallest level $\alpha$ where replication success can be established defines $p_S$. In contrast to BF$_S$, $p_S$ always exists and can be easily computed using only the variance ratio $c$, the arithmetic and the harmonic mean of the squared $z$-statistics. Similar to BF$_S$, the property that $p_S \geq \max\{p_o, p_r\}$ (Held, 2020b, Section 3.1) ensures that both studies need to be convincing on their own. However, while $p_S$ measures compatibility between the estimates only indirectly through conflict with the sceptical prior, BF$_S$ directly evaluates how likely the replication data are to occur under the posterior of the effect conditional on the original data.

Figure 4 shows BF$_S$, BF$_R$, and $p_S$ in a raster plot for a grid of $z_0$ and $z_r$ and with $c = 1$. In regions where $z_0$ and $z_r$ are similar, BF$_R$ has lower values compared to BF$_S$ (e.g. if $z_0 = 3$ and $z_r = 2.5$, BF$_R = 1/15$ and BF$_S = 1/5.2$). They only coincide when BF$_R = 1$ (e.g. if $z_0 = 3$ and $z_r = 1.4$), which is (for any $c$) at the hyperbola fulfilling

$$z_r = \frac{-z_0}{\sqrt{c}} \pm \sqrt{(1 + c)\{\log(1 + c) + z_0^2\}},$$

indicated by the dashed green lines in the left and middle panels. In regions where $z_0$ and $z_r$ differ by a large amount, BF$_R$ has values greater than one, indicating that the replication data favour the null hypothesis (e.g. if $z_0 = 2$ and $z_r = -2$, BF$_R = 10$), whereas BF$_S$ does not exist. For original findings which were only suggestive (i.e. $|z_0|$ is small), it is impossible to achieve replication success at a reasonable level with BF$_S$ and $p_S$, even though the result from the replication would be very convincing on its own. In contrast, BF$_R$ can indicate a large degree of replication success, provided that the replication result is compatible with the original result (e.g. BF$_R = 1/222$ for $z_0 = 1$ and $z_r = 5$, whereas BF$_S = 1$, $p_S = 0.17$). Note, that the replication-paradox occurs for BF$_R$ for some of the values shown in Figure 4, e.g. if $z_0 = -1$ and $z_r = 4.5$, BF$_R = 1/9.2$ provides strong evidence against the null, although the signs of the estimates differ, whereas BF$_S = 1$ indicates no replication success. However, in some very
extreme scenarios it is also possible that the replication paradox occurs for $BF_S$, e.g. if $z_0 = 10$ and $z_r = -30$ then $BF_S = 1/137$. It seems that $BF_S$ is more robust to the replication-paradox than $BF_R$ since the replication data are evaluated under the composite null hypothesis $H_S$ rather than the precise null hypothesis $H_0$. Nevertheless, the replication-paradox can only be completely avoided when restrictions on the parameter space are imposed such as discussed in Section 2.5. We can further see that when $c = 1$, $p_S$ treats $z_r$ and $z_0$ as exchangeable, whereas this is not the case for $BF_R$ and $BF_S$. For instance, $z_0 = 2$ and $z_r = 3$ or $z_0 = 3$ and $z_r = 2$ both lead to $p_S = 0.048$, while $BF_S = 1/2.2$ and $BF_R = 1/50$ for the former and $BF_S = 1/2.8$ and $BF_R = 1/4.1$ for the latter. Finally, we can see that in scenarios where the replication effect estimate goes in the same direction but with smaller magnitude than the original estimate, $BF_S$ reacts more sensitively to incompatibility than $p_S$, e.g. if $z_0 = 4.5$ and $z_r = 2$, which for $c = 1$ is equivalent to a relative effect estimate of $d = \frac{\hat{\theta}_r}{\hat{\theta}_0} = 0.44$, we obtain $BF_S = 1/1.1$, whereas $p_S = 0.034$.

### 2.8 Distribution under the null

The distribution of $BF_S$ can be easily approximated with stochastic simulation. Of particular interest is the distribution under the null hypothesis when the underlying effect is zero. Figure 5 shows the percentage of simulations for which $BF_S$ was smaller than one, equal to one, or undefined (A) and density estimates of the null distribution of $BF_S$ (B), $p_S$ (C), and $BF_R$ (D). The truncation of the distribution of $BF_S$ was chosen because it allows for easier comparison of the distributions below one for different values of $c$. We see that with increasing precision of the replication study (i.e. larger $c$), the mass of the distributions of $BF_S$ and $p_S$ shifts further towards one and a half, respectively, while the distribution of $BF_R$ shifts further towards values above one. Similarly, the proportion of $BF_S$ smaller than one decreases, e.g. it is 8% for $c = 1/4$, 6% for $c = 1$, and only 2% for $c = 4$. The proportion of $BF_S$ equal to one also decreases, whereas the proportion of undefined $BF_S$ increases with increasing $c$. Thus for all three methods the risk of wrongly claiming replication success decreases when the precision of the replication effect estimate is increased.

Simulation under the null can also be used to calibrate $BF_S$, similar to the approach suggested in Vlachos and Gelfand (2003). For a given variance ratio $c$, we can determine the threshold for replication success $\gamma_S$ such that $\alpha = \Pr(BF_S < \gamma_S \mid H_0)$, the probability for in-
correctly claiming replication success (type I error) is low. For example, if we set $c = 1$ and

want to control the type I error at $\alpha = 0.05 \times 0.025 = 0.00125$ (the conventional level for type I error control of two independent experiments with two-sided testing in the first and one-sided testing in the second), the threshold turns out to be $\gamma_S = 1/2.4$. Type I error calibration of $BF_S$ as well as $BF_R$ and $p_S$ is shown in Figure 6. Note, that for $c = 1$ the null distribution of $p_S$ is known (Held, 2020a) and simulation would not be needed for type I error calibration of this method.
2.9 Distribution under the alternative

A further operating characteristic of BF_S that can be studied is the probability that replication success is established given that the alternative is true (power) using the type I error calibrated threshold. We consider two types here: the power is either computed assuming that the underlying true effect corresponds to its estimate from the original study (conditional power) or the power is computed under the posterior predictive distribution of the replication effect estimate (predictive power) (Spiegelhalter et al., 1986; Weiss, 1997; Micheloud and Held, 2020).

Figure 7 shows conditional and predictive power of BF_S, BF_R, and p_S as a function of the variance ratio c and for several values of z_o. In general, uncertainty about replication success is higher for predictive power, leading it to be closer to 50% in all cases. As can also be seen, if the original result was not convincing on its own (e.g., if z_o = 1.75), it is impossible to achieve replication success with the reverse-Bayes methods BF_S and p_S, leading to low or zero power (conditional and predictive). This is not the case for BF_R, for which high power can also be obtained for small z_o if c is sufficiently large. For larger z_o and small to medium c, on the other hand, BF_S and p_S show power gains compared to BF_R. For example for z_o = 2.75 and c < 3 the power of BF_S and p_S is substantially larger compared to BF_R, while it is similar for c ≥ 3. In conclusion, the reverse-Bayes methods seem to be advantageous compared to BF_R if the original finding was already convincing on its own and the estimation precision in the replication is not much higher than in the original study. For higher precision, the power differences between the methods disappear.
3 Applications

3.1 Reproducibility Project: Cancer Biology

We first examine two examples from the Reproducibility Project: Cancer Biology (Errington et al., 2014). The first example concerns the replication of Ward et al. (2010), where one of the findings was that samples from patients with acute myeloid leukaemia showed an increased level of the oncometabolite 2HG/glutamate when they showed a certain mutation (IDH1/IDH2) compared to samples without this mutation (standardised mean difference SMD = 4.36 with 95% confidence interval [2.62, 6.09]). The replication study (Showalter et al., 2017) estimated the effect of the mutation to be slightly smaller (standardised mean difference SMD = 3.09 with 95% confidence interval [1.35, 4.80]), however, the meta-analytic Q-test does not indicate incompatibility of the effect estimates ($p_Q = 0.31$).

Plots A and B in Figure 8 show the results of our analysis of this study. Note that the standard errors of the SMD effect estimates were computed by dividing the difference between the confidence interval limits by twice the 97.5% quantile of the standard normal distribution. We obtain a replication Bayes factor of $BF_R = 1/200$, a one-sided sceptical $p$-value of $p_S = 0.002$, and a sceptical Bayes factor of $BF_S = 1/35$. Thus all methods indicate that the original finding was replicated to a high degree in this case.

The second example is the replication of Delmore et al. (2011). One of the investigations in this study was a comparison of the survival of mice which were treated with the drug JQ1 versus mice in a control group that received no treatment. The original study (Delmore et al., 2011) estimated a hazard ratio of $HR = 25.93$ with 95% confidence interval from 5.48 to 122.58, whereas the replication (Aird et al., 2017) resulted in an estimated hazard ratio of $HR = 3.75$ with 95% confidence interval from 1.19 to 11.81. Although both effect estimates are significantly different from zero (at the conventional 5% level), their sizes differ by an order of magnitude and the meta-analytic Q-test provides some evidence for incompatibility ($p_Q = 0.05$).

Plots C and D in Figure 8 show the results of our analysis of the Aird et al. (2017) data. The HRs were first log-transformed, resulting in approximate normality. The standard errors of the logHRs were computed by dividing the difference between the log-transformed confidence interval limits by twice the 97.5% quantile of the standard normal distribution. The replication Bayes factor then turns out to be $BF_R = 1/1.1$, indicating that the replication data are almost equally likely predicted by either the null hypothesis or the posterior of the effect from the original study. Similarly, the sceptical Bayes factor is $BF_S = 1/1.1$ which suggests that the replication data are not compelling enough that replication success can be established at a reasonable level. The one-sided sceptical $p$-value, on the other hand, is $p_S = 0.031$, indicating some conflict between the sceptical prior and the observed data from the replication. This highlights an important difference between the two approaches: Bayes factors allow for direct quantification of evidence for one hypothesis versus another, whereas $p$-values can do so only indirectly. Thus, even though the replication data in Aird et al. (2017) are unlikely under the sceptical prior, they are also very unlikely under the advocacy prior, and therefore a reverse-Bayes analysis using Bayes factors leads to a different conclusion than an analysis based on $p$-values.

3.2 Social Sciences Replication Project

Next, we consider the data from the Social Sciences Replication Project (Camerer et al., 2018), provided in Table 2. The effect estimates were given as correlations ($r$) in all cases. Application of the Fisher z-transformation, $\hat{\theta} = \tanh^{-1}(r)$, leads to approximate normality with the variance of the transformed estimates being a simple function of the sample size $n$, i.e. $\text{Var}(\hat{\theta}) = 1/(n - 3)$ (Fisher, 1921). The replication Bayes factors under normality agree reasonably well with the ones reported in the supplement of Camerer et al. (2018) where the Bayesian analysis is described (see https://osf.io/nsxgj/), despite that one-sided alternatives and
The sceptical Bayes factor

Samuel Pawel, Leonhard Held

\[
BF_{R} = 1/200
\]

Figures 8: Examples taken from Showalter et al. (2017) (above) and Aird et al. (2017) (below). Plots A and C show a standard Bayesian analysis of the data from the replication study with a prior corresponding to the posterior from the original study and 95% credible interval indicated at the top of the plots. Plots B and D show a reverse-Bayes analysis of both studies, BF_S is indicated by a cross (×).

Table 2: Results for data from Social Sciences Replication Project (Camerer et al., 2018). Confidence intervals for individual effects were computed on Fisher z-scale and back-transformed to correlation scale (r). The p-values \(p_o\), \(p_r\), \(p_S\) are one-sided (in the direction of the original estimate), and \(p_Q\) is the p-value from the Q-test assessing the differences between the effect estimates.

| Original study                 | \(r_o\) [95% CI] | \(r_r\) [95% CI] | \(p_o\) | \(p_r\) | \(p_S\) | \(p_Q\) | \(BF_S\) | BF_F | BF_R |
|------------------------------|-----------------|-----------------|--------|--------|--------|--------|--------|------|------|
| Hauser et al. (2014)          | 0.82 [0.68, 0.90] | 0.83 [0.63, 0.93] | < 0.0001  | < 0.0001 | 0.86 | < 0.0001 | < 1/1000 | < 1/1000 |
| Avison et al. (2012)          | 0.96 [0.88, 0.99] | 0.83 [0.53, 0.94] | < 0.0001  | < 0.0001 | 0.062 | 0.0003 | 1/78 | 1/284 |
| Wilson et al. (2014)          | 0.67 [0.41, 0.83] | 0.59 [0.34, 0.77] | < 0.0001  | < 0.0001 | 0.60 | 0.002 | 1/45 | < 1/1000 |
| Derex et al. (2013)           | 0.52 [0.29, 0.70] | 0.36 [0.13, 0.56] | < 0.0001  | 0.01 | 0.29 | 0.01 | 1/8.5 | 1/31 |
| Gneezy et al. (2014)          | 0.22 [0.08, 0.36] | 0.18 [0.09, 0.27] | 0.001 | 0.0001 | 0.64 | 0.19 | 1/6.9 | 1/474 |
| Karpicke and Blunt (2011)     | 0.60 [0.36, 0.77] | 0.38 [0.12, 0.60] | < 0.0001  | 0.03 | 0.19 | 0.01 | 1/3.6 | 1/12 |
| Morewedge et al. (2010)       | 0.45 [0.12, 0.69] | 0.35 [0.16, 0.52] | 0.004 | 0.0003 | 0.58 | 0.036 | 1/3.9 | 1/160 |
| Kovacs et al. (2010)          | 0.45 [0.06, 0.72] | 0.59 [0.44, 0.70] | 0.13 | < 0.0001 | 0.44 | 0.03 | 1/3.2 | < 1/1000 |
| Duncan et al. (2012)          | 0.67 [0.25, 0.88] | 0.44 [0.25, 0.59] | 0.002 | < 0.0001 | 0.26 | 0.036 | 1/3.1 | < 1/1000 |
| Nishi et al. (2015)           | 0.20 [0.06, 0.33] | 0.12 [0.03, 0.20] | 0.002 | 0.005 | 0.31 | 0.046 | 1/2.5 | 1/8.2 |
| Janssen et al. (2010)         | 0.43 [0.45, 0.76] | 0.34 [0.04, 0.59] | < 0.0001  | 0.013 | 0.061 | 0.017 | 1/16 | 1/16 |
| Balafoutas and Sutter (2012)  | 0.28 [0.05, 0.48] | 0.15 [0.02, 0.27] | 0.009 | 0.011 | 0.31 | 0.085 | 1/1.6 | 1/3.9 |
| Pyc and Rawson (2010)         | 0.38 [0.05, 0.63] | 0.15 [0.04, 0.26] | 0.011 | 0.004 | 0.18 | 0.11 | 1/1.2 | 1/4 |
| Rand et al. (2012)            | 0.14 [0.04, 0.24] | 0.03 [-0.02, 0.07] | 0.004 | 0.12 | 0.047 | 0.19 | 9.6 |
| Ackerman et al. (2010)        | 0.27 [0.00, 0.50] | 0.06 [-0.02, 0.14] | 0.024 | 0.063 | 0.14 | 0.21 | 3.2 |
| Sparrow et al. (2011)         | 0.37 [0.14, 0.56] | 0.05 [-0.08, 0.18] | 0.0009 | 0.23 | 0.016 | 0.24 | 29 |
| Shah et al. (2012)            | 0.27 [0.00, 0.50] | -0.02 [-0.09, 0.06] | 0.023 | 0.65 | 0.043 | 0.63 | 25 |
| Kidd and Castano (2013)       | 0.27 [0.06, 0.46] | -0.03 [-0.10, 0.05] | 0.006 | 0.77 | 0.009 | 0.72 | 72 |
| Gervais and Norenzayan (2012) | 0.29 [0.03, 0.51] | -0.04 [-0.12, 0.05] | 0.014 | 0.79 | 0.02 | 0.73 | 36 |
| Lee and Schwarz (2010)        | 0.39 [0.09, 0.62] | -0.05 [-0.16, 0.07] | 0.006 | 0.78 | 0.009 | 0.74 | 65 |
| Ramirez and Beslock (2011)    | 0.79 [0.54, 0.91] | -0.10 [-0.31, 0.13] | < 0.0001 | 0.80 | < 0.0001 | 0.79 | > 1000 |
initial Cauchy priors were used, as well as computations were performed with summaries of the raw data rather than effect estimates transformed to correlations. There are two notable exceptions: The study by Janssen et al. (2010), where the reported analysis indicates overwhelming evidence for the null compared to the advocate (BF_R > 1000), while our computed BF_R provides inconclusive evidence (BF_R = 1/1.6), and the study by Dereix et al. (2013) where the reported analysis indicates overwhelming evidence for the advocate compared to the null (BF_R < 1/1000), while our computed BF_R provides only strong evidence (BF_R = 1/31). In both cases non-standard analyses were carried out, *i.e.* a Bayesian Mann-Whitney test for the former and an encompassing prior approach for the latter. We therefore advise to interpret our results in these cases as numerical examples rather than as definitive answers about the replicability of these two studies.

For the study pairs where BF_S suggests strong replication success, p_S and BF_R suggest the same in every case. However, there are also study pairs where there appears to be disagreement among the methods. Discrepancies between p_S and BF_S usually happen in situations where the replication shows an effect estimate that, although incompatible with the sceptic, is also incompatible with the advocate. For example in the Janssen et al. (2010) replication, both effect estimates are substantially larger than zero (r_o = 0.63 with p_o < 0.0001 and r_r = 0.34 with p_r = 0.013), yet the Q-test provides some evidence for their incompatibility (p_Q = 0.061), which explains why p_S = 0.017, but BF_S = 1/1.6 only.

Discrepancies between BF_R and BF_S can arise when the replication finding provides overwhelming evidence against the null, whereas the original finding was less compelling. An example which illustrates this situation is the replication of Kovacs et al. (2010). The original study provided only moderate evidence against the null (p_o = 0.013 or equivalently BF_o = 1/3.2), whereas the replication finding was more compelling (p_r < 0.0001). By construction BF_S can only be as small as BF_o which happens in this case (BF_S = 1/3.2). This means that the original finding was substantiated at the highest level of evidence possible. The BF_R, on the other hand, is not limited by the moderate level of evidence from the original study and indicates decisive evidence for the advocate (BF_R < 1/1000) as there is no evidence for conflict of the effect estimates (p_Q = 0.44). This illustrates that in order to achieve a reasonable degree of replication success, BF_S requires both studies to be convincing on their own, whereas BF_R only requires a compelling replication result provided that both estimates are compatible.

### 4 Discussion

We presented a novel reverse-Bayes method for the statistical analysis of replication studies. The method consists of a two-step procedure that naturally fits the replication setting: First, we determine a prior distribution for the unknown effect such that the original result is no longer convincing in terms of its Bayes factor. This prior represents the view of a sceptic who does not believe in the original claim. In the second step, the sceptic is challenged by an advocate of the original finding. This is operationalised by a Bayes factor for the replication effect estimate that contrasts the sceptical prior to the posterior of the effect based on the original result when a uniform initial prior was chosen. If this second Bayes factor indicates that the data favour the advocate over the sceptic at a higher level than the sceptic’s objection from step one, the advocate managed to convince the sceptic about the credibility of the original claim and we declare replication success. To remove the dependence on choosing the Bayes factor level at which the original finding is no longer convincing, we determine the highest level where replication success can still be declared and denote this the sceptical Bayes factor.

The sceptical Bayes factor is thus an extension of the sceptical p-value (Held, 2020b) to the Bayesian hypothesis testing framework. In contrast to the sceptical p-value, closed form expressions are only available when certain conditions are met (when original and replication effect estimates are sufficiently large and have the same variance), otherwise numerical root-finding algorithms must be used. While the sceptical p-value always exists, BF_S does
not. This happens when the result of the replication is so inconclusive that a sceptic of the original finding cannot be convinced at any level of evidence. While the sceptical \( p \)-value only quantifies the conflict between the sceptical prior and the observed replication data, the BF\(_S\) also takes into account how likely the replication data are under the advocacy prior, which is the posterior of the effect conditional on the original finding and a uniform prior. Hence, in cases where the replication estimate goes in the same direction but with much smaller magnitude, the sceptical \( p \)-value might still indicate some replication success, while BF\(_S\) would indicate less replication success since the data are also unlikely under the advocacy prior (see the example from Aird et al. (2017) in Section 3.2). In the (rare) scenario where there is incompatibility because the replication estimate goes in the same direction but with much larger magnitude, \( p_S \) will indicate a higher degree of replication success than if there was no conflict, whereas BF\(_S\) will not. The direct evaluation of the compatibility therefore allows for stronger statements about the degree of replication success compared to \( p_S \).

We compared BF\(_S\) to the replication Bayes factor BF\(_R\) (Verhagen and Wagenmakers, 2014), which quantifies how likely the replication data are under a point null hypothesis compared to the posterior of the effect based on the original study. Similar to BF\(_S\), the BF\(_R\) takes into account whether the effect estimates from original and replication study are compatible as well as whether the replication finding shows evidence against the null. In contrast to BF\(_S\), BF\(_R\) does not require the original finding to be compelling on its own. The BF\(_R\) can be considered standard forward-Bayes evidence updating tailored to the replication setting. The nature of BF\(_S\), on the other hand, lies not in evidence synthesis but rather in challenging and substantiation of an original finding through a reverse-Bayes argument.

The presented approach also comes with some limitations and could be extended in many ways. First, BF\(_S\) cannot be interpreted as a standard Bayes factor. We have presented a way to calibrate BF\(_S\) through type I error control, yet future work is required to explore other calibration approaches that allow for a more quantitative interpretation. Second, the set of effect estimates for which replication success is achieved at a specified level could be determined, similar to the concept of the “support interval” (Wagenmakers et al., 2020). Such a set could give a better idea about which effect estimates would have been able to substantiate the original claim at a specified level. Third, in many cases we might be interested in the direction of the effect and the replication setting is no exception since usually the goal is to replicate an effect in the same direction as in the original study. As already discussed, our approach could also be modified to work in this case by imposing restrictions on the parameters under the hypothesis of the advocate. Fourth, in many cases not just one but several replication studies are conducted for one original study (e.g., as in Klein et al., 2014). The Bayesian framework allows us to easily extend BF\(_S\) to the “many-to-one” replication setting as the marginal likelihoods in BF\(_R\) are also straightforward to compute for a sample of replication effect estimates. Finally, a multivariate generalisation would allow for effects in the form of random vectors with approximate multivariate normal likelihood which is then combined with a sceptical \( g \)-prior (Liang et al., 2008). The normal prior could also be replaced with other distributions, for example the (multivariate) Cauchy distribution which is often the preferred prior choice for default Bayes factor hypothesis tests (Jeffreys, 1961). The \( g \) parameter of the \( g \)-prior or the scale parameter of the Cauchy prior would then take over the role of the relative sceptical prior variance.

An important aspect that has not been discussed so far is the design of new replication studies. An appropriate sample size is of particular importance for a replication to be meaningful. For sample size determination related to Bayesian hypothesis testing and the normal-normal model, there exist theoretical results that could be used for our purpose (see Weiss, 1997; De Santis, 2004). A more general strategy is to use stochastic simulation (Schönbrodt and Wagenmakers, 2017). Typically a design prior (O’Hagan et al., 2001) is used rather than the prior used in the final analysis. The design prior should reflect the researchers’ best knowledge of the effect under investigation to facilitate efficient design. Then the sample size is determined such that the prior predictive distribution of the replication effect estimate results...
in a BFₜ below a specified cut-off with desired probability (e.g. 80%). The “relative cut-off” approach of Weiss (1997) could also be used. Instead of specifying a cut-off value, the desired type I error is specified which then determines the cut-off and the sample size such that the type I error is controlled under the null and the power is attained under the alternative. Future work is required to investigate in further detail the issue of sample size planning based on BFₜ.

Hypothesis testing is, however, only one side of the coin and we should not forget parameter estimation. Especially in the Bayesian framework, estimation-based approaches for the evaluation of replication studies are still underdeveloped, future research should also focus on these aspects. For a thorough assessment of replication attempts, no single metric seems to be able to answer all important questions completely. Instead, we recommend that researchers conduct a comprehensive statistical evaluation of replication success. We advocate the reverse-Bayes approach as a key part of such analyses, as it provides a conceptually sound way to challenge and substantiate an original finding, which combines traditional notions of replication success.

Software and data

All analyses were performed in the R programming language version 4.0.2 (R Core Team, 2020). The code to reproduce this manuscript is available at https://gitlab.uzh.ch/samuel.pawel/BFScode. We used the implementation of the Lambert W function from the package lamW (Adler, 2015), graphics were created with the ggplot2 package (Wickham, 2016), the sceptical p-value and related power calculations were computed using the package ReplicationSuccess downloaded from https://r-forge.r-project.org/projects/replication/ (Held, 2020b). All methods are implemented in the R package BayesRep of which the development version is available at https://gitlab.uzh.ch/samuel.pawel/BayesRep. We plan to make the package available on CRAN in the future.

Data from the Social Sciences Replication Project (Camerer et al., 2018) were downloaded from https://osf.io/abu7k/. Data for the two examples from the Reproducibility Project: Cancer Biology (Errington et al., 2014) were extracted from the forest-plots in Figure 4 and Figure 4B in Showalter et al. (2017) and Aird et al. (2017), respectively.

Acknowledgements

We thank Guido Consonni, Malgorzata Roos, and Maria Bekker-Nielsen Dunbar for helpful discussion and comments on drafts of the manuscript.

Funding

This work was supported by the Swiss National Science Foundation (project number 189295, http://p3.snf.ch/Project-189295). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Adler, A. (2015). lamW: Lambert-W Function. URL https://CRAN.R-project.org/package=lamW. R package version 1.3.3.

Aird, F., Kandela, I., and Mantis, C. (2017). Replication study: BET bromodomain inhibition as a therapeutic strategy to target c-myc. eLife, 6. doi:10.7554/elife.21253.
Bayarri, M. and Mayoral, A. (1999). Graphical models for hierarchical computations in the analysis and design of replications. *Revista de la Real Academia de las Ciencias Exactas, Fisicas y Naturales*, 93(3):281–293.

Bayarri, M. and Mayoral, A. (2002a). Bayesian analysis and design for comparison of effect-sizes. *Journal of Statistical Planning and Inference*, 103(1-2):225–243. doi:10.1016/s0378-3758(01)00223-3.

Bayarri, M. J. and Mayoral, A. M. (2002b). Bayesian design of “successful” replications. *The American Statistician*, 56:207–214. doi:10.1198/0003130025.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of $P$ values and evidence. *Journal of the American Statistical Association*, 82(397):112. doi:10.2307/2289131.

Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory*. John Wiley & Sons, Inc. doi:10.1002/9780470316870.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altméj, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351:1433–1436. doi:10.1126/science.aaf0918.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B., Pfeiffer, T., Altméj, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E., and Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behavior*, 2:637–644. doi:10.1038/s41562-018-0399-z.

Colquhoun, D. (2017). The reproducibility of research and the misinterpretation of $p$-values. *Royal Society Open Science*, 4(12). doi:10.1098/rsos.171085.

Consonni, G. (2019). Sufficiently skeptical intrinsic priors for the analysis of replication studies. Unpublished notes.

Corless, R. M., Gonnet, G. H., Hare, D. E. G., Jeffrey, D. J., and Knuth, D. E. (1996). On the Lambert W function. *Advances in Computational Mathematics*, 5(1):329–359. doi:10.1007/bf02124750.

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., Berniinas, R., Boudesseau, J., Colombo, M., Cushman, F., Diaz, R., van Dongen, N. N. N., Dranseika, V., Earp, B. D., Torres, A. G., Hannikainen, I., Hernández-Conde, J. V., Hu, W., Jaquet, F., Khalifa, K., Kim, H., Kneer, M., Knobe, J., Kurthy, M., Lantian, A., yi Liao, S., Machery, E., Moerenhout, T., Mott, C., Phelan, M., Phillips, J., Rambharose, N., Reuter, K., Romero, F., Sousa, P., Sprenger, J., Thalabard, E., Tobia, K., Viciana, H., Wilkenfeld, D., and Zhou, X. (2018). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*. doi:10.1007/s13164-018-0400-9.

De Santis, F. (2004). Statistical evidence and sample size determination for Bayesian hypothesis testing. *Journal of Statistical Planning and Inference*, 124(1):121–144. doi:10.1016/s0378-3758(03)00198-8.

Delmore, J. E., Issa, G. C., Lemiesz, M. E., Rahl, P. B., Shi, J., Jacobs, H. M., Kasratis, E., Gilpatrick, T., Paranal, R. M., Qi, J., Chesi, M., Schinzel, A. C., McKeown, M. R., Heffernan, T. P., Vakoc, C. R., Bergsage, P. L., Ghobrial, I. M., Richardson, P. G., Young, R. A., Hahn, W. C., Anderson, K. C., Kung, A. L., Bradner, J. E., and Mitsiades, C. S. (2011). BET bromodomain inhibition as a therapeutic strategy to target c-myc. *Cell*, 146(6):904–917. doi:10.1016/j.cell.2011.08.017.
Derex, M., Beugin, M.-P., Godelle, B., and Raymond, M. (2013). Experimental evidence for the influence of group size on cultural complexity. *Nature*, 503(7476):389–391. doi:10.1038/nature12774.

Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70(3):193–242. doi:10.1037/h0044139.

Errington, T. M., Iorns, E., Gunn, W., Tan, F. E., Lomax, J., and Nosek, B. A. (2014). An open investigation of the reproducibility of cancer biology research. *eLife*, 3. doi:10.7554/elife.04333.

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32.

Good, I. J. (1950). *Probability and the Weighting of Evidence*. Griffin, London, UK.

Goodman, S. N. (1992). A comment on replication, *p*-values and evidence. *Statistics in Medicine*, 11(7):875–879. doi:10.1002/sim.4780110705.

Greenland, S. (2006). Bayesian perspectives for epidemiological research: I. foundations and basic methods. *International Journal of Epidemiology*, 35:765–775. doi:10.1093/ije/dyi312.

Greenland, S. (2011). Null misinterpretation in statistical testing and its impact on health risk assessment. *Preventive Medicine*, 53:225–228. doi:10.1016/j.ypmed.2011.08.010.

Harms, C. (2019). A Bayes factor for replications of ANOVA results. *The American Statistician*, 73(4):327–339. doi:10.1080/00031305.2018.1518787.

Hedges, L. V. and Schauer, J. M. (2019). More than one replication study is needed for unambiguous tests of replication. *Journal of Educational and Behavioral Statistics*, 44(5):543–570. doi:10.3102/1076998619852953.

Held, L. (2013). Reverse-Bayes analysis of two common misinterpretations of significance test. *Clinical Trials*, 10:236–242. doi:10.1177/1740774512467707.

Held, L. (2019). The assessment of intrinsic credibility and a new argument for *p* < 0.005. *Royal Society Open Science*, 6(3):181534. doi:10.1098/rsos.181534.

Held, L. (2020a). The harmonic mean *χ*²-test to substantiate scientific findings. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(3):697–708. doi:10.1111/rssc.12410.

Held, L. (2020b). A new standard for the analysis and design of replication studies (with discussion). *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(2):431–448. doi:10.1111/rssa.12493.

Held, L. and Ott, M. (2018). On *p*-values and Bayes factors. *Annual Review of Statistics and Its Application*, 5(1):393–419. doi:10.1146/annurev-statistics-031017-100307.

Janssen, M. A., Holahan, R., Lee, A., and Ostrom, E. (2010). Lab experiments for the study of social-ecological systems. *Science*, 328(5978):613–617. doi:10.1126/science.1183532.

Jeffreys, H. (1961). *Theory of Probability*. Oxford: Clarendon Press, third edition.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795. doi:10.1080/01621459.1995.10476572.
W. L., Nelson, A. J., Nier, J. A., Packard, G., Pilati, R., Rutrich, A. M., Schmidt, K., Skorinko, J. L., Smith, R., Steiner, T. G., Storbeck, J., Van Swol, L. M., Thompson, D., van ’t Veer, A. E., Ann Vaughn, L., Vranka, M., Wichman, A. L., Woodzicka, J. A., and Nosek, B. A. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology*, 45:142–152. doi:10.1027/1864-9335/a000178.

Kovacs, A. M., Teglas, E., and Endress, A. D. (2010). The social sense: Susceptibility to others’ beliefs in human infants and adults. *Science*, 330(6012):1830–1834. doi:10.1126/science.1190792.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423. doi:10.1198/016214507000001337.

Ly, A., Etz, A., Marsman, M., and Wagenmakers, E.-J. (2018). Replication Bayes factors from evidence updating. *Behavior Research Methods*, 51(6):2498–2508. doi:10.3758/s13428-018-1092-x.

Mathur, M. B. and VanderWeele, T. (2017). New statistical metrics for multisite replication projects. doi:10.31219/osf.io/w89s5. To appear in Journal of the Royal Statistical Society: Series A (Statistics in Society).

Matthews, R. A. J. (2001a). Methods for assessing the credibility of clinical trial outcomes. *Drug Information Journal*, 35:1469–1478. doi:10.1177/009286150103500442.

Matthews, R. A. J. (2001b). Why should clinicians care about Bayesian methods? *Journal of Statistical Planning and Inference*, 94:43–71. doi:10.1016/S0378-3758(00)00232-9.

Matthews, R. A. J. (2018). Beyond ‘significance’: principles and practice of the analysis of credibility. *Royal Society Open Science*, 5(1). doi:10.1098/rsos.171047.

Micheloud, C. and Held, L. (2020). Power calculations for replication studies. URL https://arxiv.org/abs/2004.10814.

O’Hagan, A., Stevens, J. W., and Montmartin, J. (2001). Bayesian cost-effectiveness analysis from clinical trial data. *Statistics in Medicine*, 20(5):733–753. doi:10.1002/sim.861.

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

Patil, P., Peng, R. D., and Leek, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science*, 11:539–544. doi:10.1177/1745691616646366.

Pawel, S. and Held, L. (2020). Probabilistic forecasting of replication studies. *PLOS ONE*, 15(4):e0231416. doi:10.1371/journal.pone.0231416.

Pericchi, L. (2020). Discussion on the meeting on ‘signs and sizes: understanding and replicating statistical findings’. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(2):449–469. doi:10.1111/rssa.12544.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schönbrodt, F. D. and Wagenmakers, E.-J. (2017). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, 25(1):128–142. doi:10.3758/s13423-017-1230-y.
The sufficiently sceptical relative prior variance

The sufficiently sceptical relative prior variance at level $\gamma$ is the value $g_\gamma \in [0, g]$ that fulfills the condition

$$BF_0(g_\gamma) = \gamma.$$ (7)

Substituting and rearranging terms in (7), we obtain

$$\sqrt{1 + g_\gamma} \cdot \exp \left\{- \frac{1}{2} \cdot \frac{g_\gamma}{1 + g_\gamma} \cdot z_0^2 \right\} = \gamma \iff \frac{1}{\gamma} \cdot \exp \left\{- \frac{z_0^2}{2} \right\} = \frac{1}{\sqrt{1 + g_\gamma}} \cdot \exp \left\{- \frac{1}{2} \cdot \frac{z_0^2}{1 + g_\gamma} \right\}.$$
Squaring both sides and multiplying by \(-z_0^2\), this becomes

\[
\iff -\frac{z_0^2}{\gamma^2} \cdot \exp \{ -z_0^2 \} = -\frac{z_0^2}{1 + g_{\gamma}} \exp \left\{ -\frac{z_0^2}{1 + g_{\gamma}} \right\}.
\]

(8)

This is a transcendental equation that cannot be explicitly solved in terms of elementary functions. However, if we set \(q = -z_0^2/(1 + g_{\gamma})\) then (8) becomes

\[
-\frac{z_0^2}{\gamma^2} \cdot \exp \{ -z_0^2 \} = q \cdot \exp \{ q \}.
\]

The solution for \(q\) (and consequently for \(g_{\gamma}\)) can be explicitly computed using the Lambert \(W\) function (Corless et al., 1996) which is available in standard numerical software

\[
q = W \left( -\frac{z_0^2}{\gamma^2} \cdot \exp \{ -z_0^2 \} \right)
\]

\[
g_{\gamma} = \begin{cases} 
-\frac{z_0^2}{q} - 1 & \text{if } -\frac{z_0^2}{q} \geq 1 \\
\text{undefined} & \text{else}
\end{cases}
\]

(9)

For some \(z_0\), equation (8) can also be satisfied for negative \(g_{\gamma}\), which is why we need to add the condition \(-z_0^2/q \geq 1\) in equation (9), such that \(g_{\gamma}\) is a valid relative variance. Note that the Lambert \(W\) function is a multivalued function. Specifically, for real arguments, the two branches \(W_0\) and \(W_{-1}\) can be distinguished. Applying \(W_{-1}\) in our case returns \(g_{\gamma} \in [0, \infty)\) that fulfils (7), while \(W_0\) returns \(g_{\gamma} \in [\infty, \infty)\) that fulfils (7) due to the Jeffreys-Lindley’s paradox.

### B Computation of the sceptical Bayes factor

From the definition of the sceptical Bayes factor

\[
BF_S \equiv \inf_{g_{\gamma} \in [0, \infty]} \left\{ \gamma : BF_r(g_{\gamma}) \leq \gamma \right\},
\]

it is apparent that \(BF_S\) is either

1. undefined, if \(BF_r(g) > BF_o(g)\) for all \(g \in [0, \infty]\)
2. \(BF_S = BF_o\), if \(BF_r(g) < BF_o\)
3. \(BF_S = \inf_{g_{\gamma} \in [0, \infty]} \{ \gamma : BF_r(g_{\gamma}) = \gamma \}\), the height of the lowest intersection of \(BF_o(\cdot)\) and \(BF_r(\cdot)\) in \(g_{\gamma} \in [0, \infty]\) otherwise

Whether \(BF_S\) attains the lower bound \(BF_o\) (condition 2) can be checked by evaluating if \(BF_r(g) \geq BF_o\) and setting \(BF_S = BF_o\) if it is the case.

For condition 3, we know that the intersections between \(BF_o(\cdot)\) and \(BF_r(\cdot)\) satisfy

\[
BF_o(g_*) = BF_r(g_*)
\]

\[
\sqrt{\frac{1+c}{1+cg_*}} \cdot \exp \left\{ \frac{1}{2} \left( \frac{z_r^2}{1+cg_*} - \frac{(z_r-z_0\sqrt{c})^2}{1+c} \right) \right\} = \sqrt{1+g_*} \cdot \exp \left\{ -\frac{1}{2} \cdot \frac{g}{1+g_*} \cdot z_0^2 \right\}
\]

and the \(g_*\) fulfilling this can be computed using root-finding algorithms. Note that for the special case \(c = 1\) this can be rearranged to

\[
\frac{1}{1+g_*} \cdot \exp \left\{ \frac{1}{2} \cdot \frac{z_r^2 + z_0^2}{1+g_*} \right\} = \frac{1}{\sqrt{2}} \cdot \exp \left\{ -\frac{z_0^2}{2} \left( 1 + \left( 1 - d \right)^2 / 2 \right) \right\}
\]
where \( d = \hat{\theta}_r / \hat{\theta}_o = z_r / z_o \) is the relative effect estimate. Multiplying both sides by \(-z_A^2 = -(z_r^2 + z_o^2) / 2 \) (minus the arithmetic mean of the squared z-statistics \( z_r^2 \) and \( z_o^2 \)) we obtain

\[
\iff - \frac{z_A^2}{1 + g_*} \cdot \exp \left\{ - \frac{z_A^2}{1 + g_*} \right\} = - \frac{z_A^2}{2} \cdot \exp \left\{ - \frac{z_0^2}{2} \left( 1 + \frac{(1 - d)^2}{2} \right) \right\}.
\]

For \( q = -z_A^2 / (1 + g_*) \), this reads

\[
q \cdot \exp \{q\} = - \frac{z_A^2}{\sqrt{2}} \cdot \exp \left\{ - \frac{z_0^2}{2} \left( 1 + \frac{(1 - d)^2}{2} \right) \right\}
\]

and hence application of the Lambert \( W \) function leads to the solutions

\[
q = W \left( - \frac{z_A^2}{\sqrt{2}} \cdot \exp \left\{ - \frac{z_0^2}{2} \left( 1 + \frac{(1 - d)^2}{2} \right) \right\} \right)
\]

\[
g_* = \begin{cases} 
\frac{-z_A^2}{q} - 1 & \text{if } \frac{-z_A^2}{q} \geq 1 \\
\text{undefined} & \text{else}
\end{cases}
\]

with the condition that \(-z_A^2 / q \geq 1\) such that \( g_* \) is a valid relative variance, as the equation may otherwise be satisfied for negative \( g_* \). Since the argument to \( W(\cdot) \) is real and negative (if at least one of \( z_r, z_o > 0 \)), the branches \( W_1 \) and \( W_0 \) provide the two different solutions that can fulfill the equation, provided the argument is not smaller than \(-e^{-1}\) (which would mean that there are no intersections).