Reverse-Bayes Methods for Replication Studies and Beyond

Dissertation

zur
Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der
Mathematisch-naturwissenschaftlichen Fakultät
der
Universität Zürich

von

Samuel Pawel

von
Flawil SG

Promotionskommission

Prof. Dr. Leonhard Held (Vorsitz)
Prof. Dr. Reinhard Furrer
Prof. Dr. Guido Consonni

Zürich, 2023

Abstract

The fact that a scientific finding can be replicated in an independent replication study is central for its credibility. However, various large-scale replication failures have shown that the replicability of scientific findings is often lower than expected. This "replication crisis" has led to several methodological reforms in the past decade, an increased conduct of replication studies being one of them. Yet despite this rise of replication research, there is no consensus on which statistical methods should be used for the design and analysis of replication studies; Various methods already exist and various new ones have been proposed in response to the crisis. This thesis centres around the proposal from Held (2020), which is based on a reverse-Bayes approach. The key idea is to reverse the traditional "forward" use of Bayes' theorem ("prior + likelihood $\rightarrow$ posterior"), starting instead with a pre-specified posterior and deducing the corresponding prior ("posterior + likelihood $\rightarrow$ prior"). This allows to challenge the original finding by determining a sceptical prior for the underlying effect size such that the resulting posterior is no longer convincing. This prior then represents the position of a sceptic who remains unconvinced by the original study. Whether or not the sceptics' position is justified can then be assessed in light of the new data from the replication study. The larger the conflict between the replication data and the sceptic, the larger the degree of replication success. This procedure can be summarized in a single quantitative measure of replication success, termed the sceptical p-value.

The first and largest part of the thesis consists of extending this procedure. A first extension replaces tail probabilities by Bayes factors as measures of evidence. The sceptical prior is now determined such that the original finding is no longer convincing in terms of a Bayes factor. In contrast to tail probabilities, Bayes factors have a more natural interpretation and allow for direct quantification of evidence for one hypothesis versus another. Similarly as with the sceptical p-value, the procedure leads to a single measure quantifying the degree of replication success called the sceptical Bayes factor. A second extensions recalibrates the original procedure to produce more appropriate inferences in terms of effect size. The recalibration is chosen such that for borderline significant original studies, replication success can only be achieved if the replication effect estimate is larger than the original one. The third extension provides a framework for Bayesian design of replication studies. The framework allows combining the data from the original study with external knowledge, which leads to potentially more efficient designs compared to classical approaches.

The second part of the thesis is concerned with reverse-Bayes approaches in general. The reverse-Bayes idea was first proposed in the 1950s, but it has mostly been forgotten. To increase awareness and show potential use cases, reverse-Bayes history and methods are summarized in a comprehensive review. Furthermore, a short commentary on a recently proposed reverse-Bayes method draws connections to other reverse-Bayes methods.

The last part of the thesis revolves around research integrity issues in methodological research. Questionable research practices, such as selective reporting of results, are often seen as main cause for the low replicability in applied research. These practices can similarly harm methodological research, but this is often not recognized. To raise awareness, an illustrative simulation study is conducted in which it is shown how a novel method can easily be presented as superior over established competitor methods if questionable research practices are employed.

Key words: Bayesian inference, meta-science, replication studies

Dedicated to my mother Christa.
You are deeply missed.

Thesis outline

Preface

This thesis is submitted under the PhD program "Epidemiology and Biostatistics" at the University of Zurich for the degree of Doctor of Philosophy. The research contained in this thesis was conducted between October 2019 and December 2022. Financial support was provided by the Swiss National Science Foundation through the project "Reverse-Bayes design and analysis of replication studies" (project #189295) awarded to Leonard Held.

First and foremost I want to thank my supervisor Leo for giving me the opportunity to do this PhD, for showing me an open-minded and pragmatic approach to statistics, and for giving me the freedom to explore and develop into an independent researcher. I also want to thank the other two members of my PhD committee, Guido and Reinhard, for always providing excellent advice and being great collaborators. Another thank you goes to Robert Matthews who also is a great collaborator and without whom this PhD project would never exist as it was him who resurrected the reverse-Bayes approach from the dead almost twenty years ago. Furthermore, I want to thank Eric-Jan Wagenmakers for reviewing this thesis and for giving me the opportunity to do a very interesting and productive six months research stay in Amsterdam.

I also want to thank my friends and colleagues from the Epidemiology, Biostatistics, and Prevention Institute from the University of Zurich (in alphabetical order): Ainesh, Alexandra, Annina, Charlotte, Babette, Bálint, Dafne, Dominik, Eveline, Eva, Franscesca, Felix, Julia, Klaus, Kelly, Lucas, Lisa, Luisa, Goscha, Manuela, Monika, Muriel, Manja, Maria, Marielena, Martin, Mina, Nadja, Ruedi, Rachel, Sandra, Sona, Steffi, Tala, Torsten, Ulrike. Another "thanks" goes to my friends from the Department of Psychological Methods from the University of Amsterdam (in alphabetical order): Adam, Alexander, Alexandra, Alessandra, Alejandro, Angelika, Bruno, Don, František, Frederik, Jason, Johnny, Joris, Jill, Julia, Lukas, Maarten, Michelle, Nora, Omid, Quentin, René, Serjan, Suzanne, Ting. I also want to thank my good friends from outside academia (in alphabetical order): Chronis, Dani, Eleftheria, Flo, Fabi, Giuachin, Mirela, Peter. A special thanks goes to Ada. Finally, I thank my family Beni, Christa, Elisabeth, Fabian, Harry, Markus, Maja, Noemi for their support.

Zürich, December 2022

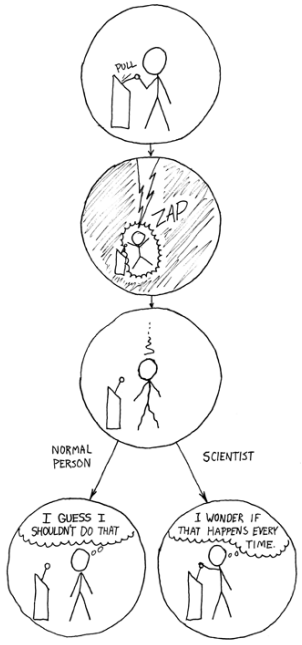
Samuel Pawel

Introduction

1 Replication studies

How can we know if a finding from a study is really true? For example, how can we know if the protective effect of a vaccine found in a study is real? The answer to this question is of highest importance to scientists and decision makers, but typically we can never know for sure as any study result comes with uncertainty.

One way to come a little closer to the truth, however, is to repeat that original study with new subjects. Such a *replication study* may then yield similar results which would make us more confident about the original finding, or it may yield conflicting results which would lower our confidence. Replication studies are thus an essential part of the scientific process as they provide a means for substantiating genuine research findings and refuting research findings which occurred merely by chance. For this reason, "successful" replication is often a requirement, for example, for acceptance of newly proposed scientific theories (e. g., a new physical model) or the implementation of policies based on scientific knowledge (e. g., market approval of a drug). The results from replication studies may thus have real world consequences, such as deciding whether we should get vaccinated (a consequence that anyone who experienced the COVID-19 pandemic is very aware of).



Replication studies as illustrated by Randall Munroe (https://xkcd.com/242/).

Despite their tremendous importance, the traditional academic system has made it unattractive for researchers to conduct replication studies; the currencies of science –publications, citations, and grant money– are typically easier to acquire by conducting novel research. This is because until relatively recently, top journals in many fields refused to publish replication studies as they were considered not "innovative" enough. To build a successful career, researchers thus often had no other choice than to concentrate their efforts on producing novel and eye-catching research results.

The perceived value of replication studies, however, has changed over the past decade. Earlier criticisms of low research standards (Altman, 1994; Ioannidis, 2005) were backed up by empirical evidence. For instance, pharmaceutical companies reported surprisingly low replication rates from pre-clinical research (Begley and Ellis, 2012) followed by later studies estimating that billions are wasted each year on flawed and non-replicable research in medicine and the life sciences (Chalmers et al., 2014; Freedman et al., 2015; Glasziou and Chalmers, 2018). Similarly, reports of fraud (Wicherts, 2011) and questionable research practices (Wagenmakers et al., 2011; Simmons et al., 2011; John et al., 2012) sparked intense discussion about the need for higher research standards in psychology and the social sciences. These discussions eventually culminated in large-scale replication projects conducted by huge researcher consortia in fields such as psychology (Open Science Collaboration, 2015; Klein et al., 2014, 2018; Protzko et al., 2020), economics (Camerer et al., 2016), the social sciences (Camerer et al., 2018), experimental philosophy (Cova et al., 2018), or cancer biology (Errington et al., 2021).

Most of these large-scale replication projects confirmed what many researchers had feared; carefully conducted replication studies often show less impressive results than their original counterparts, and the replicability of research findings is surprisingly low on average. This realization led many to declare science as being in a "replication crisis". Debates arose about whether or not the crisis really existed, and who or what was to blame (Gilbert et al., 2016; Amrhein et al., 2019b). Even the popular press became interested (e. g., Carey, 2015; Kovic, 2016; Achenbach and McGinley, 2017; Devlin, 2018), so that also in the eyes of the public the credibility of science became seriously threatened.

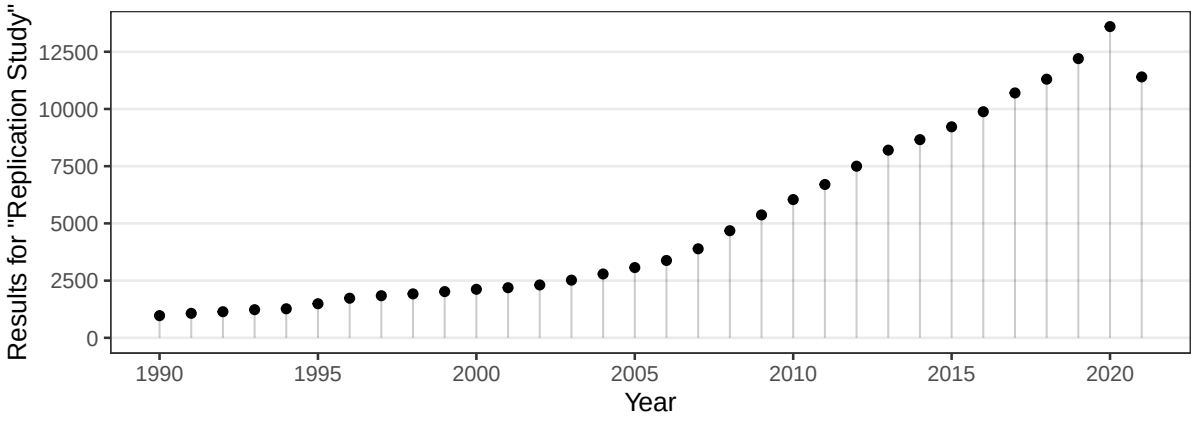


Figure 1: Number of results per year for search term "Replication Study" on Google Scholar. The search was conducted on 18 October 2022.

In the midst of this doomsday, various reforms were implemented to safe the reputation of science and to prevent a second crisis from happening (for an overview see e. g., Munafò et al., 2017). For instance, many journals adopted the "registered report" format (Chambers and Tzavella, 2021) in which a study proposal is peer reviewed *before* the study is conducted, and which, if accepted, gives provisional publication acceptance regardless of the study outcome. Similarly, digital infrastructure platforms, such as zenodo (`https://zenodo.org/`) or the open science framework (`https://osf.io/`), were created to facilitate preregistration, preprints, code, and data sharing, all of which have substantially increased over the last decade (Kidwell et al., 2016; Nosek et al., 2018; Rawlinson and Bloom, 2019). The practice of conducting replication studies has also gained popularity, and several journals and funders are now explicitly promoting and funding replication research (NWO, 2016; NSL, 2018; Nat, 2022). Figure 1 illustrates this trend via the yearly number of results for "Replication Study" on Google Scholar over the last three decades. We see that the numbers have been rapidly growing, especially after the mid 2000s (with a minor drop in 2021, perhaps because of research slowing down due to the COVID-19 pandemic).

1.1 The statistics of replication studies

Despite the increased interest in replication studies, the research community has not yet agreed on one important question: When is a replication study successful? For this reason, replication researchers typically report the results from different methods for assessing replication success. For example, Open Science Collaboration (2015) state "[t]here is no single standard for evaluating replication success. We evaluated [replicability] using significance and P values, effect sizes, subjective assessments of replication teams, and meta-analyses of effect sizes" (p. 11). In the following, I will give an overview about these and other methods which have been used in practice.

Table 1: Study level summary statistics for original and replication study. The cumulative distribution function of the standard normal distribution is denoted by $\Phi(\cdot)$, and the $1-\alpha$ quantile of the standard normal distribution is denoted by $\Phi^{-1}(1-\alpha) = z_\alpha$. Confidence intervals and p-values are based on a normal approximation.

	Original study	Replication study
effect estimate	$\hat{\theta}_o$	$\hat{\theta}_r$
standard error	σ_o	σ_r
$(1-\alpha)$ confidence interval	$[\hat{\theta}_o \pm z_{\alpha/2}\sigma_o]$	$[\hat{\theta}_r \pm z_{\alpha/2}\sigma_r]$
z-value	$z_o = \hat{\theta}_o/\sigma_o$	$z_r = \hat{\theta}_r/\sigma_r$
p-value (two-sided)	$p_o = 2\{1-\Phi(\lvert z_o\rvert)\}$	$p_r = 2\{1-\Phi(\lvert z_r\rvert)\}$

Most methods for analyzing replication studies can be formulated in terms of study level summary statistics as shown in Table 1. All of these are routinely reported in researcher articles, and if one of them is missing they typically can, under some assumptions, be back-calculated from the others. Using summary statistics is also often the only possible way for conducting the analysis as the raw data from the original study may not be available to the

replicators. The most important statistic is the effect estimate $\hat{\theta}$. It provides an estimate of the underlying effect size θ which quantifies the effect or association of an intervention/exposure with an outcome variable. Typical effect sizes are mean differences and correlations (for continuous outcomes), odds ratios (for binary outcomes), or hazard ratios (for time to event outcomes). Depending on the effect size type, a transformation might be required so that the assumption of approximately normally distributed effect estimates around the unknown effect size θ (for large enough sample sizes) is justifiable. This could be, for instance, the Fisher z-transformation for correlations, or the log-transformation for odds/hazard ratios (Cooper et al., 2019, chapter 11). The associated standard error σ represents the statistical uncertainty of the estimate. Under the assumption of (asymptotic) normality, confidence intervals for θ and p-values for testing the null hypothesis $H_0 \colon \theta = 0$ can be computed as shown in Table 1.

Table 2: Statistical criteria for assessing replication success which have been used in practice (Open Science Collaboration, 2015; Camerer et al., 2016, 2018; Cova et al., 2018; Errington et al., 2021).

Criterion type	Description
Significance	The original and replication p-values are smaller than a threshold α and their effect estimates show the same direction ($p_o < \alpha$, $p_r < \alpha$, and $\text{sign}(\hat{\theta}_o) = \text{sign}(\hat{\theta}_r)$). Usually $\alpha = 5\%$.
Meta-analytic significance	The meta-analytic p-value is smaller than a threshold α ($p_m = 2(1 - \Phi(\lvert\hat{\theta}_m\rvert/\sigma_m)) < \alpha$) where $\hat{\theta}_m = (\hat{\theta}_o/\sigma_o^2 + \hat{\theta}_r/\sigma_r^2)\sigma_m^2$ is the pooled effect estimate with standard error $\sigma_m = (1/\sigma_o^2 + 1/\sigma_r^2)^{-1/2}$. Usually $\alpha = 5\%$.
Relative effect size	The effect estimate of the replication study goes in the same direction and is at least as large as d_{min} times the original one ($d = \hat{\theta}_r/\hat{\theta}_o \geq d_{\text{min}}$). Usually $d_{\text{min}} = 1$.
Confidence interval	The replication effect estimate is contained in the $(1 - \alpha)$ original confidence interval ($\hat{\theta}_r \in [\hat{\theta}_o \pm z_{\alpha/2}\sigma_o]$). Sometimes, also defined as the original effect estimate is contained in the $(1 - \alpha)$ replication confidence interval ($\hat{\theta}_o \in [\hat{\theta}_r \pm z_{\alpha/2}\sigma_r]$). Usually $\alpha = 5\%$.
Prediction interval (Q-test)	The replication effect estimate is contained with its $(1 - \alpha)$ prediction interval ($\hat{\theta}_r \in [\hat{\theta}_o \pm z_{\alpha/2}\sqrt{\sigma_o^2 + \sigma_r^2}]$). Usually $\alpha = 5\%$. This criterion is equivalent to $p_Q \geq \alpha$ where p_Q is the p-value from the meta-analytic Q-test[1].

Table 2 lists commonly used criteria for replication success in terms of the summary statistics from Table 1. The most popular criterion defines replication success by simultaneous statistical significance of original and replication study along with their effect estimates showing the same direction. This approach is also called the "two-trials rule" (Senn, 2008) in drug reg-

[1]The p-value from the Q-test is $p_Q = 2\{1 - \Phi(\sqrt{Q})\}$ where $Q = \sum_{i \in \{o,r\}}(\hat{\theta}_i - \hat{\theta}_m)^2\sigma_i^{-2} = (\hat{\theta}_o - \hat{\theta}_r)^2(\sigma_o^2 + \sigma_r^2)^{-1}$ is the Q-statistic and $\hat{\theta}_m$ is the fixed-effects pooled estimate as defined in Table 2.

ulation or "vote-counting" in meta-analysis (Cooper et al., 2019). The criterion can similarly be defined through one-sided p-values, so that the original and replication effect estimate are not required to show the same direction as this is taken into account by the p-values. Some replication projects (Open Science Collaboration, 2015; Errington et al., 2021) also defined replication success via simultaneous non-significance of original and replication study ($p_o > \alpha$ and $p_r > \alpha$). However, with this definition "replication success" can almost always be achieved by conducting original and replication study with just very few samples so that the p-values are large. The approach is also logically questionable as it could be seen as an instance of the "absence of evidence fallacy" (Altman and Bland, 1995) meaning that the failure to find evidence against the null hypothesis is erroneously interpreted as evidence for the null hypothesis. Meta-analytic extensions of the significance approach define replication success by significance of a combined effect estimate[2]. Typically, the assumption of a common underlying effect size is seen as reasonable so that fixed-effects meta-analysis is used for pooling. Random-effects meta-analysis has mostly been used if more than one replication study of the same original study are conducted since in this case replicators are often interested in also understanding between-replication heterogeneity (e. g., in Klein et al., 2018).

The remaining criteria in Table 2 put more emphasis on compatibility in effect size between original and replication study. For example, the relative effect estimate $d = \hat{\theta}_r / \hat{\theta}_o$ quantifies how much the replication effect estimate changed compared to the original one, and the smaller d the smaller the degree of replication success. Some projects also report a confidence interval for d (Camerer et al., 2016, 2018), while others ignore its uncertainty and make a binary cut at one to define replication success (Errington et al., 2021).

In contrast, the criteria based on confidence and prediction intervals define effect size compatibility on an absolute scale. However, the criterion based on the original confidence interval ignores the uncertainty from the replication, whereas the criterion based on the replication confidence interval ignores the uncertainty from the original study. The prediction interval criterion, on the other hand, takes into account both sources of uncertainty (Patil et al., 2016). Yet, also declaring "replication success" based on the prediction interval may be logically questionable due to its connection to the meta-analytic Q-test. That is, if the p-value from the Q-test is larger than α this is equivalent to the replication effect estimate $\hat{\theta}_r$ being contained in its $(1 - \alpha)$ prediction interval. The null hypothesis of this test is defined that the underlying effect sizes of original and replication study are the same ($H_0 \colon \theta_o = \theta_r$), so a rejection of this null hypothesis corresponds to demonstrating replication failure and not replication success. Interpreting a failure to reject the null hypothesis as evidence for it is again an instance of the absence of evidence fallacy (Hedges and Schauer, 2019). As in the case of defining replication success by simultaneous non-significance of original and replication study, the mismatch of the null hypothesis of the prediction interval criterion results in the undesirable property that "replication success" can almost always be achieved if the sample size of the studies is small enough as the prediction interval becomes wider with larger standard errors.

[2]The meta-analytic approach can also be given a Bayesian interpretation: A flat prior distribution for the underlying effect size θ is updated by the data from original and replication study. Replication success defined by a meta-analytic p-value being smaller than α is then equivalent to replication success via a Bayesian posterior probability $\Pr(\theta > 0 \,|\, \hat{\theta}_o, \hat{\theta}_r) > 1 - \alpha/2$, respectively, $\Pr(\theta < 0 \,|\, \hat{\theta}_o, \hat{\theta}_r) > 1 - \alpha/2$ depending on the direction of combined estimate.

Example: Reproducibility Project Cancer Biology

I will now illustrate the assessment of replicability on data from the Reproducibility Project: Cancer Biology (Errington et al., 2021). This large-scale project attempted to replicate 53 landmark studies from the field of cancer biology. However, due to various difficulties (e. g., missing information from the original studies or problems in conducting the experiments) only 23 of the 53 studies could be repeated. Errington et al. (2021) report that these experiments led to data on 158 quantitative effects. However, from the data which they provide only 132 quantitative effects come with original and replication standardized mean difference effect estimates along with standard errors, and only these data will be used in the subsequent analyses.

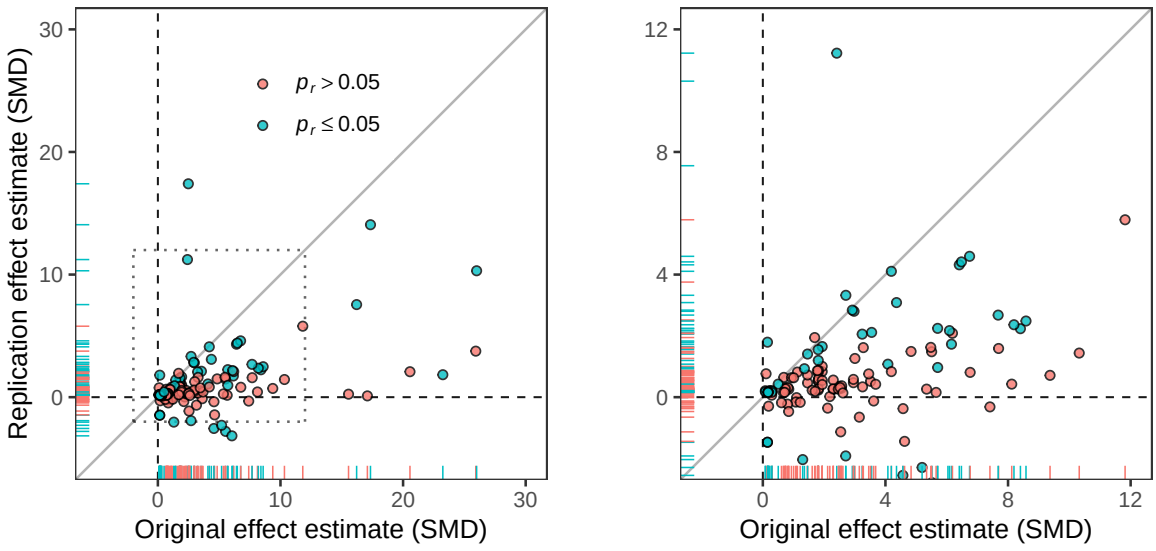


Figure 2: Results for 132 effects from the Reproducibility Project: Cancer Biology (Errington et al., 2021) for which effect estimates and standard errors are available on standardized mean difference scale. The right plot shows a zoomed-in view of the dotted area in the left plot. The *p*-values are recomputed using a normal approximation. Two study pairs with original effect estimate larger than 80 are not shown.

Figure 2 shows the original versus the replication effect estimate (on standardized mean difference scale) with the color indicating whether the replication study was statistically significant at the 5% level (two-sided). As in most other replication projects, the majority of the replications show smaller effect estimates compared to their original counterparts (mean relative effect estimate $\bar{d} = 0.38$). Many of the replications also fail to achieve statistical significance at the 5% level. Specifically, from the 94 effects which were significant in the original study only 32 were also significant in the replication study (in the same direction).

Figure 3 shows how many of the replications are successful according to the criteria from Table 2, and their combinations. For the total 132 replications, most successes occur for the meta-analytic significance (83) and the prediction interval criteria (83), followed by significance

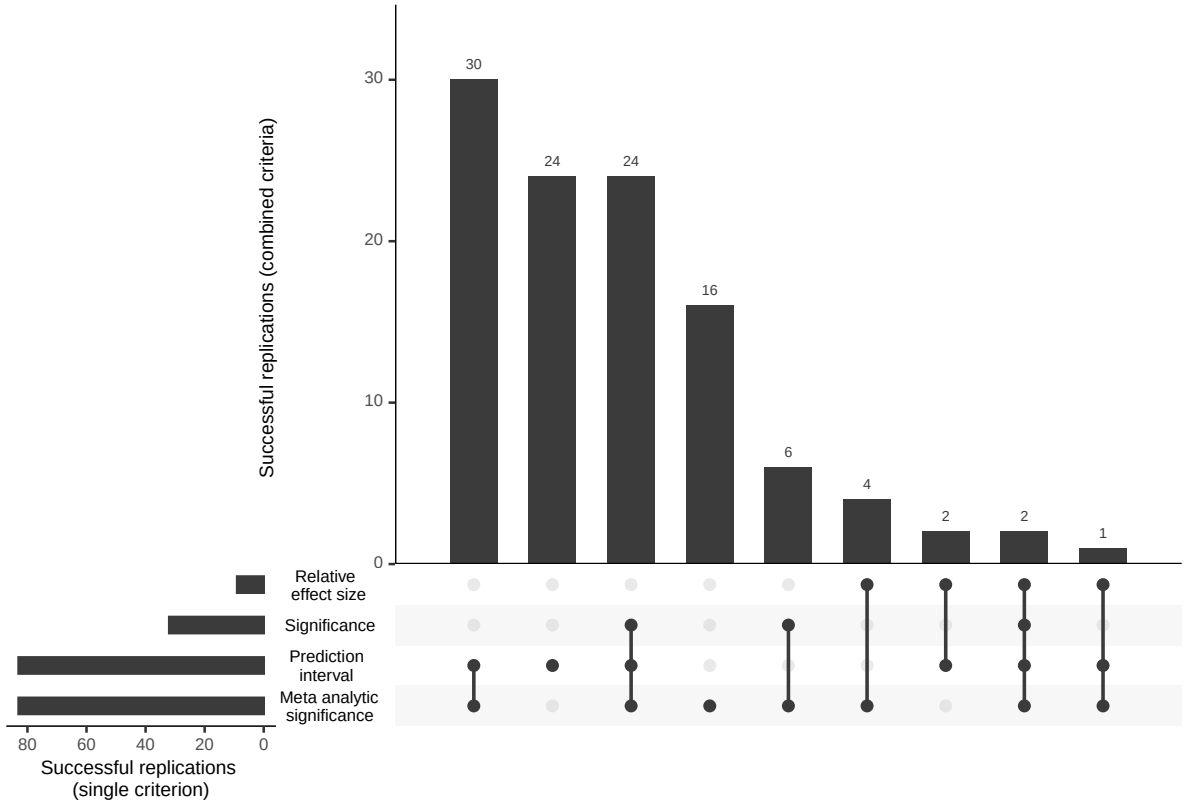


Figure 3: Upset plot for data on 132 effects from the Reproducibility Project: Cancer Biology (Errington et al., 2021) for which effect estimates and standard errors are available on standardized mean difference scale. Shown are the number of replication successes according to the different criteria from Table 2 and their combinations. A level $\alpha = 5\%$ and a relative effect size threshold $d_{\text{min}} = 1$ are used.

(32), and relative effect size (9). Looking at the combinations of the criteria, only in two replications are all of them satisfied simultaneously.

A detailed view for a subset of the data from the project is given in Table 3. For no single study pair in the table are all commonly used criteria satisfied simultaneously. For instance, the two pairs (48, 2, 1) and (16, 3, 3) at the top of the table satisfy the significance criterion ($p_o < 0.05$ and $p_r < 0.05$), meta analytic significance ($p_m < 0.05$) and the Q-test/prediction interval criterion ($p_Q > 0.05$), yet the replication effect estimate is smaller than the original one ($d < 1$). Similarly, there is no single study pair for which all criteria indicate replication failure. For example, the pair (5, 1, 3) at the bottom of the table is a clear failure with respect to the relative effect estimate ($d < 1$) and the significance criteria ($p_o > 0.05$ and $p_r > 0.05$), yet the Q-test does not indicate evidence for inconsistency of the two effect estimates ($p_Q > 0.05$).

Taken together, this analysis demonstrates that conclusions based on commonly used replication success criteria often differ. It is not always clear cut whether or not a replication is successful. Reducing replicability to a single criterion without mentioning these difficulties, as often done by the popular press (e. g., "more than half of the findings did not hold up when retested" in Carey, 2015), is a clear simplification of the matter.

Table 3: Subset of results from the Reproducibility Project: Cancer Biology (Errington et al., 2021). (P, X, E) denotes effect E from experiment X, from original paper P. Shown are original $\hat{\theta}_o$, replication $\hat{\theta}_r$, and pooled effect estimate $\hat{\theta}_m$ with 95% confidence intervals, variance ratio $c = \sigma_o^2/\sigma_r^2$, relative effect estimate $d = \hat{\theta}_r/\hat{\theta}_o$, original p-value p_o, replication p-value p_r, meta-analytic p-value p_m, Q-test p-value p_Q, and sceptical p-value p_S. A level $\alpha = 5\%$ is used as threshold for replication success via p-values, and a threshold $d_{\text{min}} = 1$ for the relative effect estimate.

(P, X, E)	$\hat{\theta}_o$ [95% CI]	$\hat{\theta}_r$ [95% CI]	$\hat{\theta}_m$ [95% CI]	c	d	p_o	p_r	p_m	p_Q	p_S
(48, 2, 1)	0.22 [0.15, 0.30]	0.20 [0.13, 0.28]	0.21 [0.16, 0.27]	0.96	0.91	< 0.0001	< 0.0001	< 0.0001	0.69	< 0.0001
(16, 3, 3)	4.36 [2.66, 6.06]	3.09 [1.42, 4.75]	3.71 [2.52, 4.90]	1.04	0.71	< 0.0001	0.00028	< 0.0001	0.29	0.0033
(19, 1, 2)	2.41 [-0.33, 5.15]	11.22 [4.21, 18.24]	3.58 [1.03, 6.13]	0.15	4.65	0.084	0.0017	0.006	0.022	0.094
(50, 1, 1)	0.50 [0.29, 0.71]	0.43 [0.10, 0.75]	0.48 [0.30, 0.66]	0.42	0.85	< 0.0001	0.0098	< 0.0001	0.7	0.016
(1, 3, 5)	5.70 [0.48, 10.92]	2.25 [0.39, 4.10]	2.63 [0.89, 4.38]	7.94	0.39	0.032	0.018	0.0031	0.22	0.25
(24, 4, 5)	2.92 [-0.36, 6.20]	2.84 [0.18, 5.50]	2.87 [0.81, 4.94]	1.52	0.97	0.081	0.037	0.0065	0.97	0.2
(6, 1, 1)	6.41 [-2.90, 15.72]	4.32 [0.16, 8.47]	4.67 [0.87, 8.46]	5.01	0.67	0.18	0.042	0.016	0.69	0.37
(29, 2, 2)	1.30 [-0.24, 2.84]	-2.04 [-4.02, -0.05]	0.05 [-1.17, 1.26]	0.60	-1.57	0.097	0.044	0.94	0.0092	0.18
(39, 1, 1)	2.54 [0.39, 4.69]	-1.13 [-2.40, 0.13]	-0.19 [-1.28, 0.90]	2.90	-0.45	0.021	0.079	0.73	0.0039	0.23
(20, 2, 2)	1.78 [0.60, 2.96]	0.87 [-0.28, 2.01]	1.31 [0.49, 2.13]	1.06	0.49	0.0032	0.14	0.0018	0.28	0.19
(28, 3, 3)	2.11 [0.08, 4.15]	-0.36 [-0.86, 0.15]	-0.21 [-0.70, 0.28]	16.21	-0.17	0.042	0.17	0.39	0.021	0.46
(44, 1, 4)	9.37 [7.81, 10.93]	0.71 [-0.34, 1.77]	3.43 [2.55, 4.30]	2.19	0.08	< 0.0001	0.19	< 0.0001	< 0.0001	0.19
(9, 2, 3)	1.80 [0.62, 2.99]	0.55 [-0.45, 1.55]	1.07 [0.31, 1.84]	1.40	0.3	0.0028	0.28	0.006	0.11	0.32
(47, 1, 5)	0.75 [-0.42, 1.92]	0.31 [-0.49, 1.12]	0.45 [-0.21, 1.12]	2.11	0.42	0.21	0.45	0.18	0.55	0.55
(37, 1, 1)	2.96 [1.14, 4.77]	0.47 [-0.81, 1.75]	1.30 [0.25, 2.34]	2.01	0.16	0.0014	0.47	0.015	0.028	0.49
(21, 1, 3)	0.61 [-0.67, 1.88]	-0.18 [-0.92, 0.56]	0.02 [-0.62, 0.66]	2.94	-0.3	0.35	0.63	0.95	0.29	0.7
(42, 2, 2)	5.34 [2.14, 8.54]	0.26 [-0.84, 1.36]	0.80 [-0.24, 1.84]	8.42	0.05	0.0011	0.64	0.13	0.0033	0.66
(15, 2, 1)	1.61 [0.46, 2.76]	0.22 [-0.83, 1.28]	0.86 [0.08, 1.64]	1.19	0.14	0.006	0.68	0.031	0.082	0.68
(41, 2, 1)	1.69 [-0.97, 4.35]	1.95 [-20.93, 24.82]	1.69 [-0.95, 4.34]	0.01	1.15	0.21	0.87	0.21	0.98	0.87
(5, 1, 3)	1.10 [-2.06, 4.26]	-0.02 [-2.33, 2.29]	0.37 [-1.49, 2.24]	1.88	-0.01	0.5	0.99	0.7	0.58	0.99

1.2 Reverse-Bayes assessment of replication studies

In response to the lack of a standard criterion for replication success, various methods have been proposed (Verhagen and Wagenmakers, 2014; Simonsohn, 2015; Anderson and Maxwell, 2016; Patil et al., 2016; Johnson et al., 2016; Etz and Vandekerckhove, 2016; van Aert and van Assen, 2017; Ly et al., 2018; Harms, 2019; Hedges and Schauer, 2019; Mathur and VanderWeele, 2020; Held, 2020; Pawel and Held, 2020; Bonett, 2020, among others). The focus of this thesis is to refine and extended the proposal from Held (2020) which combines *reverse-Bayes inference* and *Bayesian model criticism* in a method for assessing replication success. In the following, I will summarize its main ideas and technical underpinnings.

Reverse-Bayes inference

Bayesian inference is an approach to statistical inference where Bayes' theorem is used to make probability statements about unknown parameters based on the observed data. The central quantity for doing so is the distribution of the parameters conditional on the data, called the *posterior distribution*. It can be obtained from Bayes' theorem

$$f(\theta \mid \text{data}) = f(\theta) \times \frac{f(\text{data} \mid \theta)}{f(\text{data})}$$

meaning that the *prior distribution* for the parameter θ with density/probability mass function $f(\theta)$ is multiplied by the (normalized) likelihood of the data, also know as *Bayesian updating*. Parameter values which increase the likelihood of the data become more likely *a posteriori*, however, they are also weighted by their plausibility *a priori* through the prior. As such, Bayesian inference provides a formal way for combining information from the data at hand with external knowledge encoded in the prior.

However, many also consider the prior to be the weak point of Bayesian inference, as it is unclear how it should be specified in the absence of external knowledge. The *reverse-Bayes* approach, first proposed by Good (1950), is one way of dealing with this issue. The idea is to flip Bayes' theorem around

$$f(\theta) = f(\theta \mid \text{data}) \times \frac{f(\text{data})}{f(\text{data} \mid \theta)}$$

and instead "downdate" a posterior with the observed data. So, in contrast to the conventional "forward-Bayes" approach where we start with a prior, update it with the data, and end up with a posterior, the reverse-Bayes approach starts with the posterior and ends up with the prior. Reverse-Bayes inference then revolves around the question whether the resulting prior is plausible in light of external knowledge, and if so, this could be seen as support for the specified posterior.

To illustrate reverse-Bayes inference, let us return to the replication setting. Assume we want to conduct inference about the unknown effect size θ based on the effect estimate from the original study $\hat{\theta}_o$ and its standard error σ_o. We will assume that $\hat{\theta}_o$ is normally distributed around the unknown effect size θ with (known) variance equal to its squared standard error

σ_o^2, here and henceforth denoted by $\hat{\theta}_o \,|\, \theta \sim \mathrm{N}(\theta, \sigma_o^2)$. Furthermore, we specify a zero-mean normal prior with variance τ^2 for the effect size θ, representing the position of a sceptic who does not believe in the presence of a non-zero effect. The "stubbornness" of the sceptic is determined by how large the variance τ^2 is chosen. Combining the sceptical prior with the likelihood produces then a posterior which is again normal $\theta \,|\, \hat{\theta}_o, \sigma_o \sim \mathrm{N}(\mu_{\text{post}}, \sigma_{\text{post}}^2)$ with mean and variance

$$\mu_{\text{post}} = \frac{\hat{\theta}_o}{1 + \sigma_o^2/\tau^2} \qquad \text{and} \qquad \sigma_{\text{post}}^2 = \frac{1}{1/\sigma_o^2 + 1/\tau^2}.$$

The associated $(1 - \alpha)$ highest posterior density credible interval is given by

$$\left[\mu_{\text{post}} \pm z_{\alpha/2}\, \sigma_{\text{post}}\right] \tag{1}$$

and if this credible interval excludes parameter values smaller/larger than zero (depending on the orientation of the effect size) this may be interpreted as evidence[3] for a genuine effect found in the original study.

Depending on how large the prior variance τ^2 is chosen, the posterior credible interval (1) will either include or exclude zero. Different data analysts may have different degrees of scepticism and may thus choose different prior variances τ^2. As a default choice, Held (2020) proposed to use the reverse-Bayes approach from Matthews (2001), that is, to determine the *sufficiently sceptical prior variance* τ_α^2 so that the appropriate limit of the $(1 - \alpha)$ credible interval is just fixed to zero. The resulting prior then represent the beliefs of a sceptic who is just stubborn enough to not find the original study convincing at level α.

Figure 4 illustrates the derivation of the sufficiently sceptical prior variance for an original study included in the Reproducibility Project: Cancer Biology with standardized mean difference effect estimate $\hat{\theta}_o = 1.46$ and standard error $\sigma_o = 0.57$. We see that a sufficiently sceptical prior with variance $\tau_\alpha^2 = 0.68^2$ is required to render the resulting posterior no longer convincing at level $\alpha = 5\%$.

Held (2019a) showed that the sufficiently sceptical prior variance τ_α^2 for a level α is available in closed-form

$$\tau_\alpha^2 = \begin{cases} \dfrac{\sigma_o^2}{\left(z_o^2/z_{\alpha/2}^2\right) - 1} & \text{if } z_o^2 > z_{\alpha/2}^2 \\ \text{undefined} & \text{else.} \end{cases} \tag{2}$$

From (2) we see that convincing original studies (those with large absolute z-values $|z_o|$) require smaller sufficiently sceptical prior variances to render the posterior no longer convincing for the same level α. Conversely, if the original study is not convincing enough (if $|z_o| \leq z_{\alpha/2}$) the sufficiently sceptical prior variance is undefined meaning that the data provide so little evidence on their own that no scepticism is required to make them unconvincing.

This shows that the reverse-Bayes approach based on sceptical priors can be used to formally "challenge" the finding of an original study. However, once this sceptical prior is determined,

[3]For readers who do not agree with this notion of evidence, do not panic. In this thesis we will extend this reverse-Bayes procedure to use alternative measures of evidence, such as Bayes factors.

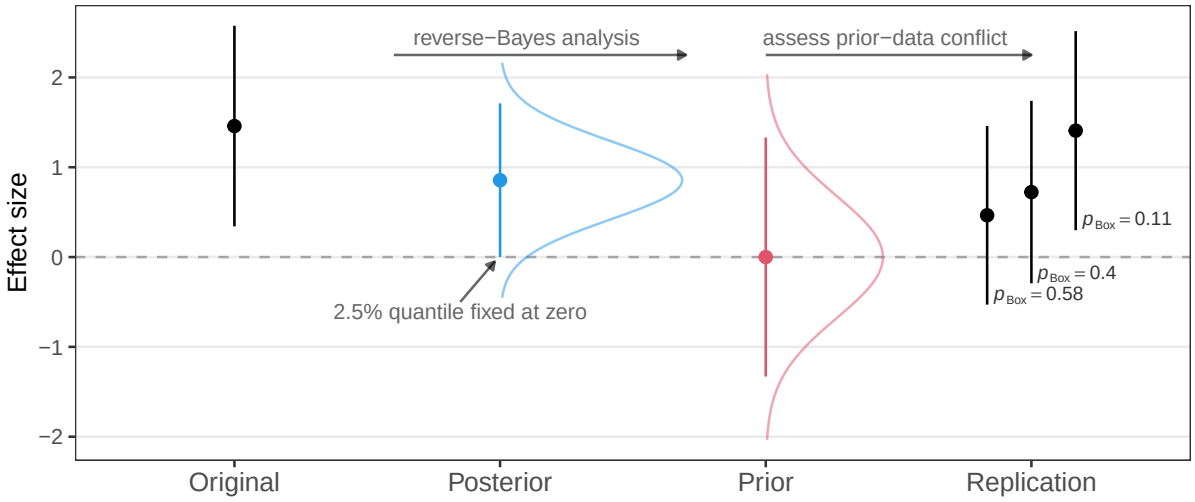


Figure 4: Illustration of reverse-Bayes assessment of replication success using data from the original study (Paper 9, Experiment 2, Effect 5) and its three replication studies from the Reproducibility Project: Cancer Biology (Errington et al., 2021). Shown are effect estimates and prior/posterior means with 95% confidence/credible interval. The original finding is challenged with a "sceptical" prior, sufficiently concentrated around zero so that the resulting posterior is no longer convincing at level $\alpha = 5\%$.

the question becomes whether it is plausible in light of external data. A natural candidate for answering the question are data from a replication study. In the following, I will show how Bayesian model criticism can be used for doing so.

Bayesian model criticism

Model criticism describes the assessment of compatibility between observed data and their assumed statistical model. If incompatibility is diagnosed, this alarms the data analyst that inferences based on the model may be invalid and modifications may be required. A formal framework for Bayesian model criticism was first introduced by Box (1980). To understand whether a Bayesian model M consisting of a joint distribution for parameter θ and data is adequate, Box gave the following fundamental decomposition of the joint distribution

$$f(\theta, \text{data} \,|\, \text{M}) = f(\theta \,|\, \text{data}, \text{M}) \times f(\text{data} \,|\, \text{M}).$$

He reasoned that inferences based on the left factor, the posterior distribution $f(\theta \,|\, \text{data}, \text{M})$, should only be trusted if the right factor, the prior predictive distribution

$$f(\text{data} \,|\, \text{M}) = \int f(\text{data} \,|\, \theta, \text{M})\, f(\theta \,|\, \text{M})\, \text{d}\theta$$

is compatible with the observed data. If the model M was indeed adequate, the empirical distribution of the observed data should be close to their predictive distribution under the model M. On the other hand, if the empirical distribution differed from the predictive distribution, this would imply that model M is inadequate due to misspecification of the likelihood $f(\text{data} \,|\, \theta, \text{M})$ and/or misspecification of the prior $f(\theta \,|\, \text{M})$.

Based on these observations, Box proposed two general approaches for conducting Bayesian model criticism. First, the predictive density of the observed data (or the value of a "checking function" applied to the observed data) can be compared to its reference distribution via a *prior predictive p-value*

$$p_{\text{Box}} = \Pr\left\{f(\text{data} \mid \text{M}) < f(\text{observed data} \mid \text{M})\right\}, \tag{3}$$

that is, the probability of obtaining data with lower predictive density ("more surprising" data) than the observed data. The lower the p-value p_{Box}, the more incompatibility between the observed data and the assumed model M. This approach was used by Held (2020), and it will soon be explained in more detail. However, Box also mentioned a second approach which has mostly been forgotten. If a second "benchmarking" model M_2 alternative to the model under investigation M_1 is available, Box proposed that the *prior predictive ratio*

$$\text{PR}_{\text{Box}} = \frac{f(\text{observed data} \mid \text{M}_1)}{f(\text{observed data} \mid \text{M}_2)},$$

the ratio of predictive densities from the observed data under both models, could be used to judge the relative adequacy of model M_1. Again, the lower the predictive ratio PR_{Box}, the less compatible the observed data with the model M_1. Bayesian model criticism approaches based on predictive ratios will be used in later parts of this thesis[4].

We now return to the replication setting. Having obtained a sceptical prior $\theta \sim \text{N}(0, \tau_\alpha^2)$ with sufficiently sceptical prior variance τ_α^2 from (2), the aim is now to assess its adequacy in light of the data from a replication study. If we are able to show that the prior is inadequate, this would demonstrate that scepticim regarding the original finding is unjustified and that the original study therefore provided evidence for a genuine effect. Under the assumption of a normal likelihood for the replication effect estimate, i. e., $\hat{\theta}_r \mid \theta \sim \text{N}(\theta, \sigma_r^2)$, the prior predictive distribution is given by $\hat{\theta}_r \mid \theta \sim \text{N}(0, \sigma_r^2 + \tau_\alpha^2)$. As the prior predictive distribution is symmetric around zero, the prior predictive p-value (3) is

$$p_{\text{Box}} = 2\left\{1 - \Phi\left(\frac{|\hat{\theta}_r|}{\sqrt{\sigma_r^2 + \tau_\alpha^2}}\right)\right\}. \tag{4}$$

Figure 4 shows the prior predictive p-values from (4) computed for three example replication studies from the Replication Project: Cancer Biology. We see that larger effect estimates show prior predictive p-values p_{Box}, indicating more prior-data conflict. This is because the standard errors from all three replications are roughly the same size, so that mostly the distance between zero and the replication effect estimate matters. The p-values suggest that there is hardly any conflict between the sceptical prior and the first two replications (those with $p_{\text{Box}} = 0.58$ and $p_{\text{Box}} = 0.4$), while the conflict seems larger for the third one (the one with $p_{\text{Box}} = 0.11$).

[4]Some readers may have noted that the predictive ratio is also the Bayes factor (Jeffreys, 1961; Good, 1958) contrasting model M_1 to M_2. The prior predictive ratio model criticism approach is therefore particularly useful in combination with reverse-Bayes procedures based on Bayes factors, as will be demonstrated in this thesis.

Held (2020) then defined replication success at level α by

$$p_{\text{Box}} \leq \alpha,$$

that is, replication success is established if there is more conflict between the sceptical prior and the replication data than there was evidence against the null hypothesis in the original study. For the examples in Figure 4, all prior predictive p-values are larger than the level $\alpha = 5\%$ used for computing the sufficiently sceptical prior variance ($\tau_\alpha^2 = 0.68^2$), so neither of them achieves replication success at level $\alpha = 5\%$. However, at a larger level, e. g., $\alpha = 10\%$, the corresponding sufficiently sceptical prior variance would be smaller ($\tau_\alpha^2 = 0.48^2$). Consequently, there would be more conflict between the prior and the replication data, so that the third replication would be successful (since the prior predictive p-value would be $p_{\text{Box}} = 0.057 < 10\%$).

The sceptical *p*-value

To remove the dependence on the level α, Held (2020) proposed to determine the smallest level at which replication success can be established. He called this level the *sceptical p-value* p_{S}, and showed that it is available in closed-form

$$p_{\text{S}} = 2\left\{1 - \Phi(|z_{\text{S}}|)\right\}$$

where

$$z_{\text{S}}^2 = \begin{cases} z_{\text{H}}^2/2 & \text{for } c = 1 \\ \left\{\sqrt{\left[z_{\text{A}}^2\left\{z_{\text{A}}^2 + z_{\text{H}}^2(c-1)\right\}\right]} - z_{\text{A}}^2\right\}/(c-1) & \text{for } c \neq 1 \end{cases}$$

with arithmetic mean $z_{\text{A}}^2 = (z_o^2 + z_r^2)/2$ and harmonic mean $z_{\text{H}}^2 = 2/(1/z_o^2 + 1/z_r^2)$ of the squared z-statistics, and variance ratio $c = \sigma_o^2/\sigma_r^2$. Replication success at level α is then equivalent to $p_{\text{S}} \leq \alpha$. For instance, the sceptical p-values of the three replication studies in Figure 4 are $p_{\text{S}} = 0.39$, $p_{\text{S}} = 0.23$, and $p_{\text{S}} = 0.075$ (from left to right), so we can see that the third replication is unsuccessful at level $\alpha = 5\%$ but successful at level $\alpha = 10\%$. However, the sceptical p-value does not necessarily have to be dichotomized but can also be interpreted as a quantitative measure of replication success, the smaller p_{S} the higher the degree of replication success.

The sceptical p-value has several interesting properties (Held, 2020, section 3): First, it is always larger than the maximum of the original and replication p-values ($p_{\text{S}} > \max\{p_o, p_r\}$), meaning that both p-values have to be smaller than α so that replication success based on $p_{\text{S}} \leq \alpha$ is possible. Second, if the p-values p_o and p_r remain fixed but the relative effect estimate $d = \hat{\theta}_r/\hat{\theta}_o$ decreases, the sceptical p-value increases, meaning that shrinkage of the replication effect estimate is penalized. The sceptical p-value hence requires *both* studies to be sufficiently convincing on their own (in terms of their p-values), similarly to the significance criterion for replication success. Unlike the significance criterion, however, if the p-values remain fixed but the replication effect estimate $\hat{\theta}_r$ becomes smaller than the original estimate $\hat{\theta}_o$, the sceptical p-value indicates less replication success. This property seems desirable in

the replication setting as a smaller replication effect estimate $\hat{\theta}_r$ may not be practically relevant anymore, despite its statistical significance.

The results from the Reproducibility Project: Cancer Biology in Table 3 illustrate these two properties. The third study from above (19, 1, 2) fails to achieve replication success at level $\alpha = 5\%$ with the sceptical p-value, even though the replication study was highly convincing –the effect estimate was almost five times as large as in the original study. Yet, as the approach requires both studies to be convincing on their own –and the original study was not significant at the 5% level– replication success at this level is impossible with the sceptical p-value. The second property is illustrated by the fifth study from above (1, 3, 5). Here the original study was convincing and also the replication study achieved significance. However, the effect estimate from the replication was roughly 60% smaller than the one from the original study, and significance is merely achieved because the standard error was much smaller (i. e., around $1/\sqrt{c} \approx 1/2.8$ times). The sceptical p-value is therefore only $p_S = 0.25$, indicating hardly any replication success.

This concludes the introduction to replication studies and the summary of the reverse-Bayes assessment of replication success. The interested reader is referred to the original article by Held (2020) for additional properties (e. g., the null distribution of the sceptical p-value) and extensions of the procedure (e. g., a one-sided version of the sceptical p-value, power and sample size calculations). However, this introduction should have prepared the reader well enough to understand the remaining parts of this thesis.

Software and data

Code and data to reproduce the analyses and recompile the thesis are available at `https://doi.org/10.5281/zenodo.XXXXXXX`. All analyses were conducted in the R programming language version 4.2.2 (R Core Team, 2022). The packages `dplyr` (Wickham et al., 2022), `ggplot2` (Wickham, 2016), `knitr` (Xie, 2022), `ReplicationSuccess` (Held, 2020), `UpSetR` (Gehlenborg, 2019), and `xtable` (Dahl et al., 2019) were used. The CC-By 4.0 licensed data from the Reproducibility Project: Cancer Biology (Errington et al., 2021) were downloaded from `https://doi.org/10.17605/osf.io/e5nvr`. The relevant variables were then extracted from the file "`RP_CB Final Analysis - Effect level data.csv`".

2 Thesis summary

This thesis consists of six papers. Paper I and II (and to a lesser extent III) focus on extending the reverse-Bayes assessment of replication success from Held (2020). Paper III presents a general Bayesian framework for replication study design. Paper IV is a review paper about reverse-Bayes methodology. Paper V is a short comment on another article which proposed a reverse-Bayes method. Paper VI lists and illustrates questionable research practices in simulation studies.

Paper I: The sceptical Bayes factor for the assessment of replication success

by Samuel Pawel, Leonhard Held
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2022, 84(3), 879–911. https://doi.org/10.1111/rssb.12491.

The reverse-Bayes approach from Held (2020) is based on challenging the original study with a sceptical prior so that there is no longer evidence for an effect. Evidence is quantified in terms of credible intervals, respectively, tail-probabilities. However, there exist also other measures of evidence, and it has been a matter of long debates which is the most appropriate (see e. g., Berger and Sellke, 1987; Casella and Berger, 1987; Royall, 1997; Berger, 2003; Benjamin et al., 2017; Lakens et al., 2018; Amrhein et al., 2019a). In this paper, we therefore extend the reverse-Bayes assessment of replication studies to use Bayes factors (Good, 1958; Jeffreys, 1961) for quantifying evidence and prior-data conflict. Similarly to the sceptical p-value, the procedure leads to a single measure for quantifying the degree of replication success, *the sceptical Bayes factor*. Systematic comparisons show that the sceptical Bayes factor shares most properties with the sceptical p-value due to the reverse-Bayes approach underlying both methods, yet in some situations conclusions may also differ because of their different ways of quantifying evidence. Specifically, it is shown that the sceptical p-values suffers from a certain type of "shrinkage paradox", which is avoided by the sceptical Bayes factor; when the p-value from the original study goes to zero, replication success based on the sceptical p-value can be achieved with any arbitrarily small replication effect estimate, whereas replication success based on the sceptical Bayes factor poses a finite limit on how much shrinkage is allowed. Technically, the procedure is more involved and closed-form solutions for the sceptical Bayes factor are only available in special situations. The method is illustrated on data from the Social Sciences Replication Project (Camerer et al., 2018), and implemented in an R package.

The idea to use Bayes factors instead of tail probabilities was suggested by Consonni (2019) and Pericchi (2020) independently in response to original article by Held (2020). L. Held then implemented a first version of the procedure for the grant application of this research project (Held, 2019b). I then worked out the technical and implementation details, including closed-form solution for the sceptical Bayes factors, asymptotic properties, type-I and type-II error rates, and non-normal extensions of the procedure. I wrote the initial draft of the manuscript and the R package. Throughout, L. Held gave high-level feedback. I presented initial results at the GMDS and CEN-IBS conference in 2020 (online), L. Held presented the final results at the ISBA world meeting 2021 (online).

Appendix II: The assessment of replication success based on relative effect size
by Leonhard Held, Charlotte Micheloud, Samuel Pawel
The Annals of Applied Statistics, 2022, 16(2), 706–720.
https://doi.org/10.1214/21-AOAS1502.

It is not clear how to numerically interpret the sceptical p-value, as it is not an ordinary p-value (which has a uniform distribution under the corresponding null hypothesis). Similarly, it is unclear which threshold should be used in case the sceptical p-value needs to be dichotomized into replication success/failure. In this article, we therefore look closer at the "success region" of the sceptical p-value in terms of the relative effect estimate $d = \hat{\theta}_r/\hat{\theta}_o$. This perspective leads to the proposal of a new default level for thresholding the sceptical p-value called the *golden level* α_G (because the golden ration appears in its derivation). The golden level is defined through the property that for an original study which was just borderline significant ($p_o = \alpha$), replication success based on $p_S \leq \alpha_G$ is only possible if the replication effect estimate is at least as large as the original one ($d \geq 1$). The behavior of the golden level seems to align with common sense; For original studies which were already convincing (in terms of their p-value) the effect estimate in the replication study is allowed to shrink, to some extent, whereas for less convincing original studies (those with p-values around the significance level) shrinkage is more strongly penalized. We find that in typical situations, replication success based on the golden level also has similar or improved frequentist properties (type-I error rate and project power) compared to the standard significance criterion. Case studies from four large-scale replication projects illustrate the properties of the method.

L. Held had the idea to apply the sceptical p-value method to the data from the four replication projects, which I collected for my master thesis. C. Micheloud and L. Held came up with the golden level. L. Held wrote an initial draft of the manuscript. C. Micheloud, L. Held, and I then iteratively worked on the manuscript.

Appendix III: Bayesian approaches to designing replication studies
by Samuel Pawel, Guido Consonni, Leonhard Held
2022. arXiv preprint. https://doi.org/10.48550/arXiv.2211.02552.

An important aspect in the design of replication studies is determining their sample size. How exactly the sample size should be determined depends on the method which will be used for analyzing the replication data. Various approaches have been proposed for doing so which are specifically tailored to certain analysis methods. In this article, we provide a general Bayesian framework which applies to any analysis method (including the sceptical p-value and the sceptical Bayes factor). We show how the data from the original study and external knowledge can be combined in a *design prior* for the underlying model parameters. Based on a design prior, predictions about the replication data can then be computed, and the replication sample size can be chosen such the probability of replication success becomes as high as desired. We illustrate Bayesian design of replication studies in the normal-normal hierarchical model which provides sufficient flexibility for specification of design priors. Data from a cross-laboratory replication project are used for illustrating our methods, which are also available in an R package.

L. Held specified in the grant application of this research project (Held, 2019b) that we will investigate power and sample size calculations for the sceptical p-value and the sceptical Bayes factor. For the first paper I already derived the power function of the sceptical Bayes factor in closed-form for two types of design priors. After its completion, I generalized the result to any design prior in the normal-normal hierarchical model, and started working on this manuscript. I presented a first draft to L. Held and G. Consonni in the beginning of 2021. G. Consonni then helped developing the methodology for multisite replication study design. I continued working on the manuscript in 2022 and also wrote the accompanying R package. Throughout, L. Held and G. Consonni gave high-level feedback.

Appendix IV: Reverse-Bayes methods for evidence assessment and research synthesis

by Leonhard Held, Robert Matthews, Manuela Ott, Samuel Pawel
Research Synthesis Methods, 2022, 13(3), 295–314.
https://doi.org/10.1002/jrsm.1538.

While the popularity of Bayesian methods has been rapidly increasing since the advent of modern computational methods in the 1990s, reverse-Bayes methods have remained largely unknown to statisticians and users of statistics alike. In this article, we review reverse-Bayes history and methods to increase awareness about the approach. Specifically, we summarize the work on reverse-Bayes by I. J. Good (Good, 1950), who first proposed the idea. We then review methods such as the *Analysis of Credibility* from Matthews (2001, 2018), its extension to Bayes factors, and the *False Positive Risk* from Colquhoun (2017). To illustrate these method, we use data from a meta-analysis on the effect of corticosteroids on COVID-19 mortality.

L. Held and M. Ott started working on this article several years ago when M. Ott was still a PhD student (around 2017). When I discovered the connection between the Analysis of Credibility and the fail-safe N method, L. Held suggested that it would fit nicely into this manuscript, and that I should start to overhaul it. I rewrote and expanded his initial draft, adding also a new section on reverse-Bayes approaches with Bayes factors, largely based on the work from paper I. We then managed to recruit R. Matthews to also contribute. From that point on the three of us iteratively worked on the manuscript and M. Ott gave high-level feedback.

Appendix V: Comment on "Bayesian additional evidence for decision making under small sample uncertainty"

by Samuel Pawel, Leonhard Held, Robert Matthews
BMC Medical Research Methodology, 2022, 22(149).
https://doi.org/10.1186/s12874-022-01635-4.

Shortly after the acceptance of paper III, the article by Sondhi et al. (2021) appeared. It proposed a novel reverse-Bayes method called *Bayesian Additional Evidence*, and we noted some flaws in the article. This prompted us to write a commentary. We show that –contrary to the statement by Sondhi et al. (2021)– there is a closed form solution for the key quantity in their approach termed "Bayesian Additional Evidence tipping point". The method is also closely

related the Analysis of Credibility by Matthews (2018). We investigate differences and similarities of the two methods, showing that the priors determined through the Bayesian Additional Evidence method are not always helpful.

R. Matthews attended us about the article from Sondhi et al. (2021). After reading it, I realized that their statement about closed-form solutions was incorrect and derived a solution. L. Held suggested to write a commentary. I wrote an initial draft of the manuscript, which R. Matthews substantially improved. The two of us iteratively worked on the manuscript, while L. Held gave mostly high-level feedback.

Appendix VI: Pitfalls and Potentials in Simulation Studies

by Samuel Pawel, Lucas Kook, Kelly Reeve
2022. arXiv preprint. https://doi.org/10.48550/arXiv.2203.13076.

Simulation studies are frequently used for evaluating statistical methods. However, several studies showed that the reporting standards in simulation studies have remained low over the years (Hoaglin and Andrews, 1975; Burton et al., 2006; Morris et al., 2019). Moreover, some authors have recently argued that also methodological research is suffering from reproducibility issues, publication bias, and a "replication crisis" due to researchers engaging in questionable research practices, such as selective reporting (Boulesteix et al., 2020). In this article, we raise awareness about these issues. We summarize possible questionable research practices in simulation studies, and show how easy it is make a method seem superior if various questionable research practices are employed. We also give recommendations which could help to alleviate these issues, most importantly we recommend researchers to write and pre-register simulation protocols.

The manuscript is co-first authored by myself and L. Kook. I had the idea to invent a "mock-method" and use questionable research practices that make it seem superior, to draw attention to the low standards in methodological research. I then wrote a first draft of the manuscript and proposed the idea to L. Kook and K. Reeve. L. Kook and myself then came up with the method "AINET" and started writing the simulation protocol. L. Kook took lead in developing the R package and simulation study code. K. Reeve provided feedback on the simulation protocol and helped polishing the manuscript. Recently, I was invited to present the results from this project at the CEN conference 2023 in Basel.

Bibliography

(2022). Replication studies hold the key to generalization [editorial]. *Nature Communications*, 13(1). doi:10.1038/s41467-022-34748-x.

Achenbach, J. and McGinley, L. (2017). Researchers struggle to replicate 5 influential cancer experiments from top labs. *The Washington Post*. URL https://www.washingtonpost.com/news/speaking-of-science/wp/2017/01/18/researchers-struggle-to-replicate-5-influential-cancer-experiments-from-top-labs/.

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

Altman, D. G. (1994). The scandal of poor medical research. *BMJ*, 308(6924):283–284. doi:10.1136/bmj.308.6924.283.

Altman, D. G. and Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *BMJ*, 311(7003):485–485. doi:10.1136/bmj.311.7003.485.

Amrhein, V., Greenland, S., and McShane, B. (2019a). Scientists rise up against statistical significance. *Nature*, 567(7748):305–307. doi:10.1038/d41586-019-00857-9.

Amrhein, V., Trafimow, D., and Greenland, S. (2019b). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician*, 73(sup1):262–270. doi:10.1080/00031305.2018.1543137.

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. doi:10.1037/met0000051.

Balafoutas, L. and Sutter, M. (2012). Affirmative action policies promote women and do not harm efficiency in the laboratory. *Science*, 335(6068):579–582. doi:10.1126/science.1211180.

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., Berger, J. O., Forte, A., and García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, 40(3):1550–1577. doi:10.1214/12-aos1013.

Bayarri, M. J. and Mayoral, A. M. (2002b). Bayesian design of "successful" replications. *The American Statistician*, 56:207–214. doi:10.1198/000313002155.

Begley, C. G. and Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533. doi:10.1038/483531a.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., et al. (2017). Redefine statistical significance. *Nature Human Behaviour*, 2(1):6–10. doi:10.1038/s41562-017-0189-z.

Berger, J. (2001). Discussion of "Why should clinicians care about Bayesian methods?" by Robert A.J. Matthews. *Journal of Statistical Planning and Inference*, 94(1):65–67. doi:10.1016/s0378-3758(00)00235-4.

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, 18(1). doi:10.1214/ss/1056397485.

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.

Bonett, D. G. (2020). Design and analysis of replication studies. *Organizational Research Methods*, 24(3):513–529. doi:10.1177/1094428120911088.

Boulesteix, A.-L., Hoffmann, S., Charlton, A., and Seibold, H. (2020). A replication crisis in methodological research? *Significance*, 17(5):18–21. doi:10.1111/1740-9713.01444.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *Journal of the Royal Statistical Society, Series A*, 143:383–430.

Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292. doi:10.1002/sim.2673.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., et al. (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., et al. (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.

Carey, B. (2015). Many psychology findings not as strong as claimed, study says. *The New York Times*. URL `https://www.nytimes.com/2015/08/28/science/many-social-science-findings-not-as-strong-as-claimed-study-says.html`.

Casella, G. and Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association*, 82(397):106–111. doi:10.1080/01621459.1987.10478396.

Chalmers, I., Bracken, M. B., Djulbegovic, B., Garattini, S., Grant, J., Gülmezoglu, A. M., Howells, D. W., Ioannidis, J. P. A., and Oliver, S. (2014). How to increase value and reduce waste when research priorities are set. *The Lancet*, 383(9912):156–165. doi:10.1016/s0140-6736(13)62229-1.

Chambers, C. D. and Tzavella, L. (2021). The past, present and future of registered reports. *Nature Human Behaviour*, 6(1):29–42. doi:10.1038/s41562-021-01193-7.

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.

Consonni, G. and La Rocca, L. (2021). The sceptic and the advocate: comparing two opinions on the mean of a normal distribution. Unpublished notes.

Cooper, H., Hedges, L. V., and Valentine, J. C., editors (2019). *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation. doi:10.7758/9781610448864.

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., Berniūnas, R., Boudesseul, J., Colombo, M., et al. (2018). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*. doi:10.1007/s13164-018-0400-9.

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., and Swinton, J. (2019). *xtable: Export Tables to LaTeX or HTML*. URL `https://CRAN.R-project.org/package=xtable`. R package version 1.8-4.

Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610. doi:10.1080/01621459.1982.10477856.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2012). Joint specification of model space and parameter space prior distributions. *Statistical Science*, 27(2). doi:10.1214/11-sts369.

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.

Devlin, H. (2018). Attempt to replicate major social scientific findings of past decade fails. *The Guardian*. URL `https://www.theguardian.com/science/2018/aug/27/attempt-to-replicate-major-social-scientific-findings-of-past-decade-fails`.

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.

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., and Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, 10. doi:10.7554/elife.71601.

Etz, A. and Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLOS ONE*, 11(2):e0149794. doi:10.1371/journal.pone.0149794.

Evans, M. and Moshonov, H. (2006). Checking for prior-data conflict. *Bayesian Analysis*, 1(4):893–914. doi:10.1214/06-ba129.

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

Freedman, L. P., Cockburn, I. M., and Simcoe, T. S. (2015). The economics of reproducibility in preclinical research. *PLOS Biology*, 13(6):e1002165. doi:10.1371/journal.pbio.1002165.

Gehlenborg, N. (2019). *UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets*. URL `https://CRAN.R-project.org/package=UpSetR`. R package version 1.4.0.

Gilbert, D. T., King, G., Pettigrew, S., and Wilson, T. D. (2016). Comment on "estimating the reproducibility of psychological science". *Science*, 351(6277):1037–1040. doi:10.1126/science.aad7243.

Glasziou, P. and Chalmers, I. (2018). Research waste is still a scandal—an essay by Paul Glasziou and Iain Chalmers. *BMJ*, page k4645. doi:10.1136/bmj.k4645.

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

Good, I. J. (1958). Significance tests in parallel and in series. *Journal of the American Statistical Association*, 53(284):799–813.

Grieve, A. P. (2016). Idle thoughts of a 'well-calibrated' Bayesian in clinical drug development. *Pharmaceutical Statistics*, 15(2):96–108. doi:10.1002/pst.1736.

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). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychological Methods*, 24(5):557–570. doi:10.1037/met0000189.

Held, L. (2019a). 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. (2019b). Research plan "reverse-Bayes design and analysis of replication studies".

Held, L. (2020). 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., Matthews, R., Ott, M., and Pawel, S. (2022a). Reverse-Bayes methods for evidence assessment and research synthesis. *Research Synthesis Methods*. doi:10.1002/jrsm.1538.

Held, L., Micheloud, C., and Pawel, S. (2022b). The assessment of replication success based on relative effect size. URL `https://www.e-publications.org/ims/submission/AOAS/user/submissionFile/47896?confirm=532335fe`. to appear in *The Annals of Applied Statistics*.

Hoaglin, D. C. and Andrews, D. F. (1975). The reporting of computation-based results in statistics. *The American Statistician*, 29(3):122–126. doi:10.1080/00031305.1975.10477393.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8):e124. doi:10.1371/journal.pmed.0020124.

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.

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5):524–532. doi:10.1177/0956797611430953.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions, Vol. 2*. Wiley.

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., and Mandal, S. (2016). On the reproducibility of psychological science. *Journal of the American Statistical Association*, 112(517):1–10. doi:10.1080/01621459.2016.1240079.

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.

Kay, R. (2015). *Statistical Thinking for Non-Statisticians in Drug Regulation*. John Wiley & Sons, Chichester, U.K., second edition. doi:10.1002/9781118451885.

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S., and Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLOS Biology*, 14(5):e1002456. doi:10.1371/journal.pbio.1002456.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, v., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., et al. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, 45:142–152. doi:10.1027/1864-9335/a000178.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Reginald B. Adams, J., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., et al. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490. doi:10.1177/2515245918810225.

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.

Kovic, M. (2016). Die Wissenschaft in der Replikationskrise. *Neue Zürcher Zeitung*. URL `https://www.nzz.ch/wissenschaft/physik/fallstricke-der-statistik-die-wissenschaft-in-der-replikationskrise-ld.86330`.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., et al. (2018). Justify your alpha. *Nature Human Behaviour*, 2(3):168–171. doi:10.1038/s41562-018-0311-x.

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.

Ly, A. and Wagenmakers, E.-J. (2021). Bayes factors for peri-null hypotheses. doi:10.48550/arXiv.2102.07162.

Marshall, R. J. (1988). Bayesian analysis of case-control studies. *Statistics in Medicine*, 7(12):1223–1230. doi:10.1002/sim.4780071203.

Mathur, M. B. and VanderWeele, T. J. (2020). New statistical metrics for multisite replication projects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(3):1145–1166. doi:10.1111/rssa.12572.

Matthews, R. A. J. (2001). 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. (2022). Power calculations for replication studies. *Statistical Science*, 37(3):369–379. doi:10.1214/21-sts828.

Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102. doi:10.1002/sim.8086.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Sert, N. P., Wagenmakers, E.-J., Ware, J. J., and Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(0021). doi:10.1038/s41562-016-0021.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., and Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606. doi:10.1073/pnas.1708274114.

NSL (2018). Achieving new insights through replicability and reproducibility. URL `https://www.nsf.gov/pubs/2018/nsf18053/nsf18053.jsp`.

NWO (2016). Make replication studies a normal part of science. URL `https://www.nwo.nl/en/researchprogrammes/replication-studies`.

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.

Protzko, J., Krosnick, J., Nelson, L. D., Nosek, B. A., Axt, J., Berent, M., Buttrick, N., DeBell, M., Ebersole, C. R., Lundmark, S., MacInnis, B., O'Donnell, M., Perfecto, H., Pustejovsky, J. E., Roeder, S. S., Walleczek, J., and Schooler, J. (2020). High replicability of newly-discovered social-behavioral findings is achievable. doi:10.31234/osf.io/n2a9x. Preprint.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Rawlinson, C. and Bloom, T. (2019). New preprint server for medical research. *BMJ*, page l2301. doi:10.1136/bmj.l2301.

Royall, R. (1997). *Statistical evidence: a likelihood paradigm*. Chapman & Hall, London New York.

Senn, S. (2008). *Statistical issues in drug development*, volume 69. John Wiley & Sons.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366. doi:10.1177/0956797611417632.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, 26:559–569. doi:10.1177/0956797614567341.

Sondhi, A., Segal, B., Snider, J., Humblet, O., and McCusker, M. (2021). Bayesian additional evidence for decision making under small sample uncertainty. *BMC Medical Research Methodology*, 21(1). doi:10.1186/s12874-021-01432-5.

Spiegelhalter, D. J., Abrams, R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. New York: Wiley.

Spiegelhalter, D. J., Freedman, L. S., and Blackburn, P. R. (1986). Monitoring clinical trials: Conditional or predictive power? *Controlled Clinical Trials*, 7(1):8–17. doi:10.1016/0197-2456(86)90003-6.

van Aert, R. C. M. and van Assen, M. A. L. M. (2017). Bayesian evaluation of effect size after replicating an original study. *PLOS ONE*, 12(4):e0175302. doi:10.1371/journal.pone.0175302.

Verhagen, J. and Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, 143:1457–1475. doi:10.1037/a0036731.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., and van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, 100(3):426–432. doi:10.1037/a0022790.

Wicherts, J. M. (2011). Psychology must learn a lesson from fraud case. *Nature*, 480(7375):7–7. doi:10.1038/480007a.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer International Publishing. doi:10.1007/978-3-319-24277-4.

Wickham, H., François, R., Henry, L., and Müller, K. (2022). *dplyr: A Grammar of Data Manipulation*. URL `https://CRAN.R-project.org/package=dplyr`. R package version 1.0.10.

Xie, Y. (2022). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. URL `https://yihui.org/knitr/`. R package version 1.40.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In Goel, P. and Zellner, A., editors, *Bayesian Inference and Decision techniques: Essays in Honor of Bruno de Finetti*, volume 6 of *Studies in Bayesian Econometrics and Statistics*, pages 233–243. Amsterdam: North-Holland.

PAPER I

The sceptical Bayes factor for the assessment of replication success

Samuel Pawel, Leonhard Held

Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2022, 84(3), 879–911. https://doi.org/10.1111/rssb.12491.

Abstract

Replication studies are increasingly conducted but there is no established statistical criterion for replication success. We propose a novel approach combining reverse-Bayes analysis with Bayesian hypothesis testing: a sceptical prior is determined for the effect size such that the original finding is no longer convincing in terms of a Bayes factor. This prior is then contrasted to an advocacy prior (the reference posterior of the effect size based on the original study), and replication success is declared if the replication data favour the advocacy over the sceptical prior at a higher level than the original data favoured the sceptical prior over the null hypothesis. The sceptical Bayes factor is the highest level where replication success can be declared. A comparison to existing methods reveals that the sceptical Bayes factor combines several notions of replicability: it ensures that both studies show sufficient evidence against the null and penalises incompatibility of their effect estimates. Analysis of asymptotic properties and error rates, as well as case studies from the Social Sciences Replication Project show the advantages of the method for the assessment of replicability.

Key words: Bayes factor, Bayesian hypothesis testing, replication studies, reverse-Bayes, sceptical *p*-value

1 Introduction

As a consequence of the so-called replication crisis, the scientific community increasingly recognises the value of replication studies, and several attempts have been made to assess replicability on a large scale (Errington et al., 2014; Klein et al., 2014; Open Science Collaboration, 2015; Camerer et al., 2016, 2018; Cova et al., 2018). Despite most researchers agreeing on the importance of replication, there is currently no agreement on a statistical criterion for replication success. Instead, a variety of statistical methods, frequentist (Simonsohn, 2015; Patil et al., 2016; Hedges and Schauer, 2019; Mathur and VanderWeele, 2020), Bayesian (Bayarri and Mayoral, 2002a,b; Verhagen and Wagenmakers, 2014; Johnson et al., 2016; Etz and Vandekerckhove, 2016; van Aert and van Assen, 2017; Ly et al., 2018; Harms, 2019), and combinations thereof (Held, 2020; Pawel and Held, 2020; Held et al., 2022b) have been proposed to quantify replication success.

Due to this lack of an established method, replication projects typically report the results of several methods and it is not uncommon for these to contradict each other. For example, both studies may find evidence against a null effect, but the individual effect estimates may still be incompatible (often the replication estimate is much smaller). Conversely, both estimates may be compatible, but there may not be enough evidence against a null effect in one of the studies.

The objective of this paper is to present a novel Bayesian method for quantifying replication success, which builds upon a previously proposed method (the *sceptical p-value* from Held, 2020) and unifies several notions of replicability. The method combines the natural fit of the reverse-Bayes approach to the replication setting with the use of Bayes factors for hypothesis testing (Jeffreys, 1961; Kass and Raftery, 1995) and model criticism (Box, 1980). In a nutshell,

replication success is declared if the replication data favour an advocacy prior for the effect size, which emerges from taking the original result at face value, over a sceptical prior, which renders the original result unconvincing.

Held (2020) proposed a reverse-Bayes approach for the assessment of replication success: The main idea is to challenge the result from an original study by determining a *sceptical prior* for the effect size, sufficiently concentrated around the null value such that the resulting posterior is rendered unconvincing (Matthews, 2001). An unconvincing posterior at level α is defined by its $(1 - \alpha)$ credible interval just including the null value. Subsequently, the replication data are used in a prior-data conflict assessment (Box, 1980; Evans and Moshonov, 2006) and replication success is concluded if there is sufficient conflict between the sceptical prior and the replication data. Specifically, replication success at level α is established if the prior predictive tail probability of the replication estimate is smaller than α. The smallest level α at which replication success can be declared corresponds to the sceptical p-value.

The method comes with appealing properties: The sceptical p-value is never smaller than the ordinary p-values from both studies, thus ensuring that they both provide evidence against the null. At the same time, it also takes into account the size of their effect estimates, penalising the case when the replication estimate is smaller than the original estimate. Held et al. (2022b) further refined the method with a recalibration that allows the sceptical p-value to be interpreted on the same scale as an ordinary p-value, as well as ensuring appropriate frequentist properties, such as type I error rate control if the replication sample size is not smaller than in the original study.

Despite the methods' Bayesian nature, it relies on tail probabilities as primary inference tool. An alternative is the Bayes factor, the principled Bayesian solution to hypothesis testing and model selection (Jeffreys, 1961; Kass and Raftery, 1995). In contrast to tail probabilities, Bayes factors have a more natural interpretation and allow for direct quantification of evidence for one hypothesis versus another. In this paper we therefore extend the reverse-Bayes procedure from Held (2020) to use Bayes factors for the purpose of quantifying evidence. This extension was suggested by Consonni (2019) and Pericchi (2020) independently. Interestingly, a similar extension of the reverse-Bayes method from Matthews (2001) was already hinted at by Berger (2001), but to date no one has attempted to realise the idea.

The inclusion of Bayes factors leads 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 combines two notions of replication success: It requires from both studies to show sufficient evidence against the null, while also 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 reference posterior of the effect conditional on the original study). This direct assessment of compatibility allows for stronger statements about the degree of replication success, and it may also lead to different conclusions in certain situations.

This paper is structured as follows: Section 2 presents the derivation of the sceptical Bayes factor. Its asymptotic and finite sample properties are then compared with other measures of

replication success in Section 3. An extension to non-normal models is presented in Section 4. Section 5 illustrates how the method works in practice using case studies from the *Social Sciences Replication Project* (Camerer et al., 2018). Section 6 provides concluding remarks about strengths, limitations and extensions of the method.

Notation and assumptions

Denote the Bayes factor comparing the plausibility of hypotheses H_1 and H_2 with respect to the observed data x by

$$\mathrm{BF}_{1:2}(x) = \frac{f(x \mid H_1)}{f(x \mid H_2)} = \frac{\int_{\Theta_1} f(x \mid \theta_1) f(\theta_1) \, \mathrm{d}\theta_1}{\int_{\Theta_2} f(x \mid \theta_2) f(\theta_2) \, \mathrm{d}\theta_2},$$

where $f(x \mid H_i)$ is the marginal likelihood of the data under H_i obtained by integrating the likelihood $f(x \mid \theta_i)$ with respect to the prior distribution $f(\theta_i)$ of the model parameters $\theta_i \in \Theta_i$ with $i = 1, 2$. Sometimes we will also write $\mathrm{BF}_{1:2}(x; \phi')$ to indicate that the Bayes factor is evaluated for a specific value ϕ' of a hyperparameter ϕ of one of the model priors. To simplify comparison with p-values we will orient Bayes factors such that lower values indicate more evidence against a null hypothesis.

Let θ denote the effect of a treatment on an outcome of interest. Let $\hat{\theta}_o$ and $\hat{\theta}_r$ denote its maximum likelihood estimates obtained from an original (subscript o) and from a replication study (subscript r), respectively. Let the corresponding standard errors be denoted by σ_o and σ_r, the z-values by $z_o = \hat{\theta}_o / \sigma_o$ and $z_r = \hat{\theta}_r / \sigma_r$, and define the variance ratio as $c = \sigma_o^2 / \sigma_r^2$ and the relative effect estimate as $d = \hat{\theta}_r / \hat{\theta}_o = z_r / (z_o \sqrt{c})$. For many effect size types the variances are inversely proportional to the sample size, i. e., $\sigma_o^2 = \kappa / n_o$ and $\sigma_r^2 = \kappa / n_r$ for some unit variance κ. The variance ratio is then the ratio of the replication to the original sample size $c = n_r / n_o$.

We adopt a meta-analytic framework and consider the effect estimates as the data, rather than their underlying samples, and assume that $\hat{\theta}_k \mid \theta \sim \mathrm{N}(\theta, \sigma_k^2)$ for $k \in \{o, r\}$, i. e., normality of the effect estimates around θ, with known variances equal to their squared standard errors. For studies with reasonable sample size, this framework usually provides a good approximation for a wide range of (suitably transformed) effect size types (Spiegelhalter et al., 2004, Chapter 2.4). For example, means and mean differences (no transformation), odds ratios, hazard ratios, risk ratios (logarithmic transformation), or correlation coefficients ("Fisher-z" transformation). We refer to the literature of meta-analysis for details about transformations of effect sizes (e. g., Cooper et al., 2019, Chapter 11.6). The normal model in combination with conjugate priors enables derivation of closed-form expressions in many cases, which allows us to easily study limiting behaviour and facilitates interpretability. In Section 4, we will present relaxations of the normality assumption, which can lead to more accurate inferences when studies have small sample sizes and/or show extreme results.

2 Reverse-Bayes assessment of replication success with Bayes factors

The idea of reversing Bayes' theorem was first proposed by Good (1950). He acknowledged that in many situations there is no obvious choice for the prior distributions involved in Bayesian analyses. On the other hand, we are often more certain which posterior inferences would convince us regarding the credibility of a hypothesis. For this reason, Good inverted Bayes' theorem and derived priors, which combined with the observed data, would lead to posterior inferences that were specified beforehand (e. g., the data favour one hypothesis over another). His reverse-Bayes inference then centred around the question whether the resulting prior is plausible, and if so, this would legitimise the posterior inference. See Figure 1 for a graphical illustration of this process.

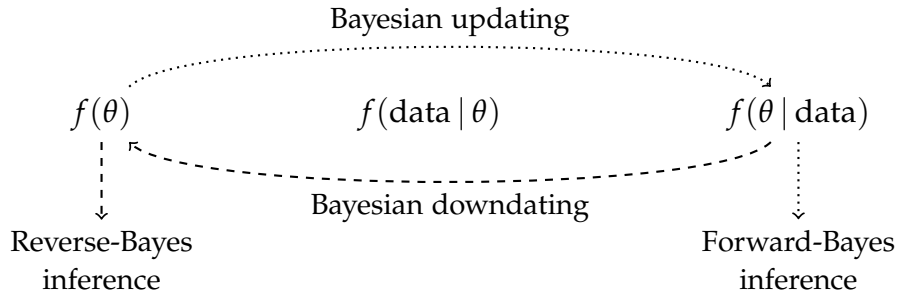


Figure 1: Schematic illustration of reverse-Bayes and forward-Bayes inference for an unknown parameter θ.

Good argued that philosophically there is nothing wrong with inferences resulting from backwards use of Bayes' theorem, since the theorem merely constrains prior and posterior to be consistent with the laws of probability (regardless of their conventional names suggesting a particular temporal ordering). Despite his advocacy, the reverse-Bayes idea remained largely unexplored until Matthews (2001) introduced the *Analysis of Credibility*, which in turn led to new developments in reverse-Bayes methodology (see Held et al. (2022a) for a recent review). Most of these approaches use the reversal of Bayes' theorem in order to challenge or substantiate the credibility of scientific claims. Usually, a posterior inference corresponding to (non-)credibility of a claim is specified, and the associated prior is then derived from the data. Inference is subsequently carried out based on this reverse-Bayes prior, e. g., the interest is often to check whether the prior is plausible in light of external evidence, an obvious candidate being data from a replication study. This can be done, for example, using methods to assess prior-data conflict (Box, 1980; Evans and Moshonov, 2006).

In this paper, we consider a reverse-Bayes procedure consisting of two stages that naturally fit the replication setting: We first determine a *sufficiently sceptical prior* for the effect θ such that the original result is no longer convincing in terms of a suitable 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 θ conditional on the original data and an uninformative/reference prior. Box 1 provides a summary of the procedure, the following sections will explain it in more detail.

1. **Original study**: For the original effect estimate $\hat{\theta}_o \,|\, \theta \sim \mathrm{N}(\theta, \sigma_o^2)$ consider the point null hypothesis $H_0\colon \theta = 0$ vs. $H_\mathrm{S}\colon \theta \neq 0$. Fix a level $\gamma \in (0,1)$ and determine the sufficiently sceptical prior under the alternative $\theta \,|\, H_\mathrm{S} \sim \mathrm{N}(0, g_\gamma \cdot \sigma_o^2)$ such that the Bayes factor contrasting H_0 to H_S is

$$\mathrm{BF}_{0:\mathrm{S}}(\hat{\theta}_o; g_\gamma) = \gamma.$$

   The prior $\theta \,|\, H_\mathrm{S}$ represents a *sceptic* who remains unconvinced about the presence of an effect at level γ.

2. **Replication study**: For the replication effect estimate $\hat{\theta}_r \,|\, \theta \sim \mathrm{N}(\theta, \sigma_r^2)$ compute the Bayes factor contrasting the sceptic $H_\mathrm{S}\colon \theta \sim \mathrm{N}(0, g_\gamma \cdot \sigma_o^2)$ to an advocate $H_\mathrm{A}\colon \theta \sim \mathrm{N}(\hat{\theta}_o, \sigma_o^2)$. Declare *replication success* at level γ if

$$\mathrm{BF}_{\mathrm{S:A}}(\hat{\theta}_r; g_\gamma) \leq \gamma,$$

   i. e., the data favour the advocate over the sceptic at a higher level than the sceptic's initial objection.

→ The *sceptical Bayes factor* BF_S is the smallest level γ at which replication success can be declared.

Box 1: Summary of reverse-Bayes assessment of replication success with Bayes factors.

2.1 Data from the original study

For the effect estimate $\hat{\theta}_o \,|\, \theta \sim \mathrm{N}(\theta, \sigma_o^2)$ from the original study consider a hypothesis test comparing the null hypothesis $H_0\colon\ \theta = 0$ to the alternative $H_\mathrm{S}\colon\ \theta \neq 0$. Specification of a prior distribution for θ under H_S is now required for Bayesian hypothesis testing. A typical choice (Jeffreys, 1961) is a local alternative, a unimodal symmetric prior distribution centred around the null value. We consider the sceptical prior $\theta \,|\, H_\mathrm{S} \sim \mathrm{N}(0, \sigma_s^2 = g \cdot \sigma_o^2)$ with relative sceptical prior variance g for this purpose (relative to the variance from the original estimate $g = \sigma_s^2/\sigma_o^2$), resembling the g-prior known from the regression literature (Zellner, 1986; Liang et al., 2008). The explicit form of the Bayes factor is then given by

$$\mathrm{BF}_{0:\mathrm{S}}(\hat{\theta}_o; g) = \sqrt{1+g} \cdot \exp\left\{-\frac{1}{2} \cdot \frac{g}{1+g} \cdot z_o^2\right\}. \tag{1}$$

The Bayes factor from equation (1) is shown in Figure 2 as a function of g and for different original z-values z_o. For fixed z_o, it is well known that this Bayes factor is bounded from below by

$$\mathrm{minBF}_o = \begin{cases} |z_o| \cdot \exp(-z_o^2/2) \cdot \sqrt{e} & \text{for } |z_o| > 1 \\ 1 & \text{for } |z_o| \leq 1 \end{cases} \tag{2}$$

which is reached at $g_{\text{minBF}_o} = \max\{0, z_o^2 - 1\}$ (Edwards et al., 1963). Further increasing the relative sceptical prior variance increases (1) indefinitely because of the Jeffreys-Lindley paradox, i. e., $\text{BF}_{0:\text{S}}(\hat{\theta}_o; g) \to \infty$ for $g \to \infty$ (Bernardo and Smith, 2000, Section 6.1.4). Hence, for a relative sceptical prior variance $g \in [0, g_{\text{minBF}_o}]$, the resulting Bayes factor will be $\text{BF}_{0:\text{S}}(\hat{\theta}_o; g) \in [\text{minBF}_o, 1]$.

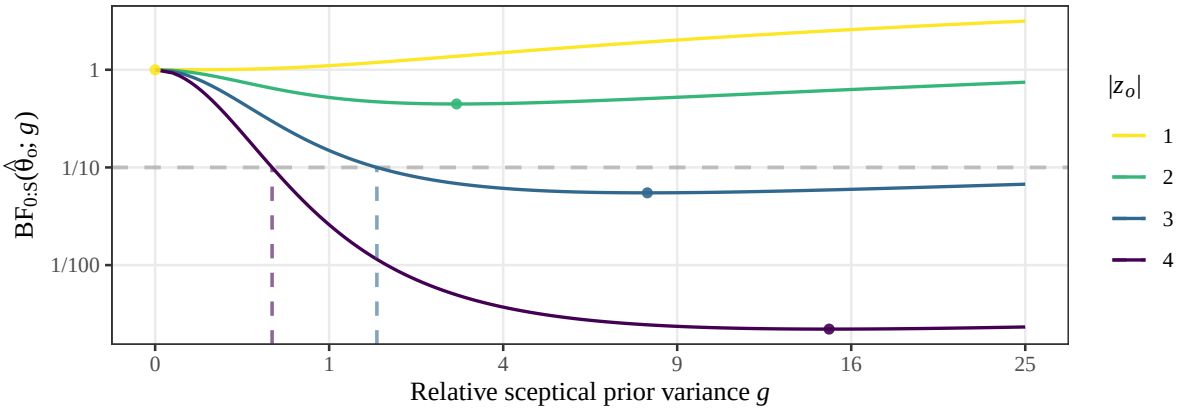


Figure 2: Bayes factor $\text{BF}_{0:\text{S}}(\hat{\theta}_o; g)$ as a function of relative sceptical prior variance g for different values of $|z_o| = |\hat{\theta}_o|/\sigma_o$. Minimum Bayes factors minBF_o are indicated by dots. Dashed vertical lines indicate sufficiently sceptical relative prior variance g_γ at level $\gamma = 1/10$, if they exist.

We now apply the reverse-Bayes idea and challenge the original finding. To do so, we fix a level γ above which the original finding is no longer convincing to us. For example, this could be $\gamma = 1/10$; the threshold for strong evidence against H_0 according to the classification from Jeffreys (1961). Suppose now there exists a $g_\gamma \leq g_{\text{minBF}_o}$ such that $\text{BF}_{0:\text{S}}(\hat{\theta}_o; g_\gamma) = \gamma$. It can be shown (Appendix A) that g_γ can be explicitly computed by

$$g_\gamma = \begin{cases} -\dfrac{z_o^2}{q} - 1 & \text{if } -\dfrac{z_o^2}{q} \geq 1 \\ \text{undefined} & \text{else} \end{cases} \tag{3}$$

$$\text{where } q = W_{-1}\left(-\frac{z_o^2}{\gamma^2} \cdot \exp\left\{-z_o^2\right\}\right)$$

with $W_{-1}(\cdot)$ the branch of the Lambert W function (Corless et al., 1996) that satisfies $W(y) \leq -1$ for $y \in [-e^{-1}, 0)$, see Appendix B for details about the Lambert W function. The sufficiently sceptical prior is then given by $\theta \,|\, H_\text{S} \sim \text{N}(0, g_\gamma \cdot \sigma_o^2)$ and it can be interpreted as the view of a sceptic who argues that given their prior belief about the effect θ, the observed effect estimate $\hat{\theta}_o$ cannot convince them about the presence of a non-null effect at level γ. An alternative data-based interpretation of sufficiently sceptical priors is to see them as the priors obtained by updating an initial uniform prior with the data from an imaginary study, which was $1/g_\gamma$ times the size of the original study, and which resulted in an effect estimate of exactly zero (Held et al., 2022a).

From Figure 2 we can see that the more compelling the original data (i. e., the larger $|z_o|$), the smaller the sufficiently sceptical relative prior variance g_γ needs to be in order to make

the result no longer convincing at level γ. In the most extreme case, when $|z_o| \to \infty$ and γ remains fixed, the sufficiently sceptical prior variance will converge to zero (Appendix B). On the other hand, if $|z_o|$ is not sufficiently large, $\text{BF}_{0:\text{S}}(\hat{\theta}_o; g)$ will either be always increasing in g (if $|z_o| \leq 1$) or it will reach a minimum above the chosen level γ. In both cases the sufficiently sceptical relative prior variance g_γ is not defined since there is no need to challenge an already unconvincing result.

A side note on the Jeffreys-Lindley paradox is worth being mentioned: If a $g_\gamma < g_{\text{minBF}_o}$ exists, there exists also a $g'_\gamma > g_{\text{minBF}_o}$ as the Bayes factor monotonically increases in $g > g_{\text{minBF}_o}$ and therefore must intersect a second time with γ, due to the paradox. This means that the more compelling the original result, the larger g'_γ needs to be chosen, such that the result becomes no longer convincing at level γ. However, priors which become increasingly diffuse do not represent increasing scepticism but rather increasing ignorance. Using (3) therefore avoids this manifestation of the Jeffreys-Lindley paradox, since it determines sceptical priors only from the class of priors that become increasingly concentrated for increasing evidence (i. e., priors with $g_\gamma \leq g_{\text{minBF}_o}$). In principle, the solution $g'_\gamma > g_{\text{minBF}_o}$ could also be computed by replacing the W_{-1} branch of the Lambert W function in (3) with the W_0 branch, but this will not be of interest to us.

2.2 Data from the replication study

In order to assess whether the original finding can be replicated in an independent study, a replication study is conducted, leading to a new effect estimate $\hat{\theta}_r$. In light of the new data, the sceptic is now challenged by an advocate of the original finding. This is formalised with another Bayes factor, which compares the plausibility of the replication effect estimate $\hat{\theta}_r \,|\, \theta \sim \text{N}(\theta, \sigma_r^2)$ under the sceptical prior $H_\text{S}\colon\ \theta \sim \text{N}(0, g \cdot \sigma_o^2)$ relative to the advocacy prior $H_\text{A}\colon\ \theta \sim \text{N}(\hat{\theta}_o, \sigma_o^2)$. The the view of an advocate is represented by H_A since this is the posterior of θ given the original estimate and a uniform prior (also the reference prior for this model). The Bayes factor is given by

$$\text{BF}_{\text{S:A}}(\hat{\theta}_r; g) = \sqrt{\frac{1/c+1}{1/c+g}} \cdot \exp\left\{-\frac{z_o^2}{2}\left(\frac{d^2}{1/c+g} - \frac{(d-1)^2}{1/c+1}\right)\right\} \tag{4}$$

so it depends on the original z-statistic z_o, the relative sceptical prior variance g, the relative effect estimate $d = \hat{\theta}_r/\hat{\theta}_o$, and the relative variance $c = \sigma_o^2/\sigma_r^2$.

Our goal is now to define a condition for *replication success* in terms of (4). It is natural to consider a replication successful if the replication data favour the advocate over the sceptic to a higher degree than the sceptic's initial objection to the original study. More formally, we say that if the Bayes factor from (4) evaluated at the sufficiently sceptical relative prior variance g_γ is not larger than the corresponding level γ used to define the sufficiently sceptical prior:

$$\text{BF}_{\text{S:A}}(\hat{\theta}_r; g_\gamma) \leq \text{BF}_{0:\text{S}}(\hat{\theta}_o; g_\gamma) = \gamma, \tag{5}$$

we have established *replication success at level* γ.

For example, if we observe $z_o = 3$ (equivalent to minimum Bayes factor $\text{minBF}_o = 1/18$) and choose a level $\gamma = 1/10$ the sufficiently sceptical relative prior variance (3) is $g_\gamma = 1.6$. If a replication is conducted with the same precision ($c = 1$) and we observe $z_r = 2.5$ (equivalent to minimum Bayes factor $\text{minBF}_r = 1/5.5$ and relative effect estimate $d = z_r/(z_o\sqrt{c}) = 0.83$), using equation (4) this would lead to $\text{BF}_{\text{S:A}}(\hat{\theta}_r; 1.6) = 1/3.5$, which means that the replication was not successful at level $\gamma = 1/10$. However, if we had chosen a less stringent level, e. g., $\gamma = 1/3$, the replication would have been considered successful since then $g_\gamma = 0.4$ and $\text{BF}_{\text{S:A}}(\hat{\theta}_r; 0.4) = 1/7.4$.

2.3 The sceptical Bayes factor

Apart from specifying a level γ, the described procedure offers an automated way to assess replication success. One way to remove this dependence is to find the smallest level γ where replication success can be established. We thus call this level the *the sceptical Bayes factor*

$$\text{BF}_\text{S} = \inf\left\{\gamma : \text{BF}_{\text{S:A}}(\hat{\theta}_r; g_\gamma) \leq \gamma\right\}, \tag{6}$$

and replication success at level γ is equivalent with $\text{BF}_\text{S} \leq \gamma$.

Figure 3 shows $\text{BF}_{\text{S:A}}(\hat{\theta}_r; g_\gamma)$ and $\text{BF}_{\text{0:S}}(\hat{\theta}_o; g_\gamma)$ as a function of g_γ for several values of z_o and d along with the corresponding BF_S. Typically, BF_S is given by the height of the intersection between $\text{BF}_{\text{S:A}}(\hat{\theta}_r; g_\gamma)$ and $\text{BF}_{\text{0:S}}(\hat{\theta}_o; g_\gamma)$ in g_γ. It may also happen that $\text{BF}_{\text{S:A}}(\hat{\theta}_r; g_\gamma)$ remains below $\text{BF}_{\text{0:S}}(\hat{\theta}_o; g_\gamma)$ for all values of g_γ, in such situations BF_S is equal to the original minimum Bayes factor minBF_o. Finally, in some pathological cases it may happen that either z_o, d, or both are so small that replication success cannot be established for any level γ and hence BF_S does not exist. This means that the replication study was unsuccessful since it is impossible for the advocate to convince the sceptic at any level of evidence.

In terms of computing the sceptical Bayes factor, it is worth noting that for the special case when the replication is conducted with the same precision as the original study ($c = 1$) and BF_S is located at the intersection of $\text{BF}_{\text{S:A}}(\hat{\theta}_r; g)$ and $\text{BF}_{\text{0:S}}(\hat{\theta}_o; g)$ in g, there is an explicit expression for BF_S

$$\text{BF}_\text{S} = \sqrt{-\frac{z_o^2}{k} \cdot \frac{1+d^2}{2}} \cdot \exp\left\{-\left(\frac{z_o^2}{2} + \frac{k}{1+d^2}\right)\right\} \tag{7}$$

with

$$k = W\left(-\frac{z_o^2}{\sqrt{2}} \cdot \frac{d^2+1}{2} \cdot \exp\left\{-\frac{z_o^2}{2}\left[1 + \frac{(1-d)^2}{2}\right]\right\}\right),$$

see Appendix C for details.

3 Properties

To study properties of the sceptical Bayes factor and facilitate comparison with other methods we will look at the requirements for replication success based on the relative effect estimate

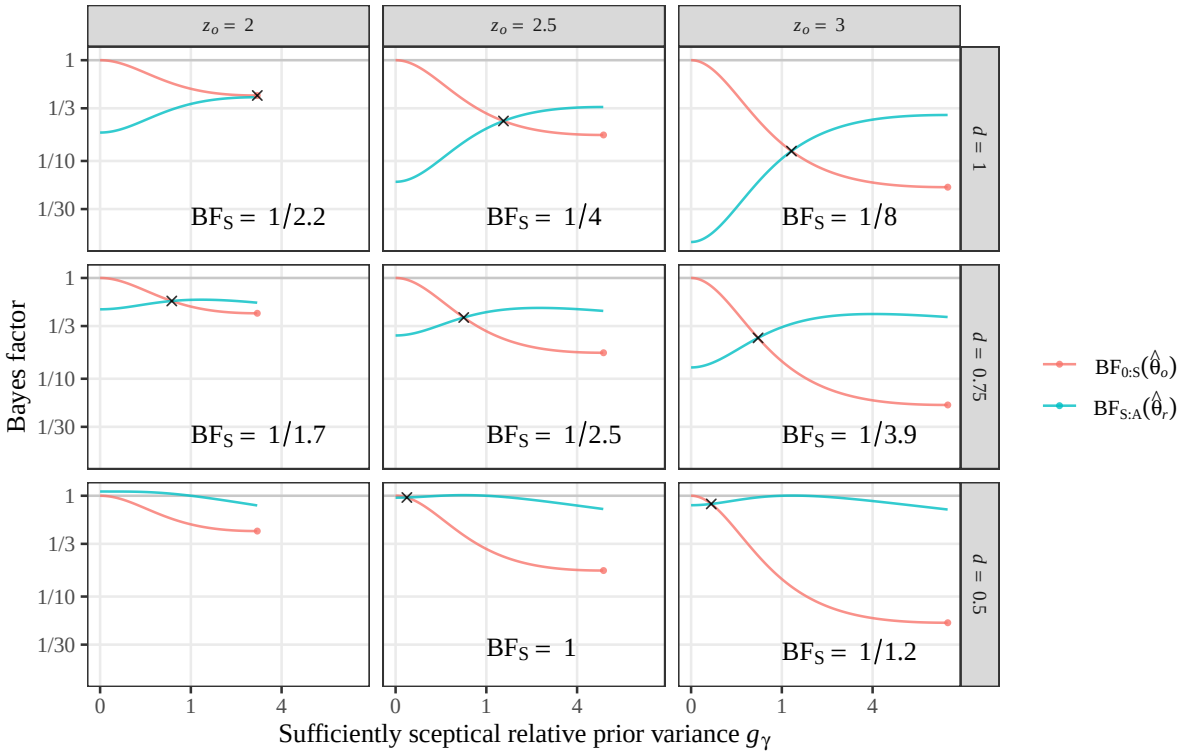


Figure 3: Bayes factors $\text{BF}_{\text{S:A}}(\hat{\theta}_r; g)$ and $\text{BF}_{\text{0:S}}(\hat{\theta}_o; g)$ as a function of the sufficiently sceptical relative prior variance g_γ. In all examples $c = \sigma_o^2/\sigma_r^2 = 1$. Minimum Bayes factors minBF_o are indicated by dots, sceptical Bayes factors BF_S are indicated by crosses where existent.

$d = \hat{\theta}_r/\hat{\theta}_o$, the variance ratio $c = \sigma_o^2/\sigma_r^2$ and the original minimum Bayes factor minBF_o (respectively the original z-value z_o). This perspective is helpful because it disentangles how the method reacts to changes in compatibility of the effect estimates (d), evidence from the original study (minBF_o), and the change in sample size of the replication compared to the original study (c).

The condition for replication success at level γ from (5) is equivalent to

$$\log\left\{\frac{1/c+1}{(1/c+g_\gamma)(1+g_\gamma)}\right\} + \frac{z_o^2}{1+1/g_\gamma} \leq z_o^2\left(\frac{d^2}{1/c+g_\gamma} - \frac{(d-1)^2}{1/c+1}\right). \tag{8}$$

On the right-hand side of (8) the Q-statistic

$$Q = \frac{(\hat{\theta}_o - \hat{\theta}_r)^2}{\sigma_o^2 + \sigma_r^2} = \frac{z_o^2(d-1)^2}{1/c+1} \tag{9}$$

appears. The Q-statistic was proposed as a measure of incompatibility among original and replication effect estimates since its distribution is known for standard meta-analytic models of effect sizes (Hedges and Schauer, 2019). The connection to the sceptical Bayes factor is such that Q acts as a penalty term in (8) and a larger value will lower the degree of replication success possible. However, as we will see, the sceptical Bayes factor goes beyond assessing

effect estimate compatibility as there is also a trade-off with the amount of evidence that the replication study provides against the null.

Applying some algebraic manipulations to (8), one can show that replication success at level γ is achieved if and only if the relative effect estimate d falls within a success region given by

$$\begin{cases} d \notin [M - \sqrt{A/B}, M + \sqrt{A/B}] & \text{if } g_\gamma < 1 \\ d \geq [1 + (1 + 1/c)\{1/2 - \log(2)/z_o^2\}]/2 & \text{if } g_\gamma = 1 \\ d \in [M - \sqrt{A/B}, M + \sqrt{A/B}] & \text{if } g_\gamma > 1 \end{cases} \tag{10}$$

where

$$\begin{aligned} M &= \frac{1/c + g_\gamma}{g_\gamma - 1} \\ A &= \log\left\{\frac{1/c + 1}{(1/c + g_\gamma)(1 + g_\gamma)}\right\} \Big/ z_o^2 + \frac{g_\gamma}{1 + g_\gamma} + \frac{1}{1 - g_\gamma} \\ B &= \frac{1 - g_\gamma}{(1/c + g_\gamma)(1/c + 1)}. \end{aligned}$$

The top-left plot in Figure 4 shows the conditions on d from (10) to achieve replication success at level $\gamma = 1/3$ as a function of the original minimum Bayes factor minBF_o and for different values of the relative variance c. It is important to note that $\gamma = 1/3$ is an arbitrary choice and in practice one should interpret the sceptical Bayes factor as a quantitative measure of replication success. Only the success regions for positive d are shown as replication success in the opposite direction is usually not of interest (see Section 3.2 for a discussion of this issue). We see that with increased precision of the replication study (larger c), the success regions shift closer to zero. This means that the method allows for more shrinkage of the replication effect estimate when the replication provides more evidence against the null (because $|z_r| = d\,|z_o|\,\sqrt{c}$ increases with increasing c). However, the success regions cannot be pushed arbitrarily close to zero but are bounded away. So when $c \to \infty$ the methods still requires the replication estimate to be sufficiently large, despite that the evidence against the null becomes overwhelming (since $|z_r| \to \infty$ as $c \to \infty$).

By definition the sceptical Bayes factor can never be smaller than minBF_o, so replication success at level γ is impossible for original studies with $\text{minBF}_o > \gamma$. This property is visible in the top-left plot in Figure 4 by the cut-off at $\text{minBF}_o = \gamma = 1/3$. In contrast, for more convincing original studies with $\text{minBF}_o < 1/3$ replication success is possible and two cases can be distinguished in terms of the success region: When $1/4.5 < \text{minBF}_o \leq 1/3$ the sufficiently sceptical relative prior variance is $g_\gamma > 1$ and thus by condition (10) the success region consists of an interval $(d_{\min}, d_{\max})$. Hence, in this case the method also penalises too large replication effect estimates. For original studies with $\text{minBF}_o < 1/4.5$, the sufficiently sceptical relative prior variance is $g_\gamma \leq 1$, so due to (10) the success region for positive d is given by $(d_{\min}, d_{\max} = \infty)$. This means that for more convincing original studies there are no upper restrictions for the relative effect estimate, whereas shrinkage of the replication estimate is still penalised.

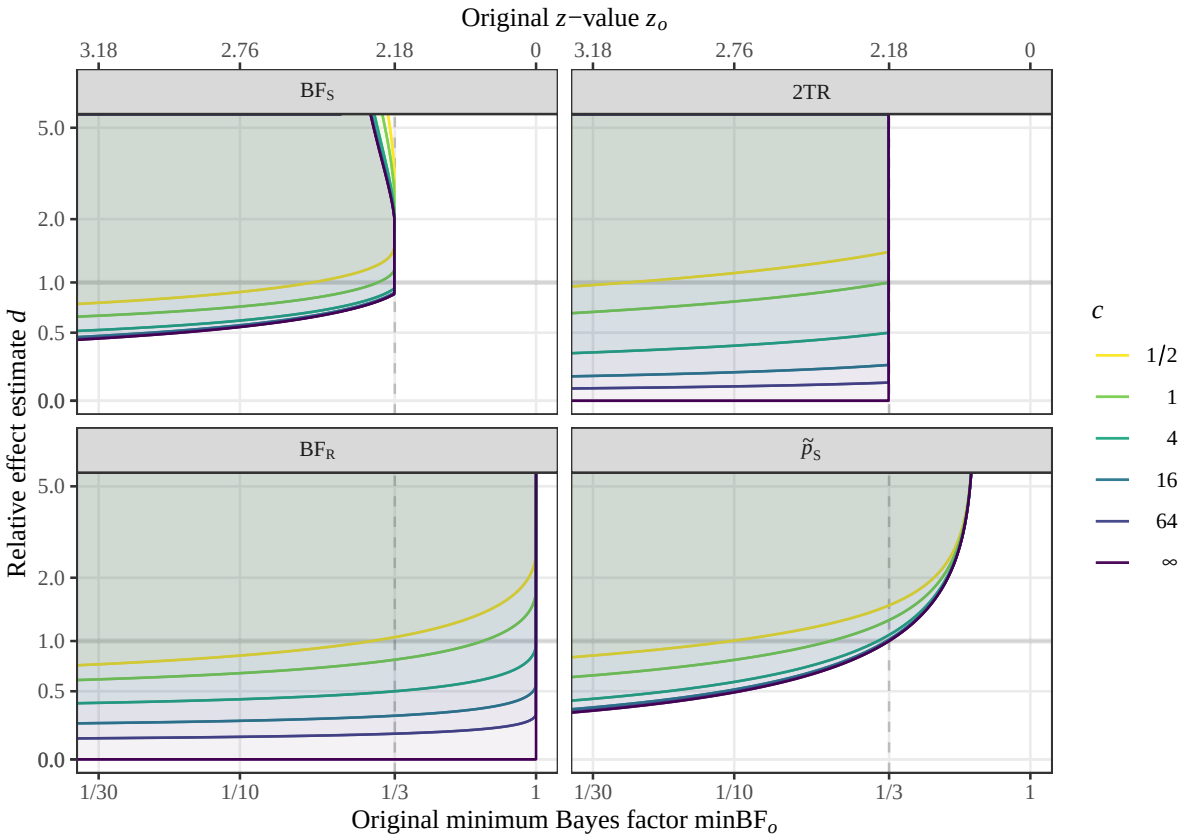


Figure 4: Required relative effect estimate $d = \hat{\theta}_r/\hat{\theta}_o$ to achieve replication success based on the sceptical Bayes factor ($\mathrm{BF_S} \leq 1/3$), the two-trials rule (2TR: $\mathrm{minBF}_o \leq 1/3$ and $\mathrm{minBF}_r \leq 1/3$), the replication Bayes factor ($\mathrm{BF_R} \leq 1/3$), and the recalibrated sceptical p-value ($\tilde{p}_\mathrm{S} \leq 1 - \Phi\{z_\gamma\}$ with $\gamma = 1/3$) as a function of the original minimum Bayes factor minBF_o (respectively the corresponding z-value z_o) for different values of the relative variance $c = \sigma_o^2/\sigma_r^2$. Shading indicates regions where replication success is possible. Only positive relative effect estimates d are shown.

3.1 Comparison with other methods

Of interest is the relationship between the sceptical Bayes factor and other measures of replication success. Here, we review and compare a classical (the two-trials rule), a forward-Bayes (the replication Bayes factor from Verhagen and Wagenmakers, 2014) and a reverse-Bayes method (the sceptical p-value from Held, 2020). These methods provide a useful benchmark as they all are based on hypothesis testing, have unique properties, and can be directly compared in terms of their replication success regions as shown in Figure 4.

The two-trials rule

Replication success is most commonly declared when both original and replication study provide compelling evidence against a null effect. This approach is also known as the *two-trials*

rule in drug development and usually a requirement for drug approval (Kay, 2015, Section 9.4). Most replication projects report p-values associated with the effect estimates as measures of evidence against the null, but also default Bayes factors have been used (see e. g., the Bayesian supplement of Camerer et al., 2018). To compare the two-trials rule with methods based on Bayes factors we will study the two-trials rule based on the minimum Bayes factor from (2), i. e., replication success at level γ is established when both $\text{minBF}_k \leq \gamma$ for $k \in \{o, r\}$, as well as $\text{sign}(\hat{\theta}_o) = \text{sign}(\hat{\theta}_r)$. This approach has a one-to-one correspondence to the usual version of two-trials rule as minimum Bayes factors and p-values both only depend on the z-values of original and replication study.

The two-trials rule guarantees that both studies provide compelling evidence against the null. Similarly, the sceptical Bayes factor requires the original study to be compelling on its own since it can never be smaller than minBF_o. However, one can easily construct examples where the sceptical Bayes factor is smaller than the minimum Bayes factor from the replication study (e. g., when $\text{minBF}_o = 1/2$, $\text{minBF}_r = 1/1.5$, and $c = 1$ we obtain $\text{BF}_\text{S} = 1/1.9$). So for the same level of replication success γ the two-trials may not flag replication success whereas the sceptical Bayes factor would.

By definition the two-trials rule can never be fulfilled when the original study was uncompelling. Assuming now that $\text{minBF}_o < \gamma$, replication success with the two-trials rule at level γ is achieved if and only if the relative effect estimate is

$$d \geq \frac{z_\gamma}{z_o \sqrt{c}} \tag{11}$$

with $z_\gamma > 1$ corresponding to $\text{minBF} = z_\gamma \exp(-z_\gamma^2/2)\sqrt{e} = \gamma$. The success region from (11) is displayed in the top-right plot of Figure 4. We see that the success regions shift closer to zero as the relative variance c increases. Also there is a cut-off at $\text{minBF}_o = \gamma = 1/3$ similarly as with the sceptical Bayes factor. In contrast to the sceptical Bayes factor, however, the two-trials can be fulfilled for any arbitrary small (but positive) relative effect estimate d, provided the relative variance c is large enough. Hence, the two-trials rule may flag success even when the replication effect estimate is much smaller than the original one.

The replication Bayes factor

Verhagen and Wagenmakers (2014) proposed the *replication Bayes factor* as a measure of replication success. It is defined as the Bayes factor comparing the point null hypothesis $H_0 \colon \theta = 0$, to the alternative that the effect is distributed according to the posterior distribution of θ after observing the original data. For the normal model considered so far and if an initial reference prior was chosen, this alternative is also the advocacy prior $H_\text{A} \colon \theta \sim \text{N}(\hat{\theta}_o, \sigma_o^2)$ and therefore the replication Bayes factor is given by

$$\text{BF}_\text{R} = \text{BF}_{\text{0:A}}(\hat{\theta}_r) = \sqrt{1+c} \cdot \exp\left\{ -\frac{z_o^2}{2} \left(d^2 \cdot c - \frac{(1-d)^2}{1/c + 1} \right) \right\}. \tag{12}$$

Similarly, as with the sceptical Bayes factor, the Q-statistic from (9) appears in (12) and acts as a penalty term, i. e., larger values of Q lower the degree of replication success. However, in

contrast to the sceptical Bayes factor, the replication Bayes factor is not limited by the evidence from the original study because $\text{BF}_\text{R} \downarrow 0$ as $c \to \infty$ provided $z_o \neq 0$ and $d \neq 0$. Moreover, we have that

$$1 \geq \text{BF}_\text{S} \geq \text{BF}_{\text{S:A}}(\hat{\theta}_r; g_{\text{BF}_\text{S}}) = \text{BF}_{\text{S:0}}(\hat{\theta}_r; g_{\text{BF}_\text{S}}) \cdot \underbrace{\text{BF}_{\text{0:A}}(\hat{\theta}_r)}_{=\text{BF}_\text{R}}.$$

So the sceptical Bayes factor is larger than the replication Bayes factor if the replication data favour the sceptical prior $H_\text{S}\colon \theta \sim \text{N}(0, g_{\text{BF}_\text{S}} \cdot \sigma_o^2)$ over the null hypothesis. They can only coincide when $\text{BF}_\text{S} = 1$ since then $g_{\text{BF}_\text{S}} = 0$.

We can also determine conditions on the relative effect estimate in terms of replication success based on $\text{BF}_\text{R} \leq \gamma$. The replication success region is given by

$$d \notin [-\sqrt{J} - H, \sqrt{J} - H] \tag{13}$$

with

$$J = \left\{1 + \frac{\log(1+c) - 2\log\gamma}{z_o^2}\right\} \cdot \frac{1/c + 1}{c}$$
$$H = \frac{1/c + 1}{1 + c}.$$

The condition (13) implies that replication success can also be achieved for negative relative effect estimates d (see Section 3.2 for a discussion of this issue). The bottom-left plot in Figure 4 shows the conditions from (13) for positive relative effect estimates. As with the two-trials rule, the success region of the replication Bayes factor can be pushed arbitrarily close to zero by increasing the relative variance c. In contrast to the two-trials rule, however, replication success can also be achieved for original studies with $\text{minBF}_o > 1/3$.

The sceptical *p*-value

Of particular interest is the relationship between the sceptical Bayes factor and the sceptical p-value (Held, 2020), as it is the outcome of a similar reverse-Bayes procedure. One also considers a sceptical prior for the effect size $\theta \sim \text{N}(0, \tau^2)$, the sufficiently sceptical prior variance at level α is then defined as $\tau^2 = \tau_\alpha^2$ such that the $(1-\alpha)$ credible interval for θ based on the posterior $\theta \,|\, \hat{\theta}_o, \tau_\alpha^2$ does not include zero. Replication success is declared if the tail probability of the replication effect estimate under its prior predictive distribution $\hat{\theta}_r \,|\, \tau_\alpha^2 \sim \text{N}(0, \tau_\alpha^2 + \sigma_r^2)$ is smaller than α. The smallest level α where replication success can be established defines the sceptical p-value. In contrast to the sceptical Bayes factor, the sceptical p-value always exists and there are closed form expressions to compute it for all values of c, i. e., $p_\text{S} = 1 - \Phi(z_\text{S})$ with

$$z_\text{S}^2 = \begin{cases} z_H^2/2 & \text{for } c = 1 \\ \frac{1}{c-1}\left\{\sqrt{z_A^2\left[z_A^2 + (c-1)z_H^2\right]} - z_A^2\right\} & \text{for } c \neq 1 \end{cases}$$

where $z_H^2 = 2/(1/z_o^2 + 1/z_r^2)$ the harmonic mean, $z_A^2 = (z_o^2 + z_r^2)/2$ the arithmetic mean of the squared z-statistics, and provided that $\text{sign}(\hat{\theta}_o) = \text{sign}(\hat{\theta}_r)$ (otherwise $p_\text{S} = \Phi(z_\text{S})$).

Similar to the two-trials rule, the sceptical p-value requires both studies to provide compelling evidence due to the property that $p_S \geq \max\{p_o, p_r\}$. The sceptical p-value also penalises the case when the replication effect estimate shrinks as compared to the original one since it monotonically increases with decreasing relative effect estimate d (Held, 2020, Section 3.1).

Held et al. (2022b) showed that replication success based on $p_S \leq \alpha_S$ is achieved when

$$d \geq \sqrt{\frac{1/c + 1/(K-1)}{K}} \tag{14}$$

with $K = z_o^2/z_{\alpha_S}^2$ where $z_{\alpha_S} = \Phi^{-1}(1-\alpha_S)$. Thresholding the sceptical p-value with the ordinary significance level α for traditional p-values leads to a very stringent criterion for replication success. For example, when $z_o = 2$, $\alpha = 0.025$, and $c = 2$, the replication effect estimate needs to be $d = 4.87$ times larger than the original one. Therefore, Held et al. (2022b) used (14) to determine the *golden level* $\alpha_S = 1 - \Phi(z_\alpha/\sqrt{\varphi})$ with $\varphi = (1+\sqrt{5})/2$ the golden ratio. The golden level ensures that borderline significant original studies ($|z_o| = z_\alpha$) can still achieve replication success provided the replication effect estimate does not shrink compared to the original one ($d \geq 1$). Instead of comparing the sceptical p-value to the golden level ($p_S < \alpha_S$), one can compute a recalibrated sceptical p-value $\tilde{p}_S = 1 - \Phi(z_S\sqrt{\varphi})$ and compare it to the ordinary significance level ($\tilde{p}_S < \alpha$).

The bottom-right plot in Figure 4 shows the success region for the recalibrated sceptical p-value. We see that increasing the precision of the replication study lowers the required minimum relative effect estimate $d_{\min}$ as for all other methods. Similarly, as with the sceptical Bayes factor, $d_{\min}$ of the sceptical p-value cannot be pushed arbitrarily close to zero. However, its limiting minimum relative effect estimate in c ($\lim_{c\to\infty} d_{\min}$) is smaller than the one from the sceptical Bayes factor when $\text{minBF}_o < 1/5.6$, while for $\text{minBF}_o > 1/5.6$ it is the other way around. So for more convincing original studies the sceptical p-value is less stringent than the sceptical Bayes factor. Due to the recalibration, the sceptical p-value also allows replication success when the $\text{minBF}_o > \gamma$. This is visible in the bottom-right plot of Figure 4 where the success region has no cut-off at $\text{minBF}_o = \gamma = 1/3$, unlike the two-trials rule and the sceptical Bayes factor.

3.2 Paradoxes in the assessment of replication success

The replication setting is different from the classical setting where data from only one study are analysed. As a result, several unique paradoxes may occur.

The replication paradox

The *replication-paradox* (Ly et al., 2018) occurs when original and replication effect estimates go in opposite directions ($\text{sign}(\hat{\theta}_o) \neq \text{sign}(\hat{\theta}_r)$) but a method flags replication success. This is undesired since effect direction is crucial to most scientific theories and research questions.

The two-trials rule and the sceptical p-value both avoid the replication paradox by using one-sided test-statistics. In contrast, the sceptical Bayes factor and the replication Bayes factor

may suffer from the paradox as their success regions from (10) and (13), respectively, include negative relative effect estimates $d < 0$. This is related to the fact that Bayes factors are quantifying relative evidence: When the replication estimate goes in the opposite direction, it will be poorly predicted by the sceptical prior H_S and the advocacy prior H_A, yet when H_S is mostly concentrated around zero (or a point-null in case of the replication Bayes factor), replication estimates going in the opposite direction may still be better predicted by H_A.

In practice, the replication paradox is hardly an issue, since replications rarely show such contradictory results, e. g., to achieve replication success at level $\gamma = 1/3$ with $\text{minBF}_o = 1/10$ and $c = 1$, the relative effect estimate needs to be $d < -7.09$ for the paradox to appear with the sceptical Bayes factor. The replication Bayes factor is more prone to the paradox because its point-null hypothesis fails more strongly to predict estimates in opposite direction, e. g., for the same numbers as before it requires $d < -2.66$.

In both cases the paradox can be overcome by truncating the advocacy prior H_A such that only effects in the same direction as the original estimate $\hat{\theta}_o$ have non-zero probability, i. e., for positive $\hat{\theta}_o$ consider $H_{A'}: \theta \sim \text{N}(\hat{\theta}_o, \sigma_o^2)\, \mathbb{1}_{(0,\infty)}(\theta)$, where $\mathbb{1}_B(x)$ is the indicator function of the set B. The Bayes factor contrasting H_S to $H_{A'}$ turns out to be

$$\text{BF}_{\text{S:A}'}(\hat{\theta}_r; g) = \text{BF}_{\text{S:A}}(\hat{\theta}_r; g) \frac{\Phi(|z_o|)}{\Phi\left\{\text{sign}(z_o)\frac{z_o(1+dc)}{\sqrt{1+c}}\right\}} \tag{15}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution (see Appendix D). Hence, (15) is the Bayes factor under the standard advocacy prior multiplied by a correction term. Determining the smallest level of replication success with (15) leads to a corrected sceptical Bayes factor, while setting $g = 0$ in (15) leads to a corrected replication Bayes factor. The correction term goes to one when the replication estimate goes in the same direction as the original one and the replication sample size increases ($d > 0$ and $c \to \infty$), but it penalises when the replication estimate goes in the opposite direction ($d < 0$ and $c \to \infty$).

This modification guarantees that the replication paradox is avoided and we recommend to compute the sceptical Bayes factor using (15) in cases where the replication paradox is likely to appear. However, truncated priors are unnatural and hard to interpret. Also the non-truncated advocacy prior penalises effect estimate incompatibility and the modification (15) will only make a difference in extreme situations. Due to its easier mathematical treatment we will focus on the standard version of the procedure in the remaining part of the paper.

The shrinkage paradox

The comparison showed that for certain methods replication success is still achievable even when the replication estimate is substantially smaller than the original one. However, a substantially smaller effect estimate in the replication does not reflect an effect size of the same practical importance as the original one and a method should thus not flag replication success. The *shrinkage paradox* occurs if a particular method may flag replication success for any arbitrarily small (but positive) relative effect estimate.

Two forms of the shrinkage paradox can formally be distinguished: the *shrinkage paradox at replication* appears when, for fixed evidence from the original study minBF_o (respectively z_o), the minimum relative effect estimate $d_{\text{min}} > 0$ required for replication success at a fixed level γ becomes arbitrarily small as the relative variance c increases:

$$d_{\text{min}} \downarrow 0 \text{ as } c \to \infty.$$

Held et al. (2022b) found that this form of the paradox occurs for the two-trials rule but not for the sceptical p-value. Similarly, the minimum relative effect estimate d_{min} of the sceptical Bayes factor is bounded away from zero, while it converges to zero for the replication Bayes factor (Appendix E). Hence, among the Bayes factor methods, the sceptical Bayes factor avoids the paradox, whereas the replication Bayes factor suffers from it.

The shrinkage paradox at replication is a serious issue since it depends on the relative variance c which can usually be directly influenced by changing the replication sample size. However, there is also a second form of the paradox which is affected only by evidence from the original study. The *shrinkage paradox at original* appears when, for fixed relative variance c, the minimum relative effect estimate $d_{\text{min}} > 0$ required for replication success at a fixed level γ becomes arbitrarily small as the evidence in the original study increases:

$$d_{\text{min}} \downarrow 0 \text{ as } z_o^2 \to \infty.$$

The replication Bayes factor and the sceptical Bayes factor do not suffer from this form of the paradox, while the two-trials rule and the sceptical p-value do (Appendix E). Hence, with the latter two methods, shrinkage of the replication effect estimate is hardly penalised when the original study was already very convincing.

3.3 Frequentist properties

Despite the fact that Bayesian methods do not rely on repeated testing, it is still often of interest to study their frequentist operating characteristics (Dawid, 1982; Grieve, 2016). This is especially important in the replication setting where regulators and funders usually require from statistical methods to have appropriate error control. We will therefore study and compare type I error rate as well as power of the sceptical Bayes factor and other methods.

Global type I error rate

The probability for replication success at level γ conditional on the original result z_o and the relative variance c can be easily computed as shown in Appendix F. Under the null hypothesis ($H_0 : \theta = 0$) the distribution of the z-values is $z_o, z_r \mid H_0 \sim \text{N}(0, 1)$ and hence the global type I error rate (T1E) based on $\text{BF}_\text{S} \leq \gamma$ is

$$\text{T1E} = 2 \int_{z_\gamma}^{\infty} \Pr(\text{BF}_\text{S} \leq \gamma \mid z_o, c) \phi(z_o) \, \text{d}z_o$$

with $\phi(\cdot)$ the standard normal density function. In a similar fashion one can compute the type I error rate of the sceptical p-value (see Section 3 in Held et al., 2022b), as well as the replication Bayes factor (Appendix G). The type I error rate of the two trials rule is simply $\text{T1E} = 2\{1 - \Phi(z_\gamma)\}^2$.

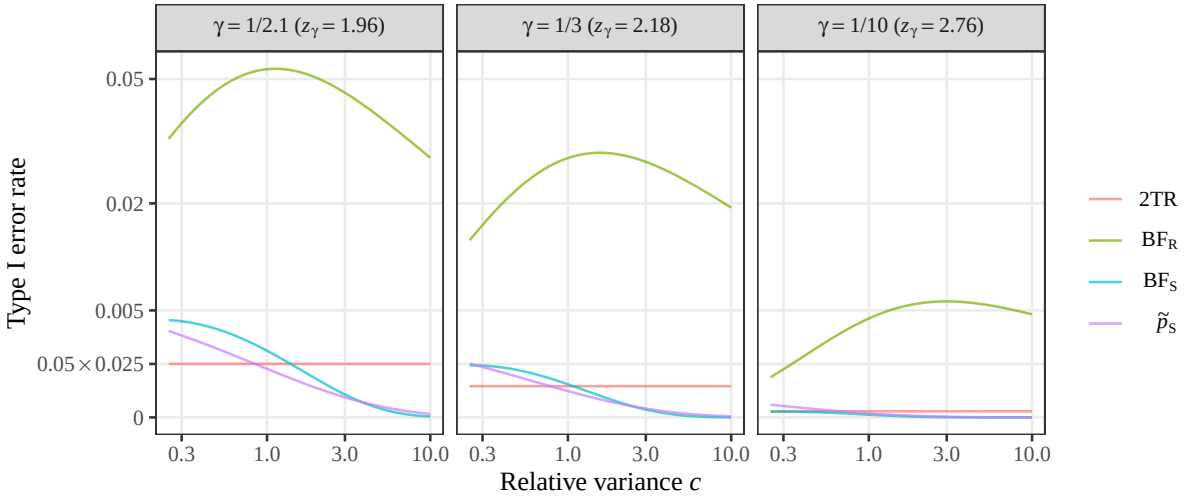


Figure 5: Type I error rate of the two-trials rule (2TR: $\text{minBF}_o \leq \gamma$ and $\text{minBF}_r \leq \gamma$), the replication Bayes factor ($\text{BF}_\text{R} \leq \gamma$), the sceptical Bayes factor ($\text{BF}_\text{S} \leq \gamma$), and the recalibrated sceptical p-value ($\tilde{p}_\text{S} \leq 1 - \Phi\{z_\gamma\}$) as a function of the relative variance $c = \sigma_o^2 / \sigma_r^2$ for different levels of replication success γ.

Figure 5 compares the type I error rates of the four methods for different levels γ. The conventional nominal $\text{T1E} = 0.05 \times 0.025$ (two independent experiments with two-sided testing in the first and one-sided testing in the second) along with the corresponding level ($z_\gamma = 1.96$ corresponding to $\gamma = 1/2.1$ and $\alpha = 0.025$) is also indicated. In contrast to the other methods, the type I error rate of the two trials rule does not depend on the relative variance c and therefore does not change for the same level γ. Type I error rates of sceptical p-value and sceptical Bayes factor are decreasing with increasing c, the former usually being slightly smaller than the latter. The point at which both become smaller than the type I error rate from the two-trials rule becomes smaller with more stringent level γ. Roughly speaking the type I error rate of the sceptical Bayes factor is controlled at the conventional level when c is slightly larger than one, while for the sceptical p-value it is controlled when c is slightly below one. Surprisingly, the type I error rate of the replication Bayes factor is non-monotone in c and far higher compared to the other methods. This suggests that a more stringent level γ should be used for the replication Bayes factor compared to the other methods to ensure appropriate type I error control.

Power conditional on the original study

Another frequentist operating characteristic is the probability to establish replication success assuming there is an underlying effect (power). While in principle original and replication

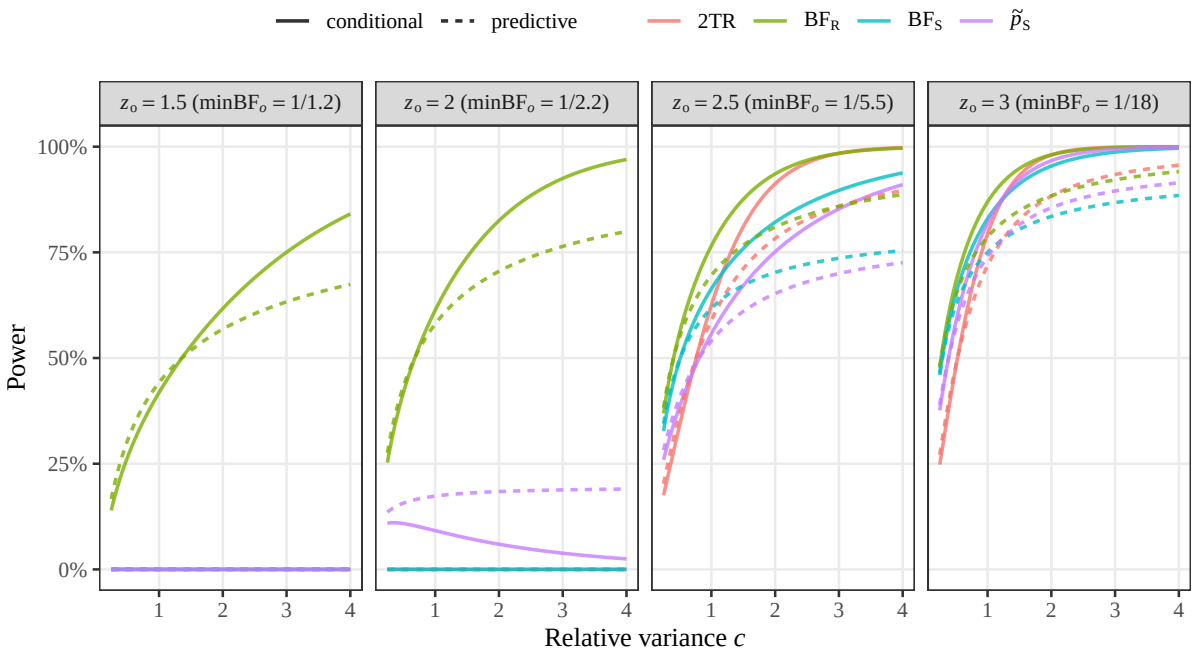


Figure 6: Power of the two-trials rule (2TR: $\text{minBF}_o \leq 1/3$ and $\text{minBF}_r \leq 1/3$), the replication Bayes factor ($\text{BF}_\text{R} \leq 1/3$), the sceptical Bayes factor ($\text{BF}_\text{S} \leq 1/3$), and the recalibrated sceptical p-value ($\tilde{p}_\text{S} \leq 1 - \Phi\{z_\gamma\}$ with $\gamma = 1/3$) as a function of the relative variance $c = \sigma_o^2/\sigma_r^2$ for different original z-values $z_o = \hat{\theta}_o/\sigma_o$ (respectively corresponding minimum Bayes factor minBF_o).

study could be powered simultaneously, we will assume the original study has already been conducted since this is the usual situation. The power to establish replication success $\text{BF}_\text{S} \leq \gamma$ can be computed using the result from Appendix F and either assuming that the underlying true effect corresponds to its estimate from the original study (*conditional power*) or using the predictive distribution of the replication effect estimate based on the advocacy prior (*predictive power*) (Spiegelhalter et al., 1986; Micheloud and Held, 2022). In practice, both forms may be too optimistic as original results are often inflated due to publication bias and questionable research practices. One solution is to shrink the original effect estimate for power calculations (Pawel and Held, 2020; Held et al., 2022b), but we will not focus on this aspect here as this would not provide much more insight but simply lower the power curves of all methods.

Figure 6 shows conditional and predictive power as a function of the relative variance c and for several values of of the original z-value z_o (respectively original minimum Bayes factor minBF_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 $z_o = 1.5$ or $z_o = 2$), it is impossible to achieve replication success with the two-trials rule, the sceptical Bayes factor, and the sceptical p-value. This is not the case for the replication Bayes factor, for which high power can also be obtained for small z_o if c is sufficiently large. However, as shown in the previous section, the higher power of the replication Bayes factor comes at the cost of a massive type I error inflation. For $z_o = 2.5$, the sceptical Bayes factor shows higher power than the two-trials rule when $c = 1$,

but the power of the two-trials rule increases faster in c and approaches the power curve of the replication Bayes factor. The power of the sceptical p-value is still a bit lower, likely due to the more stringent requirement on the minimum relative effect estimate. For $z_o = 3$, the power differences between the methods mostly disappear.

3.4 Information consistency

Bayesian hypothesis testing procedures are desired to fulfil certain asymptotic properties (Bayarri et al., 2012). Most notably, they should be *information consistent* in the sense that if data provide overwhelming support for a particular hypothesis, the procedure should indefinitely favour this hypotheses over alternative hypotheses.

There are concerns whether the sceptical Bayes factor is information consistent when we look at the asymptotics only in terms of the replication data (Consonni and La Rocca, 2021; Ly and Wagenmakers, 2021). The sceptical Bayes factor can never be smaller than the original minimum Bayes factor minBF_o. This means that it will be bounded away from zero as the replication sample size grows ($c \to \infty$), even when the data are generated from the same model in both studies. Similarly, the sceptical Bayes factor will be bounded away from zero when the replication effect estimate increases indefinitely. If these two cases constitute overwhelming evidence for replication success, they could be considered instances of the *information paradox* (Liang et al., 2008)

The key to resolving the paradox is to realise that overwhelming evidence for replication success needs to be defined through both studies, not only through the replication. Assume there is a "true" effect size $\theta_* \neq 0$ underlying both effect estimates $\hat{\theta}_i \sim \text{N}(\theta_*, \sigma_i^2)$, $i \in \{o, r\}$. Also assume the variances $\sigma_i^2 = \kappa/n$ are inversely proportional to the sample size $n = n_o = n_r$ for the same unit variance κ in both studies. Letting the sample size n go to infinity is then equivalent to $\sigma_o^2 \downarrow 0$ and $c = \sigma_r^2/\sigma_o^2 = 1$. With decreasing variances the estimates will converge to the true effect size ($\hat{\theta}_i \to \theta_*$), the relative effect estimate will converge to one ($d \to 1$), and the z-values will go to infinity ($|z_i| \to \infty$). Since $c = 1$, the sceptical Bayes factor is given by equation (7). Moreover, we are allowed to use of the approximation $W_{-1}(x) \approx \log(-x) - \log(-\log(-x))$ as the argument of the Lambert function is close to zero due to $|z_o| \to \infty$ (Corless et al., 1996, p. 350). Taken together, we have

$$\text{BF}_\text{S} = \sqrt{\frac{1 + d^2}{1 + (1 - d)^2/2 - \mathcal{O}\left\{\log(z_o^2)/z_o^2\right\}}} \cdot \exp\left\{-\frac{z_o^2\left(d^2 + 2d - 1\right)}{4(1 + d^2)} - \mathcal{O}\left(\log z_o^2\right)\right\} \tag{16}$$

Plugging $d = 1$ into (16), we see that $\text{BF}_\text{S} \downarrow 0$ as $|z_o| \to \infty$, so the sceptical Bayes factor is information consistent.

The expression for the sceptical Bayes factor (16) is also valid for other relative effect sizes d. Solving for d such that the multiplicative term of z_o^2 in the exponent changes the sign, we see that the sceptical Bayes factor goes to zero when the underlying true effect size of the replication study is at least $d > \sqrt{2} - 1 \approx 0.41$ times the size of the true effect size from the original study (or $d < -\sqrt{2} - 1 \approx -2.41$ due to the replication paradox if the advocacy prior is not truncated). This means that under the more realistic scenario where the underlying

effect sizes from original and replication are not exactly the same, the sceptical Bayes factor is still consistent when there is not more than 60% shrinkage of the replication effect size.

4 Extension to non-normal models

So far, we have always assumed approximate normality of the effect estimates $\hat{\theta}_o$ and $\hat{\theta}_r$, as well as known variances σ_o^2 and σ_r^2. This may be a problem for studies with small sample size and/or extreme results (e. g., when a study examines a rare disease with death rates close to 0%). One way of dealing with this issue is to consider the exact likelihood of the data underlying the effect estimates, and then marginalise over possible nuisance parameters (Spiegelhalter et al., 2004, Chapter 8.2.2). This leads to marginal likelihoods which are again only conditional on the effect size θ, allowing the procedures to be used analogously as described in the proceeding sections. The choice of the likelihood depends on the type of effect size θ. We will illustrate the approach for *standardised mean differences* (SMD) and *log odds ratios* (logOR), two of the most widely used types of effect sizes.

4.1 Standardised mean difference

The SMD quantifies how many standard deviation units σ, the means μ_1 and μ_2 of measurements from two groups differ, i. e.,

$$\theta = \frac{\mu_1 - \mu_2}{\sigma}.$$

Assume now that the measurements come from a normal distribution with common variance σ^2. Knowing the test-statistic t_i from the usual two-sample t-test, as well as the sample sizes in both groups n_{1i} and n_{2i} from study $i \in \{o, r\}$ is sufficient to compute the exact likelihood of the data. It is given by a non-central t-distribution with degrees of freedom $\nu_i = n_{1i} + n_{2i} - 2$ and non-centrality parameter $\theta\sqrt{n_i^*}$ with $n_i^* = (n_{1i}n_{2i})/(n_{1i} + n_{2i})$ (Bayarri and Mayoral, 2002b)

$$T_i \,|\, \theta \sim \text{NCT}_{\nu_i}\left(\theta\sqrt{n_i^*}\right). \tag{17}$$

The same framework is also applicable to test-statistics t_i from paired t-tests based on n_i paired measurements. The SMD θ represents then the standardised mean difference score and $\nu_i = n_i - 1$ and $n_i^* = n_i$ need to be used in (17).

There is no conjugate prior for the SMD θ under model (17), so it is not obvious which prior should be chosen to represent scepticism about it. We will use a zero-mean normal prior $\theta \,|\, H_S \sim \text{N}(0, \tau^2)$ so that the exact procedure is equivalent with the normal approximation as the sample size increases. For the advocacy prior we need to know the posterior distribution of the SMD θ conditional on the original study and a flat prior on θ. Exploiting the fact that the non-central t-distribution can be expressed as a location-scale mixture of a normal with an inverse-gamma distribution (Johnson et al., 1995, Chapter 31), the density of the SMD under the advocacy prior is given by

$$f(\theta \,|\, t_o) = \int_0^\infty \text{N}\left(\theta; \frac{t_o}{\sqrt{n_o^*\tau^2}}, \frac{1}{n_o^*}\right) \text{IG}\left(\tau^2; \frac{\nu_o + 1}{2}, \frac{\nu_o}{2}\right) \text{d}\tau^2,$$

where $\mathrm{N}(x; \mu, \phi)$ denotes the density function of the normal distribution with mean μ and variance ϕ evaluated at x, and similarly $\mathrm{IG}(y; a, b)$ denotes the density function of the inverse-gamma distribution with shape and rate parameters a and b evaluated at y.

Taken together, the SMD version of the method proceeds analogously as in Box 1 with the two Bayes factors replaced by

$$\mathrm{BF}_{0:\mathrm{S}}(t_o; \tau^2) = \frac{\mathrm{NCT}_{\nu_o}(t_o; 0)}{\int \mathrm{NCT}_{\nu_o}(t_o; \theta\sqrt{n_o^*})\, \mathrm{N}(\theta; 0, \tau^2)\, \mathrm{d}\theta}$$
$$\mathrm{BF}_{\mathrm{S:A}}(t_r; \tau^2) = \frac{\int \mathrm{NCT}_{\nu_r}(t_r; \theta\sqrt{n_r^*})\, \mathrm{N}(\theta; 0, \tau^2)\, \mathrm{d}\theta}{\int \mathrm{NCT}_{\nu_r}(t_r; \theta\sqrt{n_r^*}) f(\theta \mid t_o)\, \mathrm{d}\theta}$$

and using numerical integration as the integrals cannot be evaluated analytically.

4.2 Log odds ratio

In the case of binary data, we have two independent binomial samples

$$X_{1i} \mid \pi_1 \sim \mathrm{Bin}(n_{1i}, \pi_1) \qquad\qquad X_{2i} \mid \pi_2 \sim \mathrm{Bin}(n_{2i}, \pi_2)$$

for each study $i \in \{o, r\}$, and the effect of the treatment in group 1 relative to the treatment in group 2 is quantified with the logOR

$$\theta = \log \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}.$$

To obtain a marginal likelihood that only depends on θ, we need to specify a prior for either π_2 or π_1 and marginalise over it. A principled choice is the translation invariant Jeffreys prior, $\pi_1, \pi_2 \sim \mathrm{Be}(1/2, 1/2)$. The exact marginal likelihood for the data from study i is then given by

$$f(x_{1i}, x_{2i} \mid \theta) = \int_0^1 \mathrm{Bin}\left(x_{1i}; n_{1i}, \left\{1 + \exp\left[-\theta - \log\frac{\pi_2}{1-\pi_2}\right]\right\}^{-1}\right) \mathrm{Bin}(x_{2i}; n_{2i}, \pi_2)$$
$$\times\, \mathrm{Be}(\pi_2; 1/2, 1/2)\, \mathrm{d}\pi_2 \tag{18}$$

where $\mathrm{Bin}(x; n, \pi)$ denotes the probability mass function of the binomial distribution with n trials and probability π evaluated at x, and likewise $\mathrm{Be}(y; a, b)$ denotes the density function of the beta distribution with parameters a and b evaluated at y.

There is no conjugate prior for the logOR under model (18), but a pragmatic choice is to specify a zero-mean normal prior $\theta \mid H_{\mathrm{S}} \sim \mathrm{N}(0, \tau^2)$ for the sceptic, to match with the normal approximation as the sample size increases. For the advocacy prior, we need to know the posterior distribution of the logOR θ based on the original study. Using a result from Marshall (1988) combined with a change-of-variables, the exact posterior density of the logOR θ given the original data and Jeffreys priors on π_1 and π_2 is

$$f(\theta \mid x_{1o}, x_{2o}) = \begin{cases} C\, \exp\{e\theta\}\, F\left(e+f, e+g, e+f+g+h, 1-\exp\{\theta\}\right) & \text{for } \theta < 0 \\ C\, \exp\{-f\theta\}\, F\left(e+f, f+h, e+f+g+h, 1-\exp\{-\theta\}\right) & \text{for } \theta > 0 \end{cases}$$

where $F(\cdot)$ is the hypergeometric function, $e = x_{1o} + 1/2$, $f = n_{1o} - x_{1o} + 1/2$, $g = x_{2o} + 1/2$, $h = n_{2o} - x_{2o} + 1/2$, $C = \mathrm{B}(e+g, f+h)/\{\mathrm{B}(e,f)\mathrm{B}(g,h)\}$, and $\mathrm{B}(\cdot,\cdot)$ is the Beta function.

Combining the previous results, we obtain

$$\mathrm{BF}_{\mathrm{0:S}}(x_{1o}, x_{2o}; \tau^2) = \frac{f(x_{1o}, x_{2o} \mid 0)}{\int f(x_{1o}, x_{2o} \mid \theta)\, \mathrm{N}(\theta; 0, \tau^2)\, \mathrm{d}\theta}$$
$$\mathrm{BF}_{\mathrm{S:A}}(x_{1r}, x_{2r}; \tau^2) = \frac{\int f(x_{1r}, x_{2r} \mid \theta)\, \mathrm{N}(\theta; 0, \tau^2)\, \mathrm{d}\theta}{\int f(x_{1r}, x_{2r} \mid \theta) f(\theta \mid x_{1o}, x_{2o})\, \mathrm{d}\theta}$$

as an exact replacement for the Bayes factors in Box 1. Again, there are no closed form expressions for the integrals, but numerical integration needs to be used.

5 Application

The following section will illustrate application of the sceptical Bayes factor using data from the *Social Sciences Replication Project* (Camerer et al., 2018), provided in Table 1. Effect estimates were reported on the correlation scale (r), which is why we applied the Fisher z-transformation $\hat{\theta} = \tanh^{-1}(r)$. This leads to the transformed estimates having approximate variance $\mathrm{Var}(\hat{\theta}) = 1/(n-3)$ (Fisher, 1921), so the relative variance c is roughly the ratio of the replication to the original study sample size $c \approx n_r/n_o$.

For all studies except Janssen et al. (2010) and Derex et al. (2013), the exact approach for either SMD or logOR effect sizes from Section 4 is applicable. In the studies with binary data computing the exact posterior using the hypergeometric function led to numerical issues in some cases and numerical integration was used then. In most cases, the normal approximation of the likelihood seems to lead to similar numerical results for both $\mathrm{BF_S}$ and $\mathrm{BF_R}$ as compared to their counterparts based on exact likelihoods. Qualitative conclusions are the same under both approaches and we will therefore focus on the normal approximation due to better comparability with the remaining measures of replication success as all of them were computed based on approximate normal likelihoods.

For the study pairs where the sceptical Bayes factor suggests a large degree of replication success, all other methods suggest the same in every case. However, there are also cases where there appear to be discrepancies among the methods. For instance, the two-trials rule and the replication Bayes factor may indicate a larger degree of replication success compared to the sceptical p-value and sceptical Bayes factor. This happens for replications that show a substantial increase in sample size but also a much smaller effect estimate compared to the original study. For example, in Balafoutas and Sutter (2012) the sample size was about $c = 3.48$ times larger in the replication, whereas the effect estimate was only $d = 0.52$ the size of the original one. The replication is successful at $\gamma = 1/3$ with the two-trials rule ($\mathrm{minBF}_o = 1/4.2$ and $\mathrm{minBF}_r = 1/3.6$) and the replication Bayes factor ($\mathrm{BF_R} = 1/3.9$), but not with the sceptical Bayes factor ($\mathrm{BF_R} = 1/1.6$) or the sceptical p-value ($\tilde{p}_{\mathrm{S}} = 0.04 > 1 - \Phi(z_\gamma = 2.18) = 0.01$).

Table 1: Results for data from *Social Sciences Replication Project* (Camerer et al., 2018). Shown are relative variances $c = \sigma_o^2/\sigma_r^2$, relative effect estimates $d = \hat{\theta}_r/\hat{\theta}_o$ (computed on Fisher z-scale), Q-statistic $Q = (\hat{\theta}_o - \hat{\theta}_r)^2/(\sigma_o^2 + \sigma_r^2)$, minimum Bayes factors of original and replication effect estimate (minBF), recalibrated sceptical p-value ($\tilde{p}_S$), sceptical Bayes factors (BF_S) and replication Bayes factors (BF_R), the latter two computed using either a normal approximation or the exact likelihood of the data.

Original study	c	d	Q	minBF_o	minBF_r	$\tilde{p}_S$	BF_S	BF_S (exact)	BF_R	BF_R (exact)
Hauser et al. (2014)	0.51	1.04	0.03	< 1/1000	< 1/1000	< 0.0001	< 1/1000	< 1/1000	< 1/1000	< 1/1000
Aviezer et al. (2012)	0.92	0.60	3.49	< 1/1000	1/347	< 0.0001	1/78	1/10	1/284	1/41
Wilson et al. (2014)	1.33	0.83	0.28	< 1/1000	1/659	0.0001	1/45	1/35	< 1/1000	< 1/1000
Derex et al. (2013)	1.29	0.65	1.14	1/520	1/17	0.002	1/8.5		1/31	
Gneezy et al. (2014)	2.31	0.81	0.22	1/18	1/157	0.004	1/6.9	1/7.5	1/474	1/551
Karpicke and Blunt (2011)	1.24	0.58	1.75	< 1/1000	1/9.6	0.002	1/5.6	1/5	1/12	1/12
Morewedge et al. (2010)	2.97	0.76	0.30	1/7.3	1/65	0.011	1/3.9	1/4	1/160	1/156
Kovacs et al. (2010)	4.38	1.38	0.59	1/3.2	< 1/1000	0.009	1/3.2	1/3.8	< 1/1000	< 1/1000
Duncan et al. (2012)	7.42	0.57	1.29	1/12	< 1/1000	0.011	1/3.1	1/3.1	< 1/1000	< 1/1000
Nishi et al. (2015)	2.42	0.57	1.05	1/12	1/6.1	0.016	1/2.5	1/2.2	1/8.2	1/7.6
Janssen et al. (2010)	0.65	0.48	3.51	< 1/1000	1/3.3	0.003	1/1.6		1/1.6	
Balafoutas and Sutter (2012)	3.48	0.52	1.02	1/4.2	1/3.6	0.04	1/1.6	1/1.6	1/3.9	1/3.9
Pyc and Rawson (2010)	9.18	0.38	1.79	1/3.5	1/7.3	0.061	1/1.2	1/1.2	1/4	1/4
Rand et al. (2012)	6.27	0.18	3.96	1/7.1	1	0.13			9.6	9.7
Ackerman et al. (2010)	11.69	0.23	2.15	1/2.2	1/1.3	0.15			3.2	3.2
Sparrow et al. (2011)	3.50	0.13	5.80	1/26	1	0.19			29	32
Shah et al. (2012)	11.62	-0.05	4.08	1/2.2	1	0.66			25	26
Kidd and Castano (2013)	8.57	-0.10	6.83	1/5.7	1	0.77			72	69
Gervais and Norenzayan (2012)	9.78	-0.12	5.44	1/3	1	0.78			36	37
Lee and Schwarz (2010)	7.65	-0.11	6.80	1/5.4	1	0.79			65	69
Ramirez and Beilock (2011)	4.47	-0.09	19.29	< 1/1000	1	0.85			> 1000	> 1000

Discrepancies between the sceptical p-value and the sceptical Bayes factor happen in situations where the replication shows an effect estimate that, although incompatible with the sceptical prior, is also incompatible with the advocacy prior. For example in the Janssen et al. (2010) replication, both effect estimates are substantially larger than zero ($\hat{\theta}_o = 0.74$ with $\text{minBF}_o < 1/1000$ and $\hat{\theta}_r = 0.36$ with $\text{minBF}_r = 1/3.3$), yet the Q-statistic indicates some incompatibility ($Q = 3.51$), which explains why $\tilde{p}_S = 0.003$, but $\text{BF}_S = 1/1.6$ only.

Discrepancies between the replication Bayes factor and the sceptical Bayes factor arise when the replication finding provides overwhelming evidence against the null, whereas the original finding was less compelling. The replication of Kovacs et al. (2010) illustrates this situation. The original study provided only moderate evidence against the null ($\hat{\theta}_o = 0.49$ and $\text{minBF}_o = 1/3.2$), whereas the replication finding was more compelling ($\hat{\theta}_r = 0.67$ and $\text{minBF}_r < 1/1000$). By construction the sceptical Bayes factor can only be as small as the minimum Bayes factor from the original study minBF_o, which is actually attained in this case ($\text{BF}_S = 1/3.2$). The replication Bayes factor, on the other hand, is not limited by the moderate level of evidence from the original study and indicates decisive evidence for the advocate ($\text{BF}_R < 1/1000$). This illustrates that in order to achieve a reasonable degree of replication success, the sceptical Bayes factor requires the original study to be convincing, whereas the replication Bayes factor only requires a compelling replication result.

6 Discussion

We proposed a novel method for the statistical assessment of replicability combining reverse-Bayes analysis with Bayesian hypothesis testing. Compared to other methods, the sceptical Bayes factor poses more stringent requirements but also allows for stronger statements about replication success. It ensures that both studies provide sufficient evidence against a null effect, while also penalising incompatibility of their effect estimates. If the replication sample size is not too small, the sceptical Bayes factor comes with appropriate frequentist error rates, which is often a requirement from research funders and regulators. Asymptotic analysis of the method showed that it is information consistent in the sense that if the sample size in both studies increases, the sceptical Bayes factor will indicate overwhelming replication success when the underlying effect size of the replication is not much smaller than the underlying effect size of the original study. Finally, the sceptical Bayes factor is the only method in our comparison which does not suffer from any form of the shrinkage paradox, i. e., replication success can never be achieved with arbitrarily small replication effect estimates, not even when the replication sample size becomes very large or the evidence from the original study overwhelming.

In extreme scenarios the sceptical Bayes factor can suffer from the replication paradox, which means that it may flag success when the replication estimate goes in opposite direction of the original one. However, the paradox can be avoided by truncating the advocacy prior to the direction of the original estimate. It may also happen that the result of the replication is so inconclusive that replication success cannot be established at any level, so the sceptical Bayes factor does not exist. Other methods, such as the sceptical p-value or the replication Bayes factor, can be used in this situation.

The proposed method could be extended in many ways. First, 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 to easily extend the sceptical Bayes factor to the "many-to-one" replication setting as the likelihoods are also straightforward to compute for a sample of replication effect estimates. Second, a multivariate generalisation would allow for effects in the form of a vector 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. Third, based on the replication result one could also compute a posterior distribution for the effect size based on a model-average of the advocacy prior and the sceptical prior (using the variance at the sceptical Bayes factor). This distribution would provide a formal compromise between scepticism and advocacy of the original finding. Fourth, while Bayes factors are an important part in Bayesian hypothesis testing, they do not take into account the prior probabilities of the hypotheses under consideration. It would be interesting to investigate whether the reverse-Bayes approach could be used in a framework where priors are assigned jointly to the hypothesis and parameter space (Dellaportas et al., 2012). Finally, an important aspect is the design of new replication studies. An appropriate sample size is of particular importance for a replication to be informative. We will report in the future on sample size planning based on the sceptical Bayes factor.

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. Reverse-Bayes methods naturally fit to the replication setting, they avoid various paradoxes from which other methods suffers, and they combine different notions of replicability. The reverse-Bayes approach therefore leads to sensible inferences and decisions, which is why we advocate it as a key part in the assessment of replication success.

Software and data

All analyses were performed in the R programming language version 4.2.2 (R Core Team, 2022). 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 calculations were conducted using the package `ReplicationSuccess` available on the Comprehensive R Archive Network (Held, 2020). All methods are implemented in the R package `BayesRep` which is available at `https://gitlab.uzh.ch/samuel.pawel/BayesRep`.

Data on effect estimates from the *Social Sciences Replication Project* (Camerer et al., 2018) were downloaded from `https://osf.io/abu7k/`, respectively, taken from `https://osf.io/nsxgj/` for exact calculations.

Acknowledgements

We thank the anonymous referees for the helpful comments and suggestions that have considerably improved the paper. We also thank Guido Consonni, Luca La Rocca, Malgorzata Roos, Georgia Salanti, Charlotte Micheloud, 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.

Bibliography

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

Balafoutas, L. and Sutter, M. (2012). Affirmative action policies promote women and do not harm efficiency in the laboratory. *Science*, 335(6068):579–582. doi:10.1126/science.1211180.

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., Berger, J. O., Forte, A., and García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, 40(3):1550–1577. doi:10.1214/12-aos1013.

Bayarri, M. J. and Mayoral, A. M. (2002b). Bayesian design of "successful" replications. *The American Statistician*, 56:207–214. doi:10.1198/000313002155.

Berger, J. (2001). Discussion of "Why should clinicians care about Bayesian methods?" by Robert A.J. Matthews. *Journal of Statistical Planning and Inference*, 94(1):65–67. doi:10.1016/s0378-3758(00)00235-4.

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

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *Journal of the Royal Statistical Society, Series A*, 143:383–430.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., et al. (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., et al. (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.

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

Consonni, G. and La Rocca, L. (2021). The sceptic and the advocate: comparing two opinions on the mean of a normal distribution. Unpublished notes.

Cooper, H., Hedges, L. V., and Valentine, J. C., editors (2019). *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation. doi:10.7758/9781610448864.

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., Berniūnas, R., Boudesseul, J., Colombo, M., et al. (2018). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*. doi:10.1007/s13164-018-0400-9.

Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610. doi:10.1080/01621459.1982.10477856.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2012). Joint specification of model space and parameter space prior distributions. *Statistical Science*, 27(2). doi:10.1214/11-sts369.

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.

Etz, A. and Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLOS ONE*, 11(2):e0149794. doi:10.1371/journal.pone.0149794.

Evans, M. and Moshonov, H. (2006). Checking for prior-data conflict. *Bayesian Analysis*, 1(4):893–914. doi:10.1214/06-ba129.

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 Weighing of Evidence*. Griffin, London, UK.

Grieve, A. P. (2016). Idle thoughts of a 'well-calibrated' Bayesian in clinical drug development. *Pharmaceutical Statistics*, 15(2):96–108. doi:10.1002/pst.1736.

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). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychological Methods*, 24(5):557–570. doi:10.1037/met0000189.

Held, L. (2020). 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., Matthews, R., Ott, M., and Pawel, S. (2022a). Reverse-Bayes methods for evidence assessment and research synthesis. *Research Synthesis Methods*. doi:10.1002/jrsm.1538.

Held, L., Micheloud, C., and Pawel, S. (2022b). The assessment of replication success based on relative effect size. URL `https://www.e-publications.org/ims/submission/AOAS/user/submissionFile/47896?confirm=532335fe`. to appear in *The Annals of Applied Statistics*.

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.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions, Vol. 2*. Wiley.

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., and Mandal, S. (2016). On the reproducibility of psychological science. *Journal of the American Statistical Association*, 112(517):1–10. doi:10.1080/01621459.2016.1240079.

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.

Kay, R. (2015). *Statistical Thinking for Non-Statisticians in Drug Regulation*. John Wiley & Sons, Chichester, U.K., second edition. doi:10.1002/9781118451885.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, v., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., et al. (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.

Ly, A. and Wagenmakers, E.-J. (2021). Bayes factors for peri-null hypotheses. doi:10.48550/arXiv.2102.07162.

Marshall, R. J. (1988). Bayesian analysis of case-control studies. *Statistics in Medicine*, 7(12):1223–1230. doi:10.1002/sim.4780071203.

Mathur, M. B. and VanderWeele, T. J. (2020). New statistical metrics for multisite replication projects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(3):1145–1166. doi:10.1111/rssa.12572.

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

Micheloud, C. and Held, L. (2022). Power calculations for replication studies. *Statistical Science*, 37(3):369–379. doi:10.1214/21-sts828.

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 (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, 26:559–569. doi:10.1177/0956797614567341.

Spiegelhalter, D. J., Abrams, R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. New York: Wiley.

Spiegelhalter, D. J., Freedman, L. S., and Blackburn, P. R. (1986). Monitoring clinical trials: Conditional or predictive power? *Controlled Clinical Trials*, 7(1):8–17. doi:10.1016/0197-2456(86)90003-6.

van Aert, R. C. M. and van Assen, M. A. L. M. (2017). Bayesian evaluation of effect size after replicating an original study. *PLOS ONE*, 12(4):e0175302. doi:10.1371/journal.pone.0175302.

Verhagen, J. and Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, 143:1457–1475. doi:10.1037/a0036731.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer International Publishing. doi:10.1007/978-3-319-24277-4.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In Goel, P. and Zellner, A., editors, *Bayesian Inference and Decision techniques: Essays in Honor of Bruno de Finetti*, volume 6 of *Studies in Bayesian Econometrics and Statistics*, pages 233–243. Amsterdam: North-Holland.

A Sufficiently sceptical relative prior variance

The sufficiently sceptical relative prior variance at level γ is the value $g_\gamma \in [0, g_{\text{minBF}_o}]$ that fulfils the condition

$$\text{BF}_{0:S}(\hat{\theta}_o; g_\gamma) = \gamma. \tag{19}$$

Substituting (19) and rearranging terms, we obtain

$$\begin{aligned}\sqrt{1+g_\gamma} \cdot \exp\left\{-\frac{1}{2} \cdot \frac{g_\gamma}{1+g_\gamma} \cdot z_o^2\right\} &= \gamma \\ \iff \frac{1}{\gamma} \cdot \exp\left\{-\frac{z_o^2}{2}\right\} &= \frac{1}{\sqrt{1+g_\gamma}} \exp\left\{-\frac{1}{2} \cdot \frac{z_o^2}{1+g_\gamma}\right\}.\end{aligned}$$

Squaring both sides and multiplying by $-z_o^2$, this becomes

$$\iff -\frac{z_o^2}{\gamma^2} \cdot \exp\left\{-z_o^2\right\} = -\frac{z_o^2}{1+g_\gamma} \exp\left\{-\frac{z_o^2}{1+g_\gamma}\right\}. \tag{20}$$

This is a transcendental equation that cannot be explicitly solved in terms of elementary functions. However, if we set $q = -z_o^2/(1+g_\gamma)$ then (20) becomes

$$-\frac{z_o^2}{\gamma^2} \cdot \exp\left\{-z_o^2\right\} = q \cdot \exp\left\{q\right\}.$$

The solution for q (and consequently for g_γ) can be explicitly computed with

$$\begin{aligned} q &= W_{-1}\left(-\frac{z_o^2}{\gamma^2} \cdot \exp\left\{-z_o^2\right\}\right) \\ g_\gamma &= \begin{cases} -\dfrac{z_o^2}{q} - 1 & \text{if } -\dfrac{z_o^2}{q} \geq 1 \\ \text{undefined} & \text{else} \end{cases} \end{aligned} \tag{21}$$

where $W_{-1}(\cdot)$ is the branch of the Lambert W function that satisfies $W(y) \leq -1$ for $y \in [-e^{-1}, 0)$, ensuring that $g_\gamma \leq g_{\text{minBF}_o}$. See Appendix B for details about the Lambert W function. For some z_o, equation (20) can also be satisfied for negative g_γ, which is why we need to add the condition $-z_o^2/q \geq 1$ in equation (21), such that g_γ is a valid relative variance.

As z_o^2 becomes larger, the argument to the Lambert function $x = -z_o^2 \exp(-z_o^2)/\gamma^2$ will approach zero, so the approximation $W_{-1}(x) \approx \log(-x) - \log(-\log(-x))$ can be applied (Corless et al., 1996, p. 350). This leads to

$$g_\gamma \approx \frac{z_o^2}{z_o^2 + \log\gamma^2 - \log z_o^2 + \log\{z_o^2 + \log\gamma^2 - \log z_o^2\}} - 1.$$

We can see that $g_\gamma \downarrow 0$ when γ remains fixed and $z_o^2 \to \infty$, which means that the sufficiently sceptical relative prior variance converges to zero for increasingly compelling evidence from the original study.

B The Lambert W function

The Lambert W function (Corless et al., 1996) is defined as the function $W(\cdot)$ satisfying

$$W(y) \cdot \exp\{W(y)\} = y$$

and it 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 7 for an illustration).

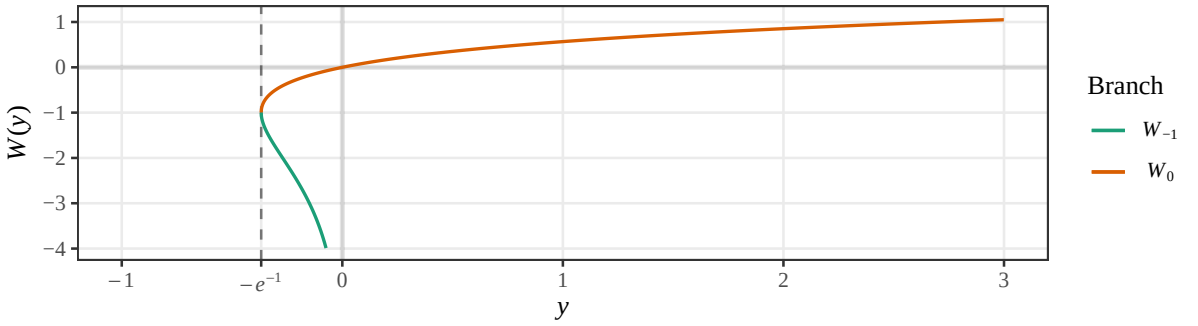


Figure 7: Lambert W function for real argument y.

C Computation of the sceptical Bayes factor

From the definition of the sceptical Bayes factor it is apparent that BF_S is either

1. undefined, if $\text{BF}_{\text{S:A}}(\hat{\theta}_r; g) > \text{BF}_{\text{0:S}}(\hat{\theta}_o; g)$ for all $g \in [0, g_{\text{minBF}_o}]$
2. $\text{BF}_\text{S} = \text{minBF}_o$, if $\text{BF}_{\text{S:A}}(\hat{\theta}_r; g_{\text{minBF}_o}) \leq \text{BF}_{\text{0:S}}(\hat{\theta}_o; g_{\text{minBF}_o})$
3. $\text{BF}_\text{S} = \inf_{g_\gamma} \{\gamma : \text{BF}_{\text{S:A}}(\hat{\theta}_r; g_\gamma) = \gamma\}$, the height of the lowest intersection of $\text{BF}_{\text{0:S}}(\hat{\theta}_o; g_\gamma) = \gamma$ and $\text{BF}_{\text{S:A}}(\hat{\theta}_r; g_\gamma)$ in g_γ otherwise

Whether BF_S attains the lower bound minBF_o (condition 2) can be checked by evaluating if $\text{BF}_{\text{S:A}}(\hat{\theta}_r; g_{\text{minBF}_o}) \leq \text{minBF}_o$ and setting $\text{BF}_\text{S} = \text{minBF}_o$ if it is the case. For condition 3, we know that the intersections satisfy

$$\text{BF}_{\text{S:A}}(\hat{\theta}_r; g_*) = \text{BF}_{\text{0:S}}(\hat{\theta}_o; g_*)$$

$$\sqrt{\frac{1/c + 1}{1/c + g_*}} \cdot \exp\left\{-\frac{z_o^2}{2}\left(\frac{d^2}{1/c + g_*} - \frac{(1-d)^2}{1/c + 1}\right)\right\} = \sqrt{1 + g_*} \cdot \exp\left\{-\frac{1}{2} \cdot \frac{g_*}{1 + g_*} \cdot z_o^2\right\}$$

which is equivalent to

$$\sqrt{\frac{1}{(1+g_*)(1/c+g_*)}} \cdot \exp\left\{-\frac{z_o^2}{2}\left(\frac{1}{1+g_*}+\frac{d^2}{1/c+g_*}\right)\right\} = \sqrt{\frac{1}{1/c+1}} \cdot \exp\left\{-\frac{z_o^2}{2}\left(1+\frac{(1-d)^2}{1/c+1}\right)\right\}. \tag{22}$$

This is a transcendental equation that has no closed-form solution for g_* in terms of elementary functions, but root-finding algorithms can be used to compute it. However, when $c = 1$, equation (22) simplifies

$$\frac{1}{1+g_*} \cdot \exp\left\{-\frac{z_o^2}{2} \cdot \frac{1+d^2}{1+g_*}\right\} = \frac{1}{\sqrt{2}} \cdot \exp\left\{-\frac{z_o^2}{2}\left(1+\frac{(1-d)^2}{2}\right)\right\}. \tag{23}$$

Multiplying (23) by $-z_o^2(1+d^2)/2$ and applying the Lambert W function leads to

$$\begin{aligned} k &= W\left(-\frac{z_o^2}{\sqrt{2}} \cdot \frac{1+d^2}{2} \cdot \exp\left\{-\frac{z_o^2}{2}\left[1+\frac{(1-d)^2}{2}\right]\right\}\right) \\ g_* &= \begin{cases} -\dfrac{z_o^2}{k} \cdot \dfrac{1+d^2}{2} - 1 & \text{if } -\dfrac{z_o^2}{k} \cdot \dfrac{1+d^2}{2} \geq 1 \\ \text{undefined} & \text{else,} \end{cases} \end{aligned} \tag{24}$$

with the condition that $-z_o^2(1+d^2)/(2k) \geq 1$ 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 $z_o \neq 0$), the branches $W_{-1}(\cdot)$ and $W_0(\cdot)$ both provide solutions that can fulfil the equation (assuming the argument is not smaller than $-e^{-1}$ which would mean that there are no intersections). It must also hold that $g_* \leq g_{\text{minBF}_o} = \max\{z_o^2 - 1, 0\}$ for g_* to be a valid sufficiently sceptical prior variance. Hence, when $|d| \leq 1$, the g_* from (24) can only be computed with the $W_{-1}(\cdot)$ branch, whereas for $|d| > 1$ and when $-k \geq (1+d^2)/2$ the solution g_* is computed from the $W_0(\cdot)$ branch. Plugging the relative prior variance g_* from (24) into the Bayes factor from (1), we obtain the expression for the sceptical Bayes factor in (7).

D Bayes factor with truncated advocacy prior

For now assume $\hat{\theta}_o > 0$. The marginal likelihood of the replication effect estimate $\hat{\theta}_r \,|\, \theta \sim \mathrm{N}(\theta, \sigma_r^2)$ under the truncated advocacy prior $H_{\mathrm{A}'}\colon \theta \sim \mathrm{N}(\hat{\theta}_o, \sigma_o^2)\mathbb{1}_{(0,\infty)}(\theta)$ is

$$
\begin{aligned}
f(\hat{\theta}_r \,|\, H_{\mathrm{A}'}) &= \int_{-\infty}^{+\infty} f(\hat{\theta}_r \,|\, \theta) f(\theta \,|\, H_{\mathrm{A}'}) \,\mathrm{d}\theta \\
&= \int_{-\infty}^{+\infty} \frac{\mathbb{1}_{(0,\infty)}(\theta)}{2\pi\sigma_r\sigma_o\Phi(z_o)} \exp\left\{-\frac{1}{2}\left[\frac{(\hat{\theta}_o - \theta)^2}{\sigma_r^2} + \frac{(\theta - \hat{\theta}_o)^2}{\sigma_o^2}\right]\right\} \mathrm{d}\theta \\
&= \frac{1}{2\pi\sigma_r\sigma_o\Phi(z_o)} \exp\left\{-\frac{1}{2}\frac{(\hat{\theta}_r - \hat{\theta}_o)^2}{\sigma_r^2 + \sigma_o^2}\right\} \underbrace{\int_0^{+\infty} \exp\left\{-\frac{1}{2}\frac{(\theta - \frac{\hat{\theta}_o/\sigma_o^2 + \hat{\theta}_r/\sigma_r^2}{1/\sigma_o^2 + 1/\sigma_r^2})^2}{(1/\sigma_o^2 + 1/\sigma_r^2)^{-1}}\right\} \mathrm{d}\theta}_{= \sqrt{\frac{2\pi}{1/\sigma_o^2 + 1/\sigma_r^2}}\,\Phi\left(\frac{z_o(1+dc)}{\sqrt{1+c}}\right)} \\
&= \frac{1}{\sqrt{2\pi(\sigma_r^2 + \sigma_o^2)}} \exp\left\{-\frac{1}{2}\frac{(\hat{\theta}_r - \hat{\theta}_o)^2}{\sigma_r^2 + \sigma_o^2}\right\} \frac{\Phi\left(\frac{z_o(1+dc)}{\sqrt{1+c}}\right)}{\Phi(z_o)}. \qquad (25)
\end{aligned}
$$

With a similar argument one can show that this result holds for any $\hat{\theta}_o$ if the last factor in (25) is changed to

$$
\frac{\Phi\left\{\mathrm{sign}(z_o)\frac{z_o(1+dc)}{\sqrt{1+c}}\right\}}{\Phi\left\{|z_o|\right\}}.
$$

By dividing the marginal likelihood of the replication data under the sceptical prior by the marginal likelihood under the truncated advocacy prior, the Bayes factor in (15) is obtained.

E The shrinkage paradox

We want to investigate what happens to the replication success regions as the relative variance c and the squared original z-value z_o^2 (a monotone transformation of the original minimum Bayes factor minBF_o) become larger. Ignoring the success regions on the wrong side of zero (due to the replication paradox), the minimum relative effect estimates $d_{\min}$ as shown in Section 3 are given by

$$
\begin{aligned}
d^{\text{BF}_\text{S}}_{\text{min}} &= \frac{1/c + g_\gamma}{g_\gamma - 1} + \sqrt{\frac{\log\left[\frac{1/c+1}{(1/c+g_\gamma)(1+g_\gamma)}\right]/z_o^2 + \frac{g_\gamma}{1+g_\gamma} + \frac{1}{1-g_\gamma}}{(1-g_\gamma)/\left[(1/c+g_\gamma)(1/c+1)\right]}} && \text{(sceptical Bayes factor)} \\
d^{\text{2TR}}_{\text{min}} &= \frac{z_\gamma}{z_o\sqrt{c}} && \text{(two-trials rule)} \\
d^{\text{BF}_\text{R}}_{\text{min}} &= \sqrt{\left[1 + \frac{\log(1+c) - 2\log\gamma}{z_o^2}\right]\frac{1/c+1}{c}} - \frac{1/c+1}{1+c} && \text{(replication Bayes factor)} \\
d^{p_\text{S}}_{\text{min}} &= \sqrt{\frac{1/c + 1/(z_o^2/z_\gamma^2 - 1)}{z_o^2/z_\gamma^2}} && \text{(sceptical } p\text{-value)}
\end{aligned}
$$

where for the sceptical Bayes factor it was assumed that $g_\gamma > 1$ (otherwise the plus before the square root term needs to be replaced by a minus).

For the sceptical Bayes factor, we obtain

$$
\lim_{c\to\infty} d^{\text{BF}_\text{S}}_{\text{min}} = \frac{g_\gamma}{g_\gamma - 1} + \sqrt{\frac{\log\left[\frac{1}{g_\gamma(1+g_\gamma)}\right]/z_o^2 + \frac{g_\gamma}{1+g_\gamma} + \frac{1}{1-g_\gamma}}{(1-g_\gamma)/g_\gamma}} \quad \text{and} \quad \lim_{z_o^2\to\infty} d^{\text{BF}_\text{S}}_{\text{min}} = \sqrt{\frac{1/c+1}{c}} - \frac{1}{c}
$$

where for the second limit we used that $\lim_{z_o^2\to\infty} g_\gamma = 0$ for a fixed level γ (Appendix A). So the sceptical Bayes factor does not suffer from any form of the shrinkage paradox. The limits for the two-trials rule are given by

$$
\lim_{c\to\infty} d^{\text{2TR}}_{\text{min}} = 0 \qquad \text{and} \qquad \lim_{z_o^2\to\infty} d^{\text{2TR}}_{\text{min}} = 0
$$

so the two-trials rule suffers from both forms of the shrinkage paradox. For the sceptical p-value, we obtain

$$
\lim_{c\to\infty} d^{p_\text{S}}_{\text{min}} = \sqrt{\frac{z_\gamma^2}{z_o^2(z_o^2/z_\gamma^2 - 1)}} \qquad \text{and} \qquad \lim_{z_o^2\to\infty} d^{p_\text{S}}_{\text{min}} = 0
$$

thus, the sceptical p-value suffers from the shrinkage paradox at original. Finally, the limits for the replication Bayes factor are

$$
\lim_{c\to\infty} d^{\text{BF}_\text{R}}_{\text{min}} = 0 \qquad \text{and} \qquad \lim_{z_o^2\to\infty} d^{\text{BF}_\text{R}}_{\text{min}} = \sqrt{\frac{1/c+1}{c}} - \frac{1/c+1}{1+c}
$$

which means that the replication Bayes factor suffers from the shrinkage paradox at replication.

F Probability of replication success with the sceptical Bayes factor

Conditional on the original study, the probability for replication success at level γ is given by the probability of (8). This event involves $z_r = dz_o\sqrt{c}$ as the only random quantity if the

original study has been completed. Assuming a normal distribution

$$z_r \,|\, z_o, c \sim \mathrm{N}(\mu_{z_r}, \sigma^2_{z_r})$$

which may depend on z_o and c encompasses the typical scenarios under which one would want to compute the probability for replication success. For example, under the null hypothesis ($H_0\colon \theta = 0$), we have $\mu_{z_r} = 0$ and $\sigma^2_{z_r} = 1$. For conditional power we assume the underlying effect size equals the original effect estimate ($\theta = \hat{\theta}_o$) and therefore $\mu_{z_r} = z_o\sqrt{c}$ and $\sigma^2_{z_r} = 1$. Finally, predictive power is obtained by using the predictive distribution based on the advocacy prior ($H_\mathrm{A}\colon \theta \sim \mathrm{N}(\hat{\theta}_o, \sigma^2_o)$) and thus $\mu_{z_r} = z_o\sqrt{c}$ and $\sigma^2_{z_r} = 1 + c$.

Applying some algebraic manipulations to (8), the probability for replication success at level γ can be computed by

$$\Pr\left(\mathrm{BF_S} \leq \gamma \,|\, z_o, c\right) = \begin{cases} \Pr(\chi^2_{1,\lambda} \geq A/[B\sigma^2_{z_r}]) & \text{for } g_\gamma < 1 \\ \Phi(\mathrm{sign}(z_o)\left\{\mu_{z_r} - D\right\}/\sigma_{z_r}) & \text{for } g_\gamma = 1 \\ \Pr(\chi^2_{1,\lambda} \leq A/[B\sigma^2_{z_r}]) & \text{for } g_\gamma > 1 \end{cases} \tag{26}$$

with non-centrality parameter $\lambda = (\mu_{z_r} - M)^2/\sigma^2_{z_r}$ and

$$\begin{aligned} A &= \log\left\{\frac{1/c + 1}{(1/c + g_\gamma)(1 + g_\gamma)}\right\} + z_o^2\left\{\frac{g_\gamma}{1 + g_\gamma} + \frac{1}{1 - g_\gamma}\right\}, \\ B &= \frac{1 - g_\gamma}{(1 + cg_\gamma)(1/c + 1)}, \\ D &= \frac{z_o^2\{1/2 + 1/(1/c + 1)\} - \log 2}{2z_o\sqrt{c}}(1 + c) \\ M &= \frac{z_o(1 + cg_\gamma)}{\sqrt{c}(g_\gamma - 1)}. \end{aligned}$$

The probability is zero, if the original z-value $|z_o|$ is not large enough such that the sufficiently sceptical relative prior variance g_γ can be computed for level γ with (3).

G Probability of replication success with the replication Bayes factor

The probability of $\mathrm{BF_R} \leq \gamma$ is equivalent to the probability of

$$\log(1 + c) - z_r^2 + \frac{(z_r - z_o\sqrt{c})^2}{1 + c} \leq 2\log\gamma \tag{27}$$

Applying some algebraic manipulations to (27) and assuming a normal distribution for z_r as in Appendix F leads to

$$\Pr\left(\mathrm{BF_R} \leq \gamma \,|\, z_o, c\right) = \Pr\left(\chi^2_{1,\lambda} \geq \left\{z_o^2 + \log(1 + c) - \log\gamma^2\right\}[1 + 1/c]/\sigma^2_{z_r}\right)$$

with non-centrality parameter $\lambda = (\mu_{z_r} + z_o/\sqrt{c})^2/\sigma^2_{z_r}$.

PAPER II

The assessment of replication success based on relative effect size

Leonhard Held, Charlotte Micheloud, Samuel Pawel

The Annals of Applied Statistics, 2022, 16(2), 706–720.
https://doi.org/10.1214/21-AOAS1502.

PAPER III

Bayesian approaches to designing replication studies

Samuel Pawel, Guido Consonni, Leonhard Held

2022. arXiv preprint. https://doi.org/10.48550/arXiv.2211.02552.

PAPER IV

Reverse-Bayes methods for evidence assessment and research synthesis

Leonhard Held, Robert Matthews, Manuela Ott, Samuel Pawel

Research Synthesis Methods, 2022, 13(3), 295–314.
https://doi.org/10.1186/s12874-022-01635-4.

PAPER V

Comment on "Bayesian additional evidence for decision making under small sample uncertainty"

Samuel Pawel, Leonhard Held, Robert Matthews

BMC Medical Research Methodology, 2022, 22(149).

https://doi.org/10.1002/jrsm.1538.

PAPER VI

Pitfalls and Potentials in Simulation Studies

Samuel Pawel, Lucas Kook, Kelly Reeve

2022. arXiv preprint. https://doi.org/10.48550/arXiv.2203.13076.