

# **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 credibility of a scientific finding is based on the fact that it can be replicated when the same study is repeated with new subjects. However, various failures to replicate major scientific findings in the social and life sciences indicate that replicability 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 increase in replication studies, there is no consensus on a fundamental question: How should replication success be defined statistically?

The first and largest part of this thesis extends a recently proposed reverse-Bayes method for quantifying replication success. The key idea is to challenge the data from the original study with a sceptical prior so that the resulting posterior no longer provides evidence for the finding from the study. While the proposed method defined evidence through tail probabilities, the extension developed here replaces them with Bayes factors. We find that the extended method combines two notions of replicability: It requires that original and replication study independently provide evidence for an effect and ensures that their effect estimates are compatible. The extended method also avoids a statistical paradox which the proposed method suffered from. Another extension proposed in this thesis recalibrates the proposed method to produce more appropriate inferences in terms of effect size. The recalibration is chosen such that for original studies which were borderline significant, replication success can only be achieved if the replication effect estimate is larger than the original one. The recalibrated method has good frequentist properties comparable to standard methods used in practice. The third and last extension provides a framework for Bayesian design of replication studies which allows for combining data from the original study with external knowledge. External knowledge, such as data from other studies, is ubiquitous in the replication setting and taking it into account can potentially lead to more efficient designs compared to classical approaches.

The second part of the thesis is concerned with reverse-Bayes methodology. 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. The review includes also some new results on connections between reverse-Bayes methods and meta-analysis. Finally, a commentary paper 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 are 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

<b>Preface</b>	<b>ix</b>
<b>Introduction</b>	<b>1</b>
1 Replication studies.....	1
2 Open questions and thesis contributions .....	14
Data and software .....	18
Bibliography .....	18

## Design and analysis of replication studies

---

<b>Paper I</b>	<b>25</b>
----------------	-----------

### **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>.

<b>Paper II</b>	<b>63</b>
-----------------	-----------

### **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>.

<b>Paper III</b>	<b>85</b>
------------------	-----------

### **Bayesian approaches to designing replication studies**

*Samuel Pawel, Guido Consonni, Leonhard Held*

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

## Reverse-Bayes methodology

---

<b>Paper IV</b>	<b>127</b>
-----------------	------------

### **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

161

### **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>.

## **Integrity in methodological research**

---

## Paper VI

169

### **Pitfalls and Potentials in Simulation Studies**

*Samuel Pawel, Lucas Kook, Kelly Reeve*

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



---

## 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 thanks 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, and 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, and Ting. I also want to thank my good friends from outside of academia (in alphabetical order): Chronis, Dani, Eleftheria, Flo, Fabi, Giuachin, Mirela, and Peter. A special thanks goes to Ada. Finally, I thank my family: Beni, Christa, Elisabeth, Fabian, Harry, Markus, Maja, and Noemi for their support.

Zürich, December 2022

Samuel Pawel



---

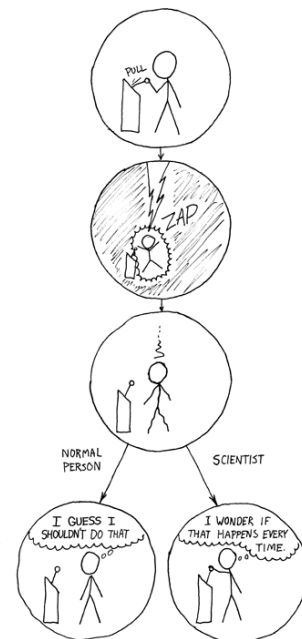
# Introduction

---

## 1 Replication studies

How can we know if a finding from a scientific 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 such questions is of high importance to scientists and decision makers. However, because any study result comes with uncertainty we can never know for sure.

One way to come a little closer to the truth is to repeat the original study with new subjects. Such a *replication study* may yield similar results which would make us more confident about the original finding, or 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 receive the recommendation to get vaccinated (a consequence that anyone who experienced the COVID-19 pandemic is very aware of).



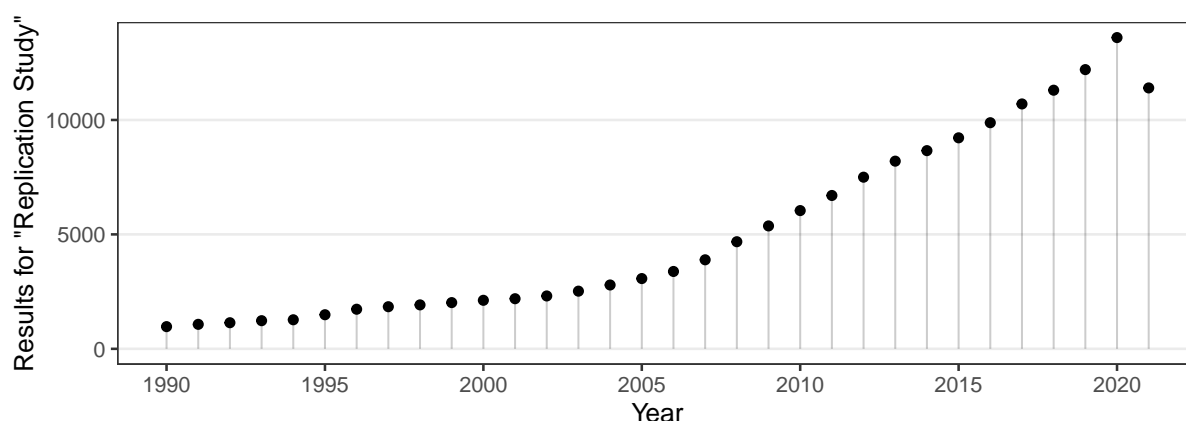
Replication studies as illustrated by Randall Munroe (<https://xkcd.com/242/>).

---

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 not considered “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 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 better 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), and 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 prevent an escalation of the situation and save the reputation of science (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. Several academic journals and funders are now explicitly promoting replication research ([NWO, 2016](#); [NSF, 2018](#); [Nature Communications, 2022](#)). Figure 1 illustrates this trend via the yearly number of results for the search term “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 considered successful? For this reason, replication researchers typically report the results from different methods for assessing replication success. For example, the [Open Science Collaboration \(2015\)](#) states “[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 an original and a 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 the normality assumption.

	Original study	Replication study
effect estimate	$\hat{\theta}_o$	$\hat{\theta}_r$
standard error	$\sigma_o$	$\sigma_r$
$(1 - \alpha)$ 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)	$p_o = 2\{1 - \Phi( z_o )\}$	$p_r = 2\{1 - \Phi( z_r )\}$

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 calculated from the others. Using summary statistics is also often the only possible way for conducting the replication success 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  $\theta$  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 type of effect size, a transformation might be required so that the assumption of approximately normally distributed effect estimates around the unknown effect size  $\theta$  (for large enough sample sizes) is justified. 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  $\sigma$  represents the statistical uncertainty of the estimate. Under the assumption of (asymptotic) normality, confidence intervals for  $\theta$  and  $p$ -values for testing the null hypothesis  $H_0: \theta = 0$  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 $\alpha$ 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 $\alpha$ ( $p_m = 2\{1 - \Phi( \hat{\theta}_m /\sigma_m)\} < \alpha$ ) 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 as the original one and is at least $d_{\min}$ times as large ( $d = \hat{\theta}_r/\hat{\theta}_o \geq d_{\min}$ ). Usually $d_{\min} = 1$ .
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 $p_Q = 2\{1 - \Phi(\sqrt{Q})\}$ with Q-statistic $Q = \sum_{i \in \{o,r\}} (\hat{\theta}_i - \hat{\theta}_m)^2/\sigma_i^2 = (\hat{\theta}_o - \hat{\theta}_r)^2/(\sigma_o^2 + \sigma_r^2)$ .

Table 2 lists commonly used criteria for replication success in terms of the summary statistics from Table 1. The most popular criterion is listed first and 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 regulation or “vote-counting” in meta-analysis (Cooper et al., 2019). The cri-

---

terion 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 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 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.<sup>1</sup> 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 primarily been used if more than one replication study of the same original study are conducted since replicators are often interested in 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 the original and the 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, and the smaller  $d$  is 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  $d_{\min} = 1$  to define replication success (Errington et al., 2021). 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 takes into account both sources of uncertainty (Patil et al., 2016). Yet, 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  $\alpha$  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 ( $H_0: \theta_o = \theta_r$ ), 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.

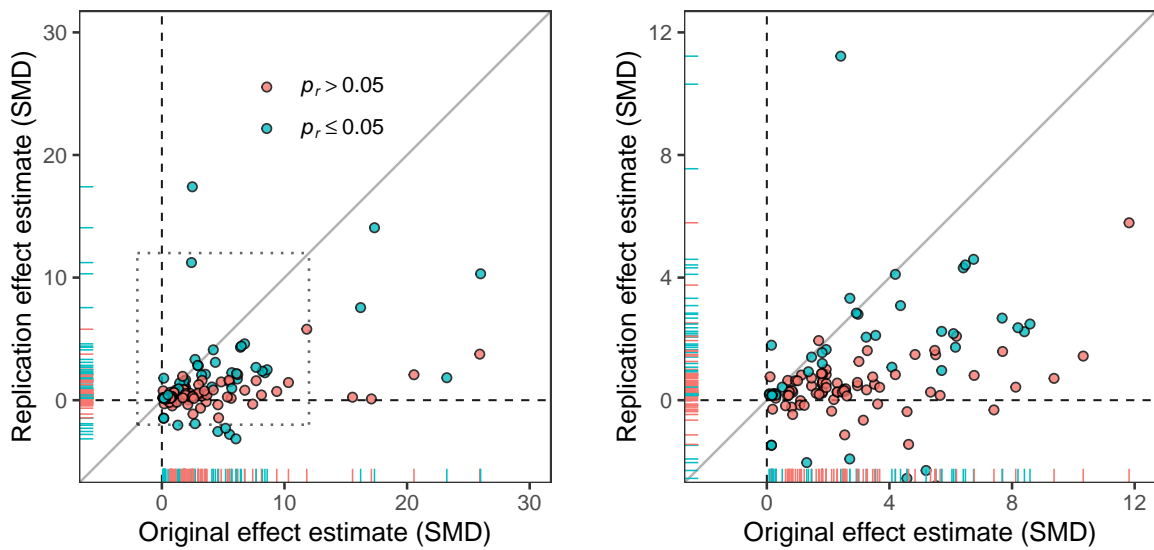
---

<sup>1</sup>The meta-analytic approach can also be given a Bayesian interpretation: A flat prior distribution for the underlying effect size  $\theta$  is updated by the data from original and replication study. Replication success defined by a meta-analytic  $p$ -value being smaller than  $\alpha$  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 are provided with original and replication standardized mean difference effect estimates along with standard errors, and only they will be used in the subsequent analyses.

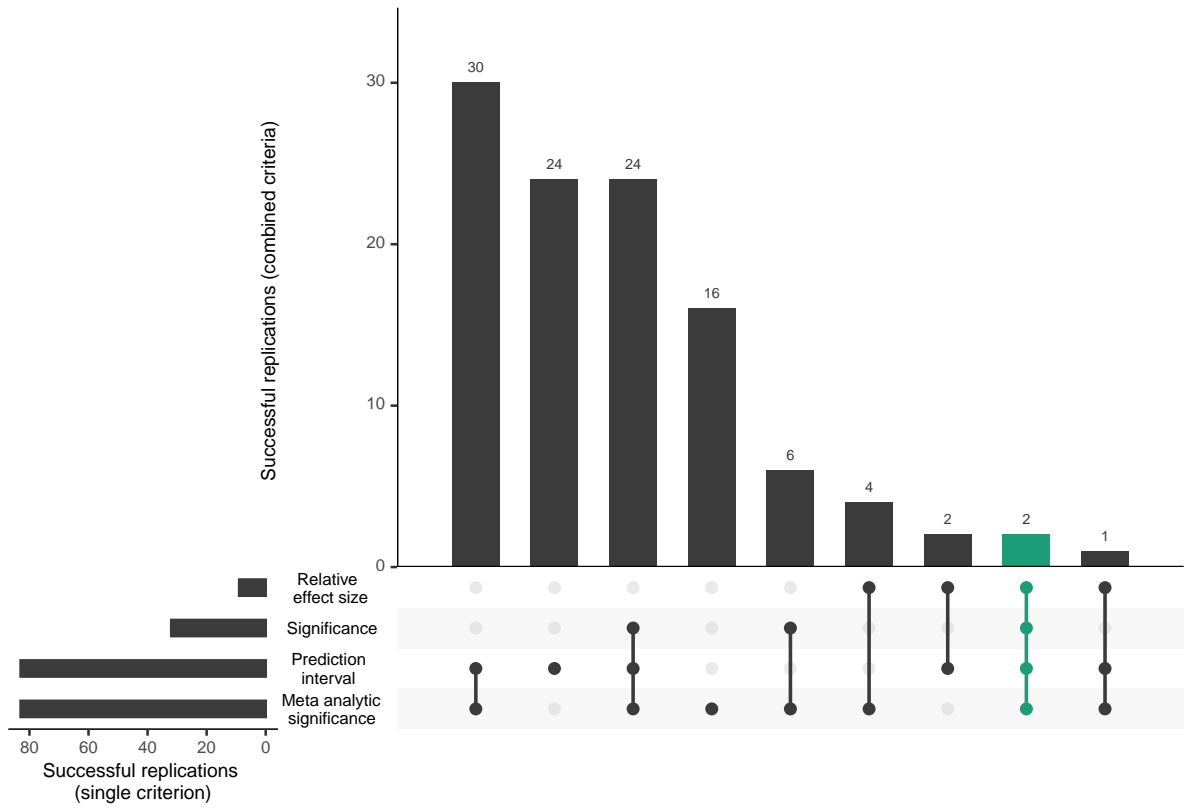


**Figure 2:** Results for 132 effects from the Reproducibility Project: Cancer Biology (Errington et al., 2021) for which effect estimates and standard errors are available on standardized mean difference (SMD) scale. The right plot shows a zoomed-in view of the dotted area in the left plot. The  $p$ -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 (except the confidence interval criteria since they do not take into account the uncertainty from both studies). For the total 132 replications, most successes occur for the meta-analytic significance (83) and the prediction interval criteria (83), followed by significance (32),





**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 threshold of  $\alpha = 5\%$  and a relative effect size threshold of  $d_{\min} = 1$  are used.

and relative effect size (9). However, looking at the combinations of the criteria, only in two replications are all criteria satisfied simultaneously (green).

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 simplification of the matter.

**Table 3:** Subset of results from the Reproducibility Project: Cancer Biology (Errington et al., 2021). (P, X, E) denotes effect number E from experiment number X, from original paper number 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  $d_{\min} = 1$  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 extend 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

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

meaning that the *prior distribution* for the parameter  $\theta$  with probability density/mass function  $f(\theta)$  is multiplied by the (normalized) likelihood of the data, also known as *Bayesian updating*. Parameter values which increase the likelihood of the data become more likely *a posteriori* but they are also weighted by their *a priori* plausibility 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.

Many consider the prior to be also the weak point of Bayesian inference since 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 reverse Bayes' theorem

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

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. For example, the posterior may be specified to indicate evidence for/against a particular hypothesis under investigation. Reverse-Bayes inference then revolves around the question of 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  $\theta$  based on the effect estimate from the original study  $\hat{\theta}_o$  and its standard error  $\sigma_o$ . We will assume that  $\hat{\theta}_o$  is normally distributed

around the unknown effect size  $\theta$  with (known) variance equal to its squared standard error  $\sigma_o^2$ , here and henceforth denoted by  $\hat{\theta}_o | \theta \sim N(\theta, \sigma_o^2)$ . Furthermore, we specify a zero-mean normal prior with variance  $\tau^2$  for the effect size  $\theta$ , representing the position of a sceptic who does not believe in the presence of a genuine effect. The “stubbornness” of the sceptic is determined by how large the variance  $\tau^2$  is chosen. Combining the sceptical prior with the likelihood produces a posterior which is again normal  $\theta | \hat{\theta}_o, \sigma_o \sim N(\mu_{\text{post}}, \sigma_{\text{post}}^2)$  with mean and variance

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

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

$$[\mu_{\text{post}} \pm z_{\alpha/2} \sigma_{\text{post}}] \quad (1)$$

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<sup>2</sup> for a genuine effect found in the original study.

Depending on how large the prior variance  $\tau^2$  is chosen, the posterior credible interval (1) will either include or exclude zero. Different researchers may have different degrees of scepticism and may thus choose different prior variances  $\tau^2$ . 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*  $\tau_\alpha^2$  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  $\alpha$ .

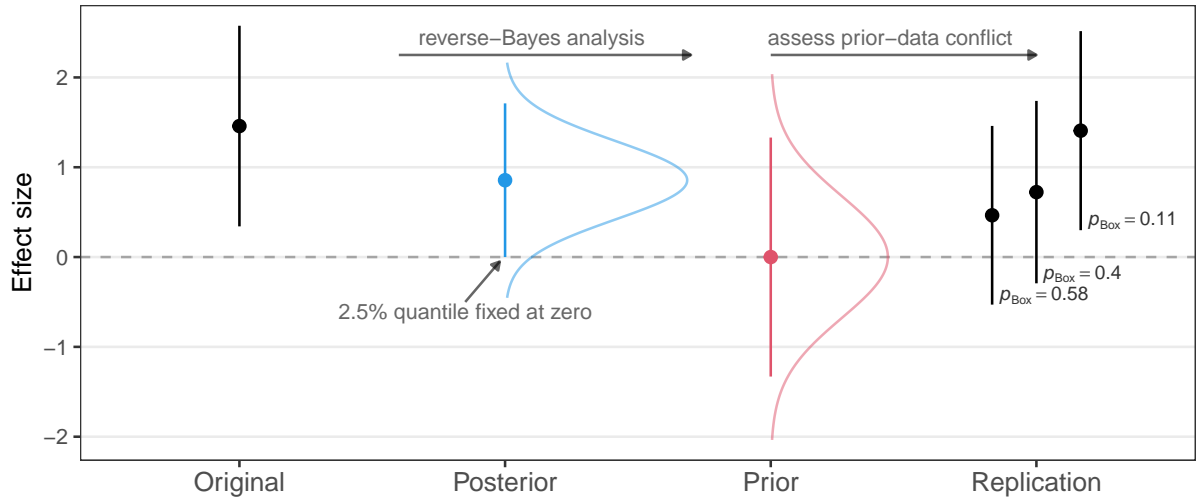
Figure 4 illustrates the derivation of the sufficiently sceptical prior variance for an original study included in the Reproducibility Project: Cancer Biology with standardized mean difference effect estimate  $\hat{\theta}_o = 1.46$  and standard error  $\sigma_o = 0.57$ . We see that the sufficiently sceptical prior with variance  $\tau_\alpha^2 = 0.68^2$  produces a posterior with  $1 - \alpha = 95\%$  credible interval fixed to zero, so that the original finding is rendered no longer convincing at level  $\alpha = 5\%$ .

[Held \(2019a\)](#) showed that the sufficiently sceptical prior variance  $\tau_\alpha^2$  for a level  $\alpha$  is available in closed-form

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

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  $\alpha$ . 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 that no scepticism is required to make them unconvincing.

<sup>2</sup>For readers who do not agree with this notion of evidence, do not panic. In this thesis (Paper I and IV) we 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  $\alpha = 5\%$ . Prior predictive  $p$ -values  $p_{\text{Box}}$  are computed for quantifying prior-data conflict.

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, the question becomes whether it is plausible in light of external data. A natural candidate for answering the question is 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  $\theta$  and data is adequate, Box gave the following fundamental decomposition of the joint distribution

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

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

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

is compatible with the observed data. If the model  $M$  was indeed adequate, the empirical distribution of the observed data should be close to its 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  $f(\text{data} | \theta, M)$  and/or misspecification of the prior  $f(\theta | M)$ .

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*

$$p_{\text{Box}} = \Pr \{f(\text{data} | M) < f(\text{observed data} | M)\}. \quad (3)$$

The  $p$ -value  $p_{\text{Box}}$  is the probability of obtaining data with lower predictive density (“more surprising” data) than the observed data, and the lower  $p_{\text{Box}}$ , the more incompatibility between the observed data and the assumed model M. This approach was used by Held (2020). 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*

$$\text{PPR}_{\text{Box}} = \frac{f(\text{observed data} | M_1)}{f(\text{observed data} | M_2)},$$

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 prior predictive ratio  $\text{PPR}_{\text{Box}}$ , the less compatible the observed data with the model  $M_1$ . Bayesian model criticism approaches based on prior predictive ratios will be used in latter parts of this thesis.<sup>3</sup>

We now return to the replication setting. Having obtained a sceptical prior  $\theta \sim N(0, \tau_\alpha^2)$  with sufficiently sceptical prior variance  $\tau_\alpha^2$  from (2), the aim is 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 scepticism regarding the original finding is unjustified and that the original study provided evidence for a genuine effect. Under the assumption of a normal likelihood for the replication effect estimate, i. e.,  $\hat{\theta}_r | \theta \sim N(\theta, \sigma_r^2)$ , the prior predictive distribution is given by  $\hat{\theta}_r | \theta \sim N(0, \sigma_r^2 + \tau_\alpha^2)$ . As the prior predictive distribution is symmetric around zero, the prior predictive  $p$ -value (3) is

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

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 smaller prior predictive  $p$ -values  $p_{\text{Box}}$ , indicating more prior-data conflict. This is because the standard errors from all three replications are roughly the same size, so that the distance between zero and the replication effect estimate matters most. The  $p$ -values suggest that there

---

<sup>3</sup>Some readers may have noted that the prior 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 later (Paper I and III).

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  $p_{\text{Box}} = 0.11$ ).

Held (2020) defined replication success at level  $\alpha$  by

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

In words, 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  $\alpha$ , Held (2020) proposed to determine the smallest level at which replication success can be established. This level is called the *sceptical  $p$ -value*  $p_S$ , and available in closed-form

$$p_S = 2 \{1 - \Phi(|z_S|)\}$$

where

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

with arithmetic mean  $z_A^2 = (z_o^2 + z_r^2)/2$  and harmonic mean  $z_H^2 = 2/(1/z_o^2 + 1/z_r^2)$  of the squared  $z$ -statistics, and variance ratio  $c = \sigma_o^2/\sigma_r^2$ . Replication success at level  $\alpha$  is then equivalent to  $p_S \leq \alpha$ . For instance, the sceptical  $p$ -values of the three replication studies in Figure 4 are  $p_S = 0.39$ ,  $p_S = 0.23$ , and  $p_S = 0.075$  (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 in this manner, 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 ( $p_S > \max\{p_o, p_r\}$ ), meaning that both  $p$ -values have to be smaller than  $\alpha$  such that replication success based on  $p_S \leq \alpha$  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 ( $p_S \uparrow 1$  as  $d \downarrow 0$  for fixed  $p_o$  and  $p_r$ ). Because of the first property, the sceptical  $p$ -value requires *both* studies to be sufficiently convincing on their



---

own (in terms of their  $p$ -values), similar to the significance criterion for replication success. However, unlike the significance criterion, 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 becomes larger so indicates less replication success due to the second property. This is 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 the replication study also 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 smaller). 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. Some additional properties (e. g., the null distribution of the sceptical  $p$ -value) and extensions (e. g., a one-sided version of the sceptical  $p$ -value, power and sample size calculations) can be found in the original article by Held (2020). However, the relevant parts of the procedure have been presented. In the following, I will discuss open questions and how they are addressed in this thesis.

## 2 Open questions and thesis contributions

This thesis consists of six papers divided into three parts. The first part (Paper I, II, and III) focuses on methodology for design and analysis of replication studies. The second part (Paper IV and V) revolves around reverse-Bayes methodology in general. The last part (Paper VI) is about integrity issues in methodological research with simulations studies.

### 2.1 Design and analysis of replication 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 there is no longer evidence for an effect. Evidence is quantified in terms of credible intervals and 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 extend the reverse-Bayes assessment of replication studies to use Bayes factors (Good, 1958; Jeffreys, 1961) for quantifying evidence and prior-data conflict. Similar 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 has similar properties as the sceptical  $p$ -value due to the reverse-Bayes approach underlying both methods (e. g., it penalizes shrinkage of the replication effect estimate and requires compelling evidence from both studies). However, it is also shown that the sceptical  $p$ -value 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 the R package BayesRep (<https://gitlab.uzh.ch/samuel.pawel/BayesRep>).

The idea to use Bayes factors instead of tail probabilities was suggested by Consonni (2019) and Pericchi (2020) independently in response to the 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. 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 annual meeting of the GMDS and CEN-IBS (German Association for Medical Informatics, Biometry and Epidemiology and the Central European Network of the International Biometric Society) in 2020. L. Held presented the final results at the ISBA (International Society for Bayesian Analysis) world meeting 2021.

## **Paper 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 or 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*  $\alpha_G$  (named “golden” because the golden ratio appears in its derivation). The golden level is defined through the property that for an original study which was 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 for replication studies with equal or larger sample size than the original study, 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 (Open Science Collaboration, 2015; Camerer et al., 2016, 2018; Cova et al., 2018) illustrate the properties of the method. This extension is now implemented as the default option in the R package `ReplicationSuccess` available on CRAN (<https://cran.r-project.org/package=ReplicationSuccess>).

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.

### **Paper 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 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 this 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 (Protzko et al., 2020) are used for illustrating our methods, which are also available in the R package `BayesRepDesign` (<https://github.com/SamCH93/BayesRepDesign>).

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. In Paper I, 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.

## 2.2 Reverse-Bayes methodology

### **Paper 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 methods, 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 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.

### **Paper 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. – 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 to 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 alerted 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 by expanding it with some logical flaws of the method.

---

The two of us iteratively worked on the manuscript, while L. Held gave mostly high-level feedback.

## 2.3 Integrity in methodological research

### **Paper 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 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 of these issues. We summarize possible questionable research practices in simulation studies, and show how easy it is to make a method seem superior if various questionable research practices are employed. We also give recommendations to alleviate these issues. Most importantly we recommend researchers write and pre-register simulation protocols.

The first authorship of this manuscript is shared with 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 `ainet` (<https://github.com/LucasKook/ainet>) and simulation study code. Myself, L. Kook, and K. Reeve then designed the final simulation study and wrote the manuscript together. Recently, I was invited to present the results from this project at the CEN (Central European Network) conference 2023 in Basel.

## **Data and software**

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”. All analyses from this chapter 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. Code and data to reproduce this thesis are available at <https://doi.org/10.5281/zenodo.XXXXXXX>.

---

## Bibliography

- 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/>.
- Altman, D. G. (1994). The scandal of poor medical research. *BMJ*, 308(6924):283–284. doi:[10.1136/bmj.308.6924.283](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1037/met0000051).
- Begley, C. G. and Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533. doi:[10.1038/483531a](https://doi.org/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](https://doi.org/10.1038/s41562-017-0189-z).
- Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, 18(1). doi:[10.1214/ss/1056397485](https://doi.org/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](https://doi.org/10.2307/2289131).
- Bonett, D. G. (2020). Design and analysis of replication studies. *Organizational Research Methods*, 24(3):513–529. doi:[10.1177/1094428120911088](https://doi.org/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](https://doi.org/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 (General)*, 143:383–430. doi:[10.2307/2982063](https://doi.org/10.2307/2982063).
- 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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1098/rsos.171085).
- Consonni, G. (2019). Sufficiently skeptical intrinsic priors for the analysis of replication studies. 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](https://doi.org/10.7758/9781610448864).
- 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](https://doi.org/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.
- 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>.
- 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](https://doi.org/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](https://doi.org/10.1371/journal.pone.0149794).



- 
- 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](https://doi.org/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](https://doi.org/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](https://doi.org/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. doi:[10.1080/01621459.1958.10501480](https://doi.org/10.1080/01621459.1958.10501480).
- Harms, C. (2019). A Bayes factor for replications of ANOVA results. *The American Statistician*, 73(4):327–339. doi:[10.1080/00031305.2018.1518787](https://doi.org/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](https://doi.org/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](https://doi.org/10.1098/rsos.181534).
- Held, L. (2019b). Research plan “reverse-Bayes design and analysis of replication studies”. URL <https://data.snf.ch/grants/grant/189295>.
- 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](https://doi.org/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*, 13(3):295–314. doi:[10.1002/jrsm.1538](https://doi.org/10.1002/jrsm.1538).
- Held, L., Micheloud, C., and Pawel, S. (2022b). The assessment of replication success based on relative effect size. *The Annals of Applied Statistics*, 16(2):706–720. doi:[10.1214/21-aos1502](https://doi.org/10.1214/21-aos1502).
- 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](https://doi.org/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](https://doi.org/10.1371/journal.pmed.0020124).
- 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](https://doi.org/10.1177/0956797611430953).
-

- 
- 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](https://doi.org/10.1080/01621459.2016.1240079).
- 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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1177/2515245918810225).
- 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., Adolphi, 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](https://doi.org/10.1038/s41562-018-0311-x).
- 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](https://doi.org/10.3758/s13428-018-1092-x).
- 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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1098/rsos.171047).
- 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](https://doi.org/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](https://doi.org/10.1038/s41562-016-0021).
- Nature Communications (2022). Replication studies hold the key to generalization [editorial]. *Nature Communications*, 13(1). doi:[10.1038/s41467-022-34748-x](https://doi.org/10.1038/s41467-022-34748-x).



- 
- 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.
- NSF (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., Consonni, G., and Held, L. (2022a). Bayesian approaches to designing replication studies. doi:10.48550/ARXIV.2211.02552.
- Pawel, S. and Held, L. (2020). Probabilistic forecasting of replication studies. *PLOS ONE*, 15(4):e0231416. doi:10.1371/journal.pone.0231416.
- Pawel, S. and Held, L. (2022). The sceptical Bayes factor for the assessment of replication success. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. doi:10.1111/rssb.12491.
- Pawel, S., Held, L., and Matthews, R. (2022b). Comment on “Bayesian additional evidence for decision making under small sample uncertainty”. *BMC Medical Research Methodology*, 22(1). doi:10.1186/s12874-022-01635-4.
- Pawel, S., Kook, L., and Reeve, K. (2022c). Pitfalls and potentials in simulation studies. doi:10.48550/ARXIV.2203.13076.
- 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., Waliczek, J., and Schooler, J. (2020). High replicability of newly-discovered social-behavioral findings is achievable. doi:10.31234/osf.io/n2a9x.
- 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 12301. doi:10.1136/bmj.12301.
- 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](https://doi.org/10.1177/0956797611417632).
- Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, 26:559–569. doi:[10.1177/0956797614567341](https://doi.org/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](https://doi.org/10.1186/s12874-021-01432-5).
- 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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1037/a0022790).
- Wicherts, J. M. (2011). Psychology must learn a lesson from fraud case. *Nature*, 480(7375):7–7. doi:[10.1038/480007a](https://doi.org/10.1038/480007a).
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer International Publishing. doi:[10.1007/978-3-319-24277-4](https://doi.org/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.

---

**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  $\alpha$  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  $\alpha$  is established if the prior predictive tail probability of the replication estimate is smaller than  $\alpha$ . The smallest level  $\alpha$  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

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

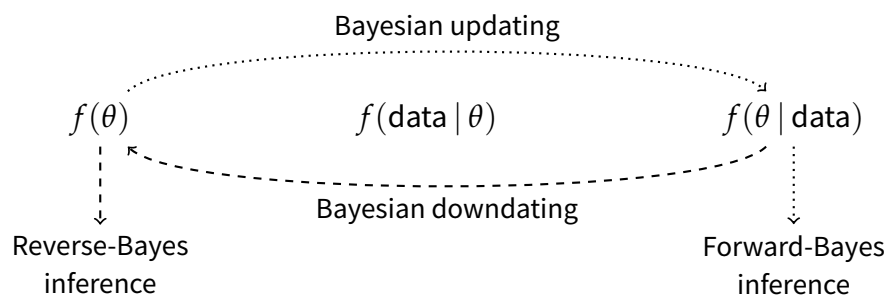
where  $f(x | H_i)$  is the marginal likelihood of the data under  $H_i$  obtained by integrating the likelihood  $f(x | \theta_i)$  with respect to the prior distribution  $f(\theta_i)$  of the model parameters  $\theta_i \in \Theta_i$  with  $i = 1, 2$ . Sometimes we will also write  $\text{BF}_{1:2}(x; \phi')$  to indicate that the Bayes factor is evaluated for a specific value  $\phi'$  of a hyperparameter  $\phi$  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  $\theta$  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  $\sigma_o$  and  $\sigma_r$ , 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  $\kappa$ . 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  $\hat{\theta}_k | \theta \sim N(\theta, \sigma_k^2)$  for  $k \in \{o, r\}$ , i. e., normality of the effect estimates around  $\theta$ , 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  $\theta$ .

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

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

The prior  $\theta | H_S$  represents a *sceptic* who remains unconvinced about the presence of an effect at level  $\gamma$ .

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

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

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

→ The *sceptical Bayes factor*  $\text{BF}_S$  is the smallest level  $\gamma$  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

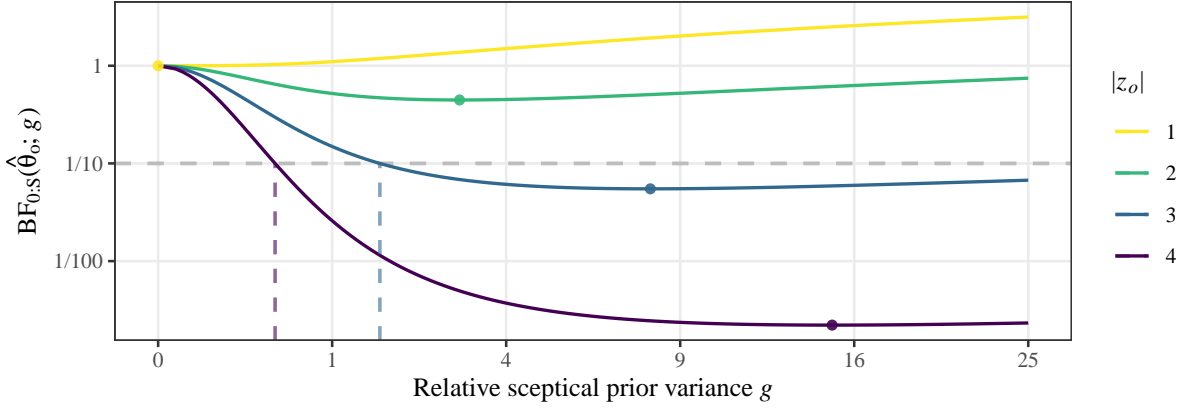
For the effect estimate  $\hat{\theta}_o | \theta \sim N(\theta, \sigma_o^2)$  from the original study consider a hypothesis test comparing the null hypothesis  $H_0: \theta = 0$  to the alternative  $H_S: \theta \neq 0$ . Specification of a prior distribution for  $\theta$  under  $H_S$  is now required 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  $\theta | H_S \sim N(0, \sigma_s^2 = g \cdot \sigma_o^2)$  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

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

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

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

which is reached at  $g_{\min \text{BF}_0} = \max\{0, z_o^2 - 1\}$  (Edwards et al., 1963). Further increasing the relative sceptical prior variance increases (1) indefinitely because of the Jeffreys-Lindley paradox, i. e.,  $\text{BF}_{0:S}(\hat{\theta}_o; g) \rightarrow \infty$  for  $g \rightarrow \infty$  (Bernardo and Smith, 2000, Section 6.1.4). Hence, for a relative sceptical prior variance  $g \in [0, g_{\min \text{BF}_0}]$ , the resulting Bayes factor will be  $\text{BF}_{0:S}(\hat{\theta}_o; g) \in [\min \text{BF}_0, 1]$ .



**Figure 2:** Bayes factor  $\text{BF}_{0:S}(\hat{\theta}_o; g)$  as a function of relative sceptical prior variance  $g$  for different values of  $|z_o| = |\hat{\theta}_o|/\sigma_o$ . Minimum Bayes factors  $\min \text{BF}_0$  are indicated by dots. Dashed vertical lines indicate sufficiently sceptical relative prior variance  $g_\gamma$  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  $\gamma$  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  $g_\gamma \leq g_{\min \text{BF}_0}$  such that  $\text{BF}_{0:S}(\hat{\theta}_o; g_\gamma) = \gamma$ . It can be shown (Appendix A) that  $g_\gamma$  can be explicitly computed by

$$g_\gamma = \begin{cases} -\frac{z_o^2}{q} - 1 & \text{if } -\frac{z_o^2}{q} \geq 1 \\ \text{undefined} & \text{else} \end{cases} \quad (3)$$

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

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  $\theta | H_S \sim N(0, g_\gamma \cdot \sigma_o^2)$  and it can be interpreted as the view of a sceptic who argues that given their prior belief about the effect  $\theta$ , the observed effect estimate  $\hat{\theta}_o$  cannot convince them about the presence of a non-null effect at level  $\gamma$ . 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  $g_\gamma$  needs to be in order to make

the result no longer convincing at level  $\gamma$ . In the most extreme case, when  $|z_o| \rightarrow \infty$  and  $\gamma$  remains fixed, the sufficiently sceptical prior variance will converge to zero (Appendix B). On the other hand, if  $|z_o|$  is not sufficiently large,  $\text{BF}_{0.5}(\hat{\theta}_o; g)$  will either be always increasing in  $g$  (if  $|z_o| \leq 1$ ) or it will reach a minimum above the chosen level  $\gamma$ . In both cases the sufficiently sceptical relative prior variance  $g_\gamma$  is not defined since there is no need to challenge an already unconvincing result.

A side note on the Jeffreys-Lindley paradox is worth being mentioned: If a  $g_\gamma < g_{\min \text{BF}_0}$  exists, there exists also a  $g'_\gamma > g_{\min \text{BF}_0}$  as the Bayes factor monotonically increases in  $g > g_{\min \text{BF}_0}$  and therefore must intersect a second time with  $\gamma$ , due to the paradox. This means that the more compelling the original result, the larger  $g'_\gamma$  needs to be chosen, such that the result becomes no longer convincing at level  $\gamma$ . However, priors which become increasingly diffuse do not represent increasing scepticism but rather increasing ignorance. Using (17) 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  $g_\gamma \leq g_{\min \text{BF}_0}$ ). In principle, the solution  $g'_\gamma > g_{\min \text{BF}_0}$  could also be computed by replacing the  $W_{-1}$  branch of the Lambert  $W$  function in (17) 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  $\hat{\theta}_r | \theta \sim N(\theta, \sigma_r^2)$  under the sceptical prior  $H_S: \theta \sim N(0, g \cdot \sigma_o^2)$  relative to the advocacy prior  $H_A: \theta \sim N(\hat{\theta}_o, \sigma_o^2)$ . The view of an advocate is represented by  $H_A$  since this is the posterior of  $\theta$  given the original estimate and a uniform prior (also the reference prior for this model). The Bayes factor is given by

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

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  $g_\gamma$  is not larger than the corresponding level  $\gamma$  used to define the sufficiently sceptical prior:

$$\text{BF}_{S:A}(\hat{\theta}_r; g_\gamma) \leq \text{BF}_{0.5}(\hat{\theta}_o; g_\gamma) = \gamma, \quad (5)$$

we have established *replication success at level  $\gamma$* .

For example, if we observe  $z_o = 3$  (equivalent to minimum Bayes factor  $\min\text{BF}_o = 1/18$ ) and choose a level  $\gamma = 1/10$  the sufficiently sceptical relative prior variance (17) 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  $\min\text{BF}_r = 1/5.5$  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  $\gamma$ , the described procedure offers an automated way to assess replication success. One way to remove this dependence is to find the smallest level  $\gamma$  where replication success can be established. We thus call this level the *sceptical Bayes factor*

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

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

Figure 3 shows  $\text{BF}_{\text{S:A}}(\hat{\theta}_r; g_\gamma)$  and  $\text{BF}_{0:S}(\hat{\theta}_o; g_\gamma)$  as a function of  $g_\gamma$  for several values of  $z_o$  and  $d$  along with the corresponding  $\text{BF}_S$ . Typically,  $\text{BF}_S$  is given by the height of the intersection between  $\text{BF}_{\text{S:A}}(\hat{\theta}_r; g_\gamma)$  and  $\text{BF}_{0:S}(\hat{\theta}_o; g_\gamma)$  in  $g_\gamma$ . It may also happen that  $\text{BF}_{\text{S:A}}(\hat{\theta}_r; g_\gamma)$  remains below  $\text{BF}_{0:S}(\hat{\theta}_o; g_\gamma)$  for all values of  $g_\gamma$ , in such situations  $\text{BF}_S$  is equal to the original minimum Bayes factor  $\min\text{BF}_o$ . 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  $\gamma$  and hence  $\text{BF}_S$  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  $\text{BF}_S$  is located at the intersection of  $\text{BF}_{\text{S:A}}(\hat{\theta}_r; g)$  and  $\text{BF}_{0:S}(\hat{\theta}_o; g)$  in  $g$ , there is an explicit expression for  $\text{BF}_S$

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

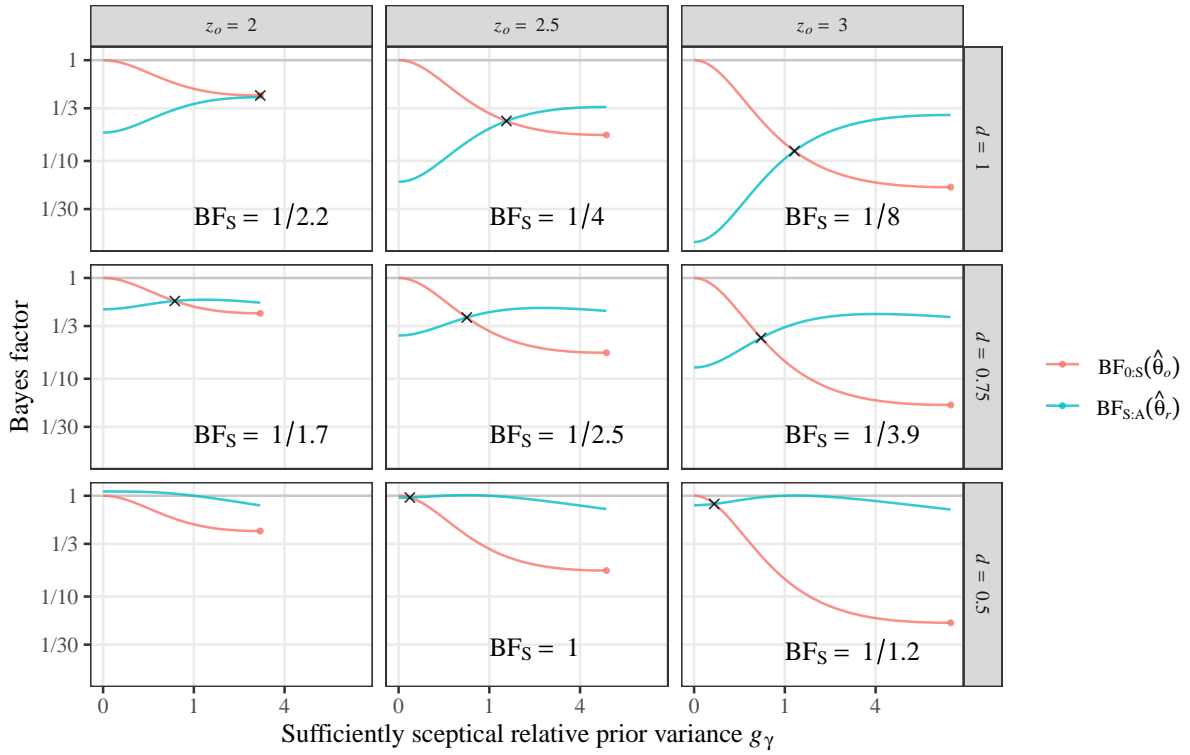
with

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

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  $BF_{S.A}(\hat{\theta}_r; g)$  and  $BF_{0.S}(\hat{\theta}_o; g)$  as a function of the sufficiently sceptical relative prior variance  $g_\gamma$ . In all examples  $c = \sigma_o^2/\sigma_r^2 = 1$ . Minimum Bayes factors  $\min BF_0$  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  $\min BF_0$  (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 ( $\min BF_0$ ), and the change in sample size of the replication compared to the original study ( $c$ ).

The condition for replication success at level  $\gamma$  from (5) is equivalent to

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

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

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

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  $\gamma$  is achieved if and only if the relative effect estimate  $d$  falls within a success region given by

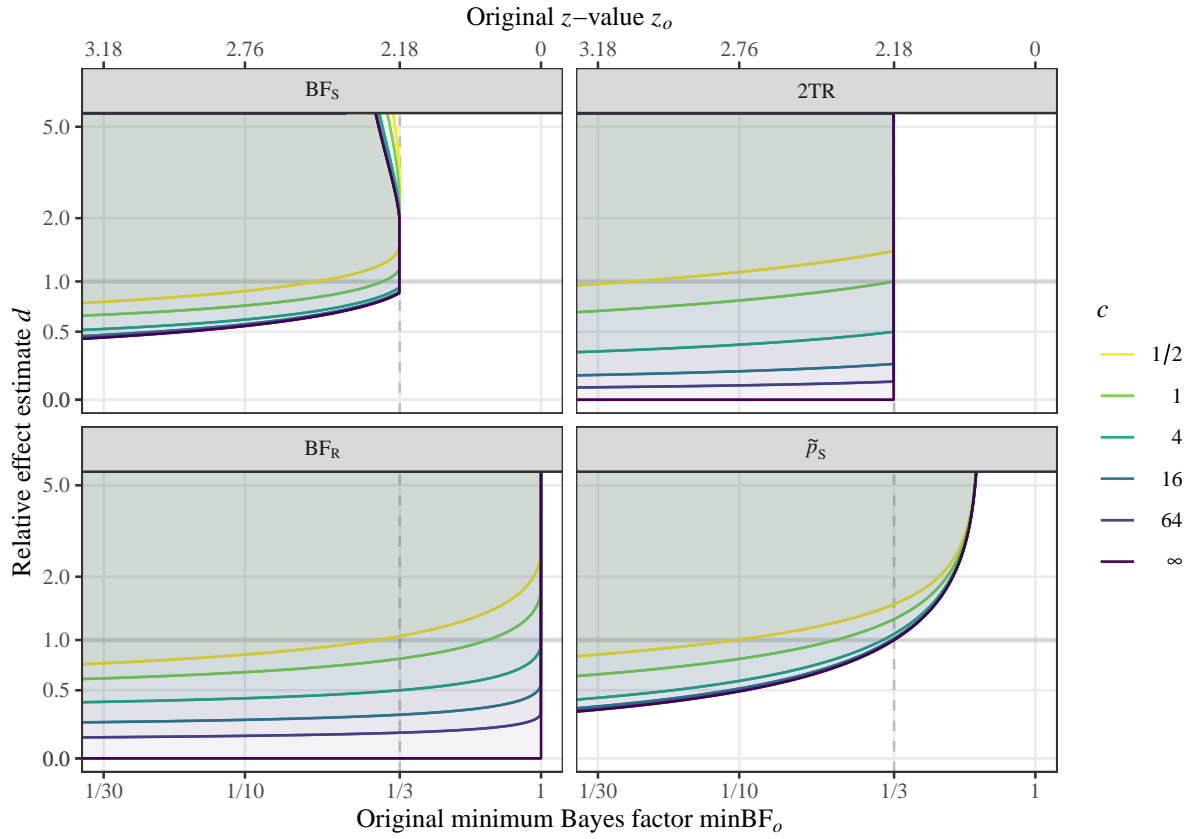
$$\begin{cases} d \notin [M - \sqrt{A/B}, M + \sqrt{A/B}] & \text{if } g_\gamma < 1 \\ d \geq [1 + (1 + 1/c)\{1/2 - \log(2)/z_o^2\}]/2 & \text{if } g_\gamma = 1 \\ d \in [M - \sqrt{A/B}, M + \sqrt{A/B}] & \text{if } g_\gamma > 1 \end{cases} \quad (10)$$

where

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

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  $\min\text{BF}_o$  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  $c \rightarrow \infty$  the methods still requires the replication estimate to be sufficiently large, despite that the evidence against the null becomes overwhelming (since  $|z_r| \rightarrow \infty$  as  $c \rightarrow \infty$ ).

By definition the sceptical Bayes factor can never be smaller than  $\min\text{BF}_o$ , so replication success at level  $\gamma$  is impossible for original studies with  $\min\text{BF}_o > \gamma$ . This property is visible in the top-left plot in Figure 4 by the cut-off at  $\min\text{BF}_o = \gamma = 1/3$ . In contrast, for more convincing original studies with  $\min\text{BF}_o < 1/3$  replication success is possible and two cases can be distinguished in terms of the success region: When  $1/4.5 < \min\text{BF}_o \leq 1/3$  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  $\min\text{BF}_o < 1/4.5$ , 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 ( $\text{BF}_S \leq 1/3$ ), the two-trials rule (2TR:  $\text{minBF}_o \leq 1/3$  and  $\text{minBF}_r \leq 1/3$ ), the replication Bayes factor ( $\text{BF}_R \leq 1/3$ ), and the recalibrated sceptical  $p$ -value ( $\tilde{p}_S \leq 1 - \Phi\{z_\gamma\}$  with  $\gamma = 1/3$ ) as a function of the original minimum Bayes factor  $\text{minBF}_o$  (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  $\gamma$  is established when both  $\min \text{BF}_k \leq \gamma$  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  $\min \text{BF}_o$ . 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  $\min \text{BF}_o = 1/2$ ,  $\min \text{BF}_r = 1/1.5$ , and  $c = 1$  we obtain  $\text{BF}_S = 1/1.9$ ). So for the same level of replication success  $\gamma$  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 unconvincing. Assuming now that  $\min \text{BF}_o < \gamma$ , replication success with the two-trials rule at level  $\gamma$  is achieved if and only if the relative effect estimate is

$$d \geq \frac{z_\gamma}{z_o \sqrt{c}} \quad (11)$$

with  $z_\gamma > 1$  corresponding to  $\min \text{BF} = z_\gamma \exp(-z_\gamma^2/2) \sqrt{e} = \gamma$ . 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  $\min \text{BF}_o = \gamma = 1/3$  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  $H_0: \theta = 0$ , to the alternative that the effect is distributed according to the posterior distribution of  $\theta$  after observing the original data. For the normal model considered so far and if an initial reference prior was chosen, this alternative is also the advocacy prior  $H_A: \theta \sim N(\hat{\theta}_o, \sigma_o^2)$  and therefore the replication Bayes factor is given by

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

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  $\text{BF}_R \downarrow 0$  as  $c \rightarrow \infty$  provided  $z_o \neq 0$  and  $d \neq 0$ . Moreover, we have that

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

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

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

$$d \notin [-\sqrt{J} - H, \sqrt{J} - H] \quad (13)$$

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  $\min \text{BF}_o > 1/3$ .

### 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  $\theta \sim N(0, \tau^2)$ , the sufficiently sceptical prior variance at level  $\alpha$  is then defined as  $\tau^2 = \tau_\alpha^2$  such that the  $(1 - \alpha)$  credible interval for  $\theta$  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  $\hat{\theta}_r | \tau_\alpha^2 \sim N(0, \tau_\alpha^2 + \sigma_r^2)$  is smaller than  $\alpha$ . The smallest level  $\alpha$  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.,  $p_S = 1 - \Phi(z_S)$  with

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

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  $p_S = \Phi(z_S)$ ).

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

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

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  $\alpha$  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  $c$  ( $\lim_{c \rightarrow \infty} d_{\min}$ ) is smaller than the one from the sceptical Bayes factor when  $\min BF_o < 1/5.6$ , while for  $\min BF_o > 1/5.6$  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  $\min BF_o > \gamma$ . This is visible in the bottom-right plot of Figure 4 where the success region has no cut-off at  $\min BF_o = \gamma = 1/3$ , 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  $\min BF_0 = 1/10$  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  $H_{A'}: \theta \sim N(\hat{\theta}_o, \sigma_o^2) \mathbb{1}_{(0, \infty)}(\theta)$ , 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

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

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  $c \rightarrow \infty$ ), but it penalises when the replication estimate goes in the opposite direction ( $d < 0$  and  $c \rightarrow \infty$ ).

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  $\min\text{BF}_o$  (respectively  $z_o$ ), the minimum relative effect estimate  $d_{\min} > 0$  required for replication success at a fixed level  $\gamma$  becomes arbitrarily small as the relative variance  $c$  increases:

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

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  $d_{\min} > 0$  required for replication success at a fixed level  $\gamma$  becomes arbitrarily small as the evidence in the original study increases:

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

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

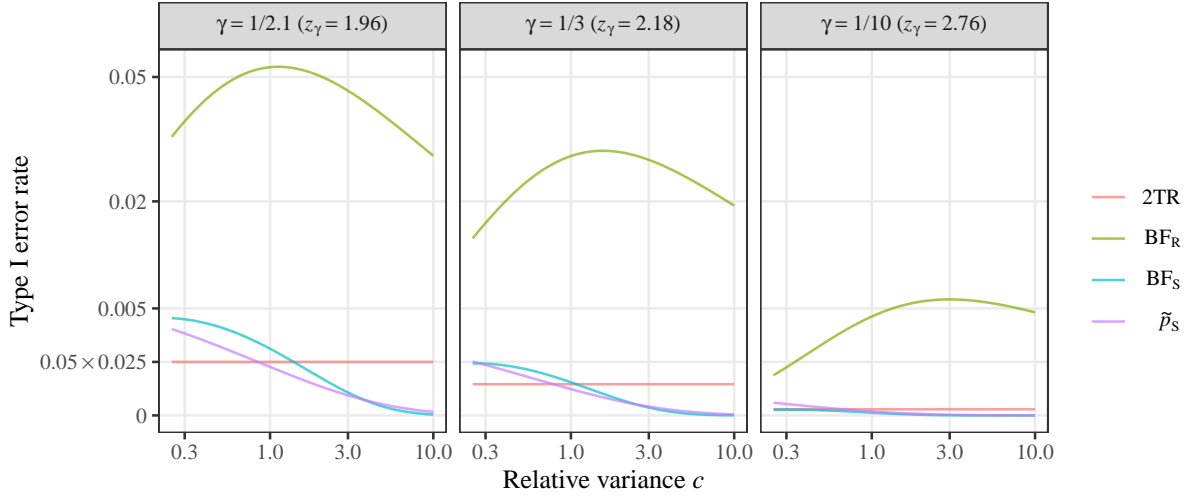
Despite the fact that Bayesian methods do not rely on repeated testing, it is still often of interest to study their frequentist operating characteristics (Dawid, 1982; Grieve, 2016). This is especially important in the replication setting where regulators and funders usually require from statistical 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  $\gamma$  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  $z_o, z_r \mid H_0 \sim N(0, 1)$  and hence the global type I error rate (T1E) based on  $\text{BF}_S \leq \gamma$  is

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

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  $\text{T1E} = 2\{1 - \Phi(z_\gamma)\}^2$ .

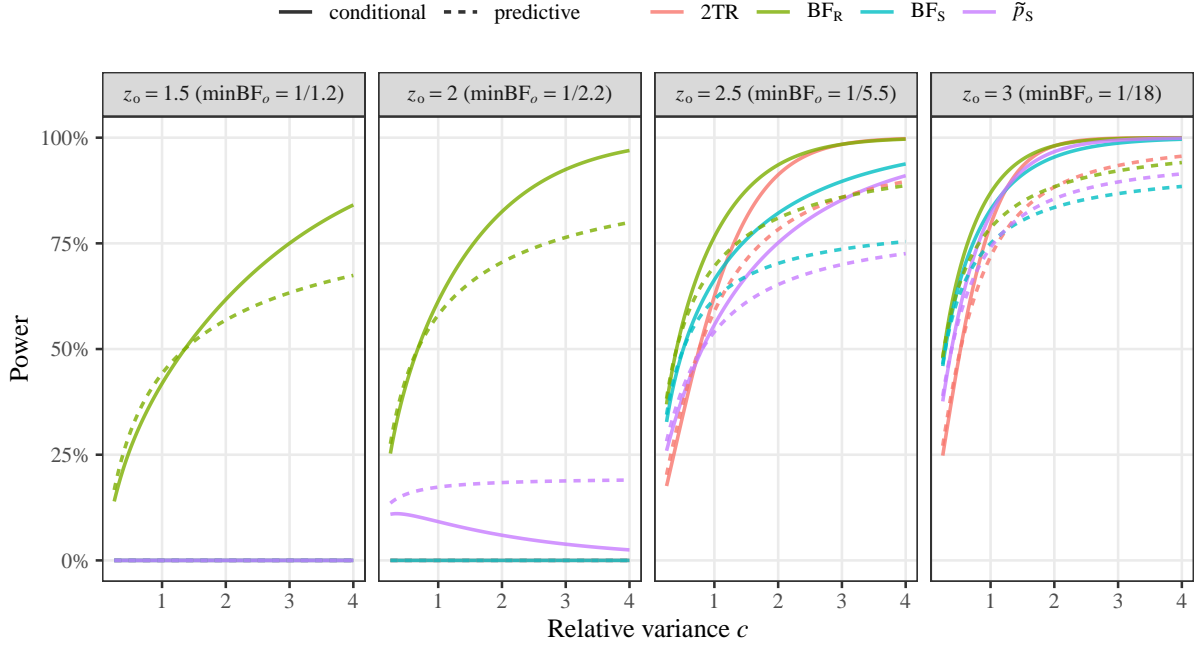


**Figure 5:** Type I error rate of the two-trials rule ( $\min \text{BF}_o \leq \gamma$  and  $\min \text{BF}_r \leq \gamma$ ), the replication Bayes factor ( $\text{BF}_R \leq \gamma$ ), the sceptical Bayes factor ( $\text{BF}_S \leq \gamma$ ), and the recalibrated sceptical  $p$ -value ( $\tilde{p}_S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

Figure 5 compares the type I error rates of the four methods for different levels  $\gamma$ . 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  $\gamma$ . 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  $\gamma$ . 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  $\gamma$  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:  $\min\text{BF}_o \leq 1/3$  and  $\min\text{BF}_r \leq 1/3$ ), the replication Bayes factor ( $\text{BF}_R \leq 1/3$ ), the sceptical Bayes factor ( $\text{BF}_S \leq 1/3$ ), and the recalibrated sceptical  $p$ -value ( $\tilde{p}_S \leq 1 - \Phi\{z_\gamma\}$  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  $\min\text{BF}_o$ ).

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  $\text{BF}_S \leq \gamma$  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 the original  $z$ -value  $z_o$  (respectively original minimum Bayes factor  $\min\text{BF}_o$ ). 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, 2022). The sceptical Bayes factor can never be smaller than the original minimum Bayes factor  $\min BF_o$ . This means that it will be bounded away from zero as the replication sample size grows ( $c \rightarrow \infty$ ), 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  $\hat{\theta}_i \sim N(\theta_*, \sigma_i^2)$ ,  $i \in \{o, r\}$ . 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  $\kappa$  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 ( $\hat{\theta}_i \rightarrow \theta_*$ ), the relative effect estimate will converge to one ( $d \rightarrow 1$ ), and the  $z$ -values will go to infinity ( $|z_i| \rightarrow \infty$ ). 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  $|z_o| \rightarrow \infty$  (Corless et al., 1996, p. 350). Taken together, we have

$$BF_S = \sqrt{\frac{1 + d^2}{1 + (1 - d)^2/2 - \mathcal{O}\{\log(z_o^2)/z_o^2\}}} \cdot \exp \left\{ -\frac{z_o^2 (d^2 + 2d - 1)}{4(1 + d^2)} - \mathcal{O}(\log z_o^2) \right\} \quad (16)$$

Plugging  $d = 1$  into (16), we see that  $BF_S \downarrow 0$  as  $|z_o| \rightarrow \infty$ , 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  $\sigma_o^2$  and  $\sigma_r^2$ . 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  $\theta$ , 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  $\theta$ . 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  $\sigma$ , the means  $\mu_1$  and  $\mu_2$  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  $\sigma^2$ . 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)

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

The same framework is also applicable to test-statistics  $t_i$  from paired  $t$ -tests based on  $n_i$  paired measurements. The SMD  $\theta$  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  $\theta$  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  $\theta | H_5 \sim \text{N}(0, \tau^2)$  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  $\theta$  conditional on the original study and a flat prior on  $\theta$ . 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

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



where  $N(x; \mu, \phi)$  denotes the density function of the normal distribution with mean  $\mu$  and variance  $\phi$  evaluated at  $x$ , and similarly  $IG(y; a, b)$  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

$$\begin{aligned} \text{BF}_{0:S}(t_o; \tau^2) &= \frac{\text{NCT}_{v_o}(t_o; 0)}{\int \text{NCT}_{v_o}(t_o; \theta \sqrt{n_o^*}) N(\theta; 0, \tau^2) d\theta} \\ \text{BF}_{S:A}(t_r; \tau^2) &= \frac{\int \text{NCT}_{v_r}(t_r; \theta \sqrt{n_r^*}) N(\theta; 0, \tau^2) d\theta}{\int \text{NCT}_{v_r}(t_r; \theta \sqrt{n_r^*}) f(\theta | t_o) d\theta} \end{aligned}$$

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

$$X_{1i} | \pi_1 \sim \text{Bin}(n_{1i}, \pi_1) \quad X_{2i} | \pi_2 \sim \text{Bin}(n_{2i}, \pi_2)$$

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  $\theta$ , we need to specify a prior for either  $\pi_2$  or  $\pi_1$  and marginalise over it. A principled choice is the translation invariant Jeffreys prior,  $\pi_1, \pi_2 \sim \text{Be}(1/2, 1/2)$ . The exact marginal likelihood for the data from study  $i$  is then given by

$$\begin{aligned} f(x_{1i}, x_{2i} | \theta) &= \int_0^1 \text{Bin} \left( x_{1i}; n_{1i}, \left\{ 1 + \exp \left[ -\theta - \log \frac{\pi_2}{1 - \pi_2} \right] \right\}^{-1} \right) \text{Bin}(x_{2i}; n_{2i}, \pi_2) \\ &\quad \times \text{Be}(\pi_2; 1/2, 1/2) d\pi_2 \end{aligned} \quad (18)$$

where  $\text{Bin}(x; n, \pi)$  denotes the probability mass function of the binomial distribution with  $n$  trials and probability  $\pi$  evaluated at  $x$ , and likewise  $\text{Be}(y; a, b)$  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  $\theta | H_S \sim N(0, \tau^2)$  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  $\theta$  based on the original study. Using a result from [Marshall \(1988\)](#) combined with a change-of-variables, the exact posterior density of the logOR  $\theta$  given the original data and Jeffreys priors on  $\pi_1$  and  $\pi_2$  is

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

where  $F(\cdot)$  is the hypergeometric function,  $e = x_{10} + 1/2$ ,  $f = n_{10} - x_{10} + 1/2$ ,  $g = x_{20} + 1/2$ ,  $h = n_{20} - x_{20} + 1/2$ ,  $C = B(e + g, f + h) / \{B(e, f)B(g, h)\}$ , and  $B(\cdot, \cdot)$  is the Beta function.

Combining the previous results, we obtain

$$\begin{aligned} \text{BF}_{0.5}(x_{10}, x_{20}; \tau^2) &= \frac{f(x_{10}, x_{20} | 0)}{\int f(x_{10}, x_{20} | \theta) N(\theta; 0, \tau^2) d\theta} \\ \text{BF}_{S:A}(x_{1r}, x_{2r}; \tau^2) &= \frac{\int f(x_{1r}, x_{2r} | \theta) N(\theta; 0, \tau^2) d\theta}{\int f(x_{1r}, x_{2r} | \theta) f(\theta | x_{10}, x_{20}) d\theta} \end{aligned}$$

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  $\text{Var}(\hat{\theta}) = 1/(n - 3)$  (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  $\text{BF}_S$  and  $\text{BF}_R$  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 ( $\min\text{BF}_o = 1/4.2$  and  $\min\text{BF}_r = 1/3.6$ ) and the replication Bayes factor ( $\text{BF}_R = 1/3.9$ ), but not with the sceptical Bayes factor ( $\text{BF}_R = 1/1.6$ ) or the sceptical  $p$ -value ( $\tilde{p}_S = 0.04 > 1 - \Phi(z_\gamma = 2.18) = 0.01$ ).

**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<sub>S</sub>) and replication Bayes factors (BF<sub>R</sub>), the latter two computed using either a normal approximation or the exact likelihood of the data.

Original study	$c$	$d$	$Q$	minBF <sub>o</sub>	minBF <sub>r</sub>	$\tilde{p}_S$	BF <sub>S</sub>	BF <sub>S</sub> (exact)	BF <sub>R</sub>	BF <sub>R</sub> (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
Derey 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  $\min\text{BF}_o < 1/1000$  and  $\hat{\theta}_r = 0.36$  with  $\min\text{BF}_r = 1/3.3$ ), 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  $\min\text{BF}_o = 1/3.2$ ), whereas the replication finding was more compelling ( $\hat{\theta}_r = 0.67$  and  $\min\text{BF}_r < 1/1000$ ). By construction the sceptical Bayes factor can only be as small as the minimum Bayes factor from the original study  $\min\text{BF}_o$ , 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.

## A Sufficiently sceptical relative prior variance

The sufficiently sceptical relative prior variance at level  $\gamma$  is the value  $g_\gamma \in [0, g_{\min \text{BF}_0}]$  that fulfils the condition

$$\text{BF}_{0.5}(\hat{\theta}_0; g_\gamma) = \gamma. \quad (19)$$

Substituting (19) and rearranging terms, we obtain

$$\begin{aligned} \sqrt{1 + g_\gamma} \cdot \exp \left\{ -\frac{1}{2} \cdot \frac{g_\gamma}{1 + g_\gamma} \cdot z_0^2 \right\} &= \gamma \\ \iff \frac{1}{\gamma} \cdot \exp \left\{ -\frac{z_0^2}{2} \right\} &= \frac{1}{\sqrt{1 + g_\gamma}} \exp \left\{ -\frac{1}{2} \cdot \frac{z_0^2}{1 + g_\gamma} \right\}. \end{aligned}$$

Squaring both sides and multiplying by  $-z_0^2$ , this becomes

$$\iff -\frac{z_0^2}{\gamma^2} \cdot \exp \{-z_0^2\} = -\frac{z_0^2}{1 + g_\gamma} \exp \left\{ -\frac{z_0^2}{1 + g_\gamma} \right\}. \quad (20)$$

This is a transcendental equation that cannot be explicitly solved in terms of elementary functions. However, if we set  $q = -z_0^2/(1 + g_\gamma)$  then (20) becomes

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

The solution for  $q$  (and consequently for  $g_\gamma$ ) can be explicitly computed with

$$\begin{aligned} q &= W_{-1} \left( -\frac{z_0^2}{\gamma^2} \cdot \exp \{-z_0^2\} \right) \\ g_\gamma &= \begin{cases} -\frac{z_0^2}{q} - 1 & \text{if } -\frac{z_0^2}{q} \geq 1 \\ \text{undefined} & \text{else} \end{cases} \end{aligned} \quad (21)$$

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  $g_\gamma \leq g_{\min \text{BF}_0}$ . See Appendix B for details about the Lambert  $W$  function. For some  $z_0$ , equation (20) can also be satisfied for negative  $g_\gamma$ , which is why we need to add the condition  $-z_0^2/q \geq 1$  in equation (21), such that  $g_\gamma$  is a valid relative variance.

As  $z_0^2$  becomes larger, the argument to the Lambert function  $x = -z_0^2 \exp(-z_0^2)/\gamma^2$  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

$$g_\gamma \approx \frac{z_0^2}{z_0^2 + \log \gamma^2 - \log z_0^2 + \log\{z_0^2 + \log \gamma^2 - \log z_0^2\}} - 1.$$

We can see that  $g_\gamma \downarrow 0$  when  $\gamma$  remains fixed and  $z_0^2 \rightarrow \infty$ , 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).

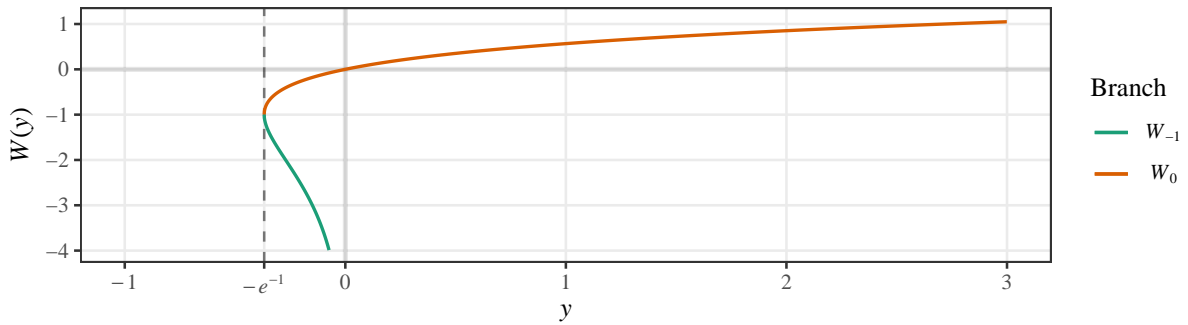


Figure 7: Lambert  $W$  function for real argument  $y$ .

## C Computation of the sceptical Bayes factor

From the definition of the sceptical Bayes factor it is apparent that  $\text{BF}_S$  is either

1. undefined, if  $\text{BF}_{\text{S:A}}(\hat{\theta}_r; g) > \text{BF}_{0:\text{S}}(\hat{\theta}_0; g)$  for all  $g \in [0, g_{\min\text{BF}_0}]$
2.  $\text{BF}_\text{S} = \min\text{BF}_0$ , if  $\text{BF}_{\text{S:A}}(\hat{\theta}_r; g_{\min\text{BF}_0}) \leq \text{BF}_{0:\text{S}}(\hat{\theta}_0; g_{\min\text{BF}_0})$
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  $\text{BF}_{0:\text{S}}(\hat{\theta}_0; g_\gamma) = \gamma$  and  $\text{BF}_{\text{S:A}}(\hat{\theta}_r; g_\gamma)$  in  $g_\gamma$  otherwise

Whether  $\text{BF}_\text{S}$  attains the lower bound  $\min\text{BF}_0$  (condition 2) can be checked by evaluating if  $\text{BF}_{\text{S:A}}(\hat{\theta}_r; g_{\min\text{BF}_0}) \leq \min\text{BF}_0$  and setting  $\text{BF}_\text{S} = \min\text{BF}_0$  if it is the case. For condition 3, we know that the intersections satisfy

$$\text{BF}_{\text{S:A}}(\hat{\theta}_r; g_*) = \text{BF}_{0:\text{S}}(\hat{\theta}_0; g_*)$$

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

which is equivalent to

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

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

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

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

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

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  $g_* \leq g_{\min\text{BF}_0} = \max\{z_o^2 - 1, 0\}$  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  $\hat{\theta}_r | \theta \sim N(\theta, \sigma_r^2)$  under the truncated advocacy prior  $H_{A'}: \theta \sim N(\hat{\theta}_o, \sigma_o^2) \mathbb{1}_{(0, \infty)}(\theta)$  is

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

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

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

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  $\text{minBF}_o$ ) 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

---


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

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

$$\lim_{c \rightarrow \infty} d_{\min}^{\text{BFS}} = \frac{g_\gamma}{g_\gamma - 1} + \sqrt{\frac{\log \left[ \frac{1}{g_\gamma(1+g_\gamma)} \right] / z_o^2 + \frac{g_\gamma}{1+g_\gamma} + \frac{1}{1-g_\gamma}}{(1-g_\gamma)/g_\gamma}} \quad \text{and} \quad \lim_{z_o^2 \rightarrow \infty} d_{\min}^{\text{BFS}} = \sqrt{\frac{1/c+1}{c}} - \frac{1}{c}$$

where for the second limit we used that  $\lim_{z_o^2 \rightarrow \infty} g_\gamma = 0$  for a fixed level  $\gamma$  (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

$$\lim_{c \rightarrow \infty} d_{\min}^{2\text{TR}} = 0 \quad \text{and} \quad \lim_{z_o^2 \rightarrow \infty} d_{\min}^{2\text{TR}} = 0$$

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

$$\lim_{c \rightarrow \infty} d_{\min}^{p\text{S}} = \sqrt{\frac{z_\gamma^2}{z_o^2(z_o^2/z_\gamma^2 - 1)}} \quad \text{and} \quad \lim_{z_o^2 \rightarrow \infty} d_{\min}^{p\text{S}} = 0$$

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

$$\lim_{c \rightarrow \infty} d_{\min}^{\text{BFR}} = 0 \quad \text{and} \quad \lim_{z_o^2 \rightarrow \infty} d_{\min}^{\text{BFR}} = \sqrt{\frac{1/c+1}{c}} - \frac{1/c+1}{1+c}$$

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  $\gamma$  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

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

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 ( $H_0: \theta = 0$ ), we have  $\mu_{z_r} = 0$  and  $\sigma_{z_r}^2 = 1$ . 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  $\sigma_{z_r}^2 = 1$ . Finally, predictive power is obtained by using the predictive distribution based on the advocacy prior ( $H_A: \theta \sim N(\hat{\theta}_o, \sigma_o^2)$ ) and thus  $\mu_{z_r} = z_o \sqrt{c}$  and  $\sigma_{z_r}^2 = 1 + c$ .

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

$$\Pr(\text{BF}_S \leq \gamma | z_o, c) = \begin{cases} \Pr(\chi_{1,\lambda}^2 \geq A / [B\sigma_{z_r}^2]) & \text{for } g_\gamma < 1 \\ \Phi(\text{sign}(z_o) \{ \mu_{z_r} - D \} / \sigma_{z_r}) & \text{for } g_\gamma = 1 \\ \Pr(\chi_{1,\lambda}^2 \leq A / [B\sigma_{z_r}^2]) & \text{for } g_\gamma > 1 \end{cases} \quad (26)$$

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

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

The probability is zero, if the original z-value  $|z_o|$  is not large enough such that the sufficiently sceptical relative prior variance  $g_\gamma$  can be computed for level  $\gamma$  with (17).

## G Probability of replication success with the replication Bayes factor

The probability of  $\text{BF}_R \leq \gamma$  is equivalent to the probability of

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

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

$$\Pr(\text{BF}_R \leq \gamma | z_o, c) = \Pr(\chi_{1,\lambda}^2 \geq \{ z_o^2 + \log(1 + c) - \log \gamma^2 \} [1 + 1/c] / \sigma_{z_r}^2)$$

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

---

## 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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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 (General)*, 143:383–430. doi:[10.2307/2982063](https://doi.org/10.2307/2982063).
- 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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1214/11-sts369).
- Dereux, 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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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*, 13(3):295–314. doi:[10.1002/jrsm.1538](https://doi.org/10.1002/jrsm.1538).
- Held, L., Micheloud, C., and Pawel, S. (2022b). The assessment of replication success based on relative effect size. *The Annals of Applied Statistics*, 16(2):706–720. doi:[10.1214/21-aos1502](https://doi.org/10.1214/21-aos1502).
- 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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/10.3758/s13428-018-1092-x).
- Ly, A. and Wagenmakers, E.-J. (2022). Bayes factors for peri-null hypotheses. *TEST*. doi:[10.1007/s11749-022-00819-w](https://doi.org/10.1007/s11749-022-00819-w).
- Marshall, R. J. (1988). Bayesian analysis of case-control studies. *Statistics in Medicine*, 7(12):1223–1230. doi:[10.1002/sim.4780071203](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1214/21-sts828).
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716. doi:[10.1126/science.aac4716](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1037/a0036731).
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer International Publishing. doi:[10.1007/978-3-319-24277-4](https://doi.org/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.





## 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>.

---



---

## Abstract

Replication studies are increasingly conducted in order to confirm original findings. However, there is no established standard how to assess replication success and in practice many different approaches are used. The purpose of this paper is to refine and extend a recently proposed reverse-Bayes approach for the analysis of replication studies. We show how this method is directly related to the relative effect size, the ratio of the replication to the original effect estimate. This perspective leads to a new proposal to recalibrate the assessment of replication success, the golden level. The recalibration ensures that for borderline significant original studies replication success can only be achieved if the replication effect estimate is larger than the original one. Conditional power for replication success can then take any desired value if the original study is significant and the replication sample size is large enough. Compared to the standard approach to require statistical significance of both the original and replication study, replication success at the golden level offers uniform gains in project power and controls the Type-I error rate if the replication sample size is not smaller than the original one. An application to data from four large replication projects shows that the new approach leads to more appropriate inferences, as it penalizes shrinkage of the replication estimate compared to the original one, while ensuring that both effect estimates are sufficiently convincing on their own.

**Key words:** Power, replication studies, sceptical  $p$ -value, shrinkage, two-trials rule, type-I error rate

## 1 Introduction

Replication studies are conducted in order to investigate whether an original finding can be confirmed in an independent study. Although replication has long been a central part of the scientific method in many fields, the so-called replication crisis ([Ioannidis, 2005](#); [Begley and Ioannidis, 2015](#)) has led to increased interest in replication over the last decade. These developments eventually culminated in large-scale replication projects that were conducted in various fields ([Errington et al., 2014](#); [Klein et al., 2014](#); [Open Science Collaboration, 2015](#); [Ebersole et al., 2016](#); [Camerer et al., 2016, 2018](#); [Cova et al., 2018](#); [Klein et al., 2018](#)).

Declaring a replication as successful is, however, not a straightforward task, and currently used approaches include statistical significance of both the original and replication study, compatibility of their effect estimates, and meta-analysis of the effect estimates. Many of the replication projects listed above also report the relative effect size, the ratio of the replication to the original effect estimate. For example, in [Camerer et al. \(2018\)](#) the replication effect estimates were only half as large as the original ones on average and even smaller in [Open Science Collaboration \(2015\)](#). This gives clear evidence of a systematic bias of the original studies and strongly suggests that the original and replication study should not be treated as exchangeable. However, all the approaches mentioned above will give the same results if the order of studies would be reversed.

In order to address this problem, a new method has recently been proposed in [Held \(2020b\)](#). The approach combines the analysis of credibility ([Matthews, 2001a,b](#)) with a prior-data conflict assessment ([Box, 1980](#)). Replication success is declared if the replication study is in conflict with a sceptical prior that would make the original study non-significant. This approach penalizes small relative effect sizes as we will see in more detail in the following.

To introduce some notation, let  $z_o = \hat{\theta}_o / \sigma_o$  and  $z_r = \hat{\theta}_r / \sigma_r$  denote the z-statistic of the original and replication study, respectively. Here  $\hat{\theta}_o$  and  $\hat{\theta}_r$  are the corresponding effect estimates (assumed to be normally distributed) of the unknown effect  $\theta$  with standard errors  $\sigma_o$  and  $\sigma_r$ , respectively. The corresponding one-sided p-values are denoted by  $p_o = 1 - \Phi(z_o)$  and  $p_r = 1 - \Phi(z_r)$ , respectively, where  $\Phi(\cdot)$  denotes the standard normal cumulative distribution function. Let  $c = \sigma_o^2 / \sigma_r^2$  denote the variance ratio of the squared standard errors of the original and replication effect estimates. The squared standard errors are usually inversely proportional to the sample size of each study, i. e.,  $\sigma_o^2 = \kappa^2 / n_o$  and  $\sigma_r^2 = \kappa^2 / n_r$  for some unit variance  $\kappa^2$ . The variance ratio  $c$  can then be identified as the relative sample size  $c = n_r / n_o$ . The relative effect size

$$d = \frac{\hat{\theta}_r}{\hat{\theta}_o} = \frac{1}{\sqrt{c}} \frac{z_r}{z_o} \quad (1)$$

quantifies the size of the replication effect estimate  $\hat{\theta}_r$  relative to the original effect estimate  $\hat{\theta}_o$ . The corresponding shrinkage of the replication effect estimate will be denoted as  $s = 1 - d$ .

Suppose the original study achieved statistical significance at one-sided level  $\alpha$ , so  $p_o \leq \alpha$ . The standard approach to assess replication success is based on significance of the replication effect estimate at the same level  $\alpha$ , i. e., the replication is considered successful if also  $p_r \leq \alpha$ . This approach is known in drug development as the two-trials rule ([Senn, 2008](#)), usually conducted at  $\alpha = 0.025$ . Let  $z_\alpha = \Phi^{-1}(1 - \alpha) > 0$  denote the z-value corresponding to the level  $\alpha$ , then significance of the replication study is achieved if  $z_r \geq z_\alpha$ , which is equivalent to the condition

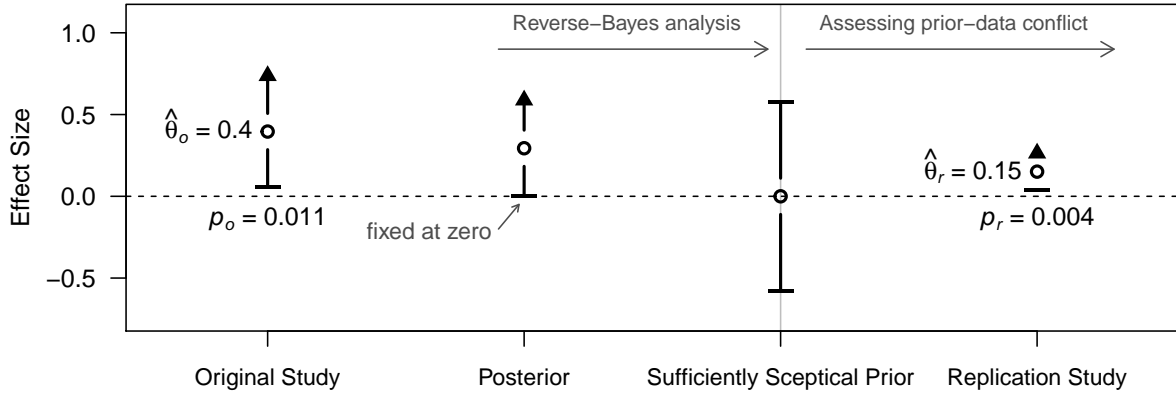
$$d \geq \frac{z_\alpha}{z_o \sqrt{c}} \quad (2)$$

on the relative effect size (1). The right hand-side goes to zero for increasing  $c$ , so if the relative sample size  $c$  is large enough, significance of the replication study can be achieved with any arbitrarily small (but positive) relative effect size  $d$ . However, declaring replication success when there is substantial shrinkage is contrary to common sense, as the replication effect estimate may not reflect an effect size of the same practical relevance as the original one, despite its statistical significance.

In this paper we first review the [Held \(2020b\)](#) approach for the assessment of replication success, followed by showing how it relates to the relative effect size (Section 2.1). This perspective is used in Section 2.2 and 2.3 to propose a recalibration of the method, the *golden level*, which leads to a more appropriate criterion for replication success compared to the two-trials rule (Section 2.4). In Section 3 we study power and Type-I error rates of the proposed method and compare it to the two-trials rule. The recalibrated method ensures that conditional power can take any desired value if the original study has been significant and the replication sample size is large enough (Section 3.1), controls the overall Type-I error if the replication sample size is not smaller than the original one (Section 3.2), and offers uniform gains in project power

compared to the two-trials rule (Section 3.3). Section 4 describes an application to data from four replication projects and Section 5 closes with some discussion.

## 2 Replication success



**Figure 1:** Example of the assessment of replication success. The original study from [Pyc and Rawson \(2010\)](#) has effect estimate  $\hat{\theta}_o = 0.4$  on Fisher's  $z$  scale (95% CI from 0.05 to 0.74) and one-sided  $p$ -value  $p_o = 0.011$ . The left part of the figure illustrates the reverse-Bayes derivation of the sufficiently sceptical prior based on the original study result and the posterior with lower credible limit fixed at zero. The comparison of the sufficiently sceptical prior with the replication study result ( $\hat{\theta}_r = 0.15$ , 95% CI from 0.04 to 0.26,  $p_r = 0.004$ ) in the right part of the figure is used to assess potential prior-data conflict.

Hereinafter we focus on the one-sided assessment of replication success to ensure that replication success can only occur if the original and replication effect estimates go in the same direction. Figure 1 illustrates the [Held \(2020b\)](#) approach based on a replication study from the *Social Sciences Replication Project* ([Camerer et al., 2018](#)): the significant original finding by [Pyc and Rawson \(2010\)](#) at one-sided level  $\alpha = 0.025$  is challenged with a sceptical prior, sufficiently concentrated around zero to make the original study result no longer convincing ([Matthews, 2001a,b](#)). Replication success is then defined as conflict between the sceptical prior and the result from the replication study in order to disprove the sceptic. Conflict is quantified by a prior-predictive tail probability  $p_{\text{Box}}$  ([Box, 1980](#)) where a small value  $p_{\text{Box}} \leq \alpha$  defines replication success. In Figure 1 the original finding is only borderline significant, so the sufficiently sceptical prior is fairly wide. Furthermore, there is substantial shrinkage ( $d = 0.15/0.4 = 0.38$ ) of the replication effect estimate and therefore hardly any conflict with the sufficiently sceptical prior (one-sided  $p_{\text{Box}} = 0.31$ ). We are thus not able to declare replication success at level 2.5%.

The actual value of  $p_{\text{Box}}$  is difficult to interpret as it depends on the level  $\alpha$  and does not even exist if the original  $p$ -value  $p_o$  exceeds  $\alpha$ . However, [Held \(2020b\)](#) showed that if both  $\text{sign}(z_o) = \text{sign}(z_r)$  and

$$(z_o^2/z_{\alpha_S}^2 - 1)(z_r^2/z_{\alpha_S}^2 - 1) \geq c \quad (3)$$

hold, replication success at level  $\alpha_S$  is achieved, where  $z_{\alpha_S} = \Phi^{-1}(1 - \alpha_S)$ . The requirement (3) can be assessed for any value of  $\alpha_S > \max\{p_o, p_r\}$  and of particular interest is the smallest possible value of  $\alpha_S$  where (3) holds, the so-called *sceptical p-value*  $p_S$ . We are thus interested in the value  $z_S^2$  that fulfills

$$(z_o^2/z_S^2 - 1)(z_r^2/z_S^2 - 1) = c. \quad (4)$$

There is a unique solution of (4) which defines the one-sided sceptical  $p$ -value  $p_S = 1 - \Phi(z_S)$  where  $z_S := +\sqrt{z_S^2}$ , provided  $\text{sign}(z_o) = \text{sign}(z_r)$  holds. Replication success at level  $\alpha_S$  is then achieved if  $p_S \leq \alpha_S$ . In the introductory example based on the original study by [Pyc and Rawson \(2010\)](#), the sceptical  $p$ -value turns out to be  $p_S = 0.11$ .

The sceptical  $p$ -value has a number of interesting properties, see [Held \(2020b, Section 3.1\)](#) for details. In particular,  $p_S > \max\{p_o, p_r\}$  always holds with  $p_S \downarrow \max\{p_o, p_r\}$  for  $c \downarrow 0$ . Furthermore, if the  $p$ -values  $p_o$  and  $p_r$  are fixed, the sceptical  $p$ -value  $p_S$  increases with decreasing relative effect size  $d$ . The first property ensures that both the original and the replication study have to be sufficiently convincing on their own to achieve replication success. The second property guarantees that shrinkage of the replication effect estimate is penalized.

The level for replication success  $\alpha_S$  has to be distinguished from the significance level  $\alpha$  associated with the ordinary  $p$ -value. [Held \(2020b\)](#) has used the *nominal level* for replication success ( $\alpha_S = \alpha$ ) for convenience, but in the following we will propose a recalibration of the procedure along with a new value for  $\alpha_S$ , the *golden level* (Section 2.2). The derivation is based on a property of the required relative effect size for replication success, if the relative sample size is very large (Section 2.1). In a nutshell, the golden level ensures that for original studies which were only borderline significant ( $p_o = \alpha$ ), replication success is only possible if the replication effect estimate is larger than the original one ( $d > 1$ ).

## 2.1 Relative effect size

Without loss of generality we now assume that  $\hat{\theta}_o > 0$  and that  $p_o < \alpha_S$  has been observed in the original study, otherwise it would be impossible to achieve replication success at level  $\alpha_S$  because  $p_S$  is always larger than  $p_o$ . The condition (3) for replication success can then be re-written as

$$z_r \geq z_{\alpha_S} \sqrt{1 + c/(K - 1)} =: z_r^{\min}, \quad (5)$$

where  $K = z_o^2/z_{\alpha_S}^2 > 1$ . The right hand-side of (5) is the minimum replication  $z$ -value  $z_r^{\min}$  required to achieve replication success. Note that  $z_r^{\min}$  increases with increasing  $c$ , so increasing the replication sample size leads to a more stringent success requirement  $z_r$  and the corresponding replication  $p$ -value  $p_r$ .

Equation (5) can be further transformed to a condition on the relative effect size (1):

$$d \geq \frac{\sqrt{1 + c/(K - 1)}}{\sqrt{cK}} =: d_{\min}. \quad (6)$$

To achieve replication success, the relative effect size must be at least as large as the right hand-side of (6), the minimum relative effect size  $d_{\min}$ , a function of  $K$  and the relative sample size  $c$ . If the relative sample size becomes very large, i. e.,  $c \rightarrow \infty$ , we have  $d_{\min} \downarrow d_{\infty}$  where

$$d_{\infty} = 1 / \sqrt{K(K-1)} \quad (7)$$

is the *limiting relative effect size*. This shows that the minimum relative effect size  $d_{\min}$  in (6) does not go to zero for increasing  $c$ , so replication success cannot be achieved if the relative effect size  $d$  is smaller or equal to  $d_{\infty}$ , no matter how large the replication study is. In contrast, the corresponding criterion (2) of the two-trials rule can be achieved for any positive relative effect size, regardless of how small, provided the replication sample size is sufficiently large.

## 2.2 The golden level

Significance of both the original and the replication study at level  $\alpha$  is a necessary but not sufficient requirement for replication success at the nominal level ( $\alpha_S = \alpha$ ). The nominal level may therefore be too stringent. It is more reasonable to calibrate the procedure in such a way that to establish replication success, original and replication study do not both necessarily need to be significant at level  $\alpha$ , provided that the replication effect estimate does not shrink compared to the original one. We therefore choose a level  $\alpha_S$  such that a borderline significant original study ( $p_o = \alpha$ ) cannot lead to replication success if there is shrinkage  $s > 0$  of the replication effect estimate. Mathematically, this translates to setting  $d_{\infty} = 1$  and  $K = z_{\alpha}^2 / z_{\alpha_S}^2$  in (7) and leads to the quadratic equation  $K(K-1) = 1$  with solution  $K = \varphi$  where  $\varphi = (\sqrt{5} + 1)/2 \approx 1.62$  is known as the golden ratio. Solving for  $z_{\alpha_S}$  gives  $z_{\alpha_S} = z_{\alpha} / \sqrt{\varphi}$  and the corresponding *golden level*

$$\alpha_S = 1 - \Phi(z_{\alpha} / \sqrt{\varphi}) \quad (8)$$

for replication success. This is our recommended default choice to assess replication success and we will study its properties in the following in more detail. For  $z_{\alpha} = 1.96$  (one-sided  $\alpha = 0.025$ ), the golden level is  $\alpha_S = 0.062$ . In the introductory example shown in Figure 1, the sceptical  $p$ -value is  $p_S = 0.11 > 0.062$ , so the replication of the [Pyc and Rawson \(2010\)](#) study was not successful.

The golden level (8) is derived from (7) with  $d_{\infty} = 1$ . However, we may also use a different value for the limiting relative effect size  $d_{\infty}$ , say  $d_{\infty} = 0.8$ . Then replication success is only possible for a borderline significant result ( $p_o = \alpha$ ) if there is less than  $1 - d_{\infty}$  (20% for  $d_{\infty} = 0.8$ ) shrinkage of the replication effect estimate. This approach is equivalent to a limiting relative effect size of 1 if the original  $p$ -value  $p_o$  is equal to a different level  $\alpha'$ , which can be derived as follows: First, solving (7) for  $d_{\infty} > 0$  gives  $K = z_{\alpha}^2 / z_{\alpha_S}^2 = 1/2 + \sqrt{1/4 + 1/d_{\infty}^2}$ . The new level  $\alpha'$  fulfills  $\varphi = z_{\alpha'}^2 / z_{\alpha_S}^2$ , so  $z_{\alpha}^2 / K = z_{\alpha'}^2 / \varphi$  and therefore

$$\alpha' = 1 - \Phi\left(z_{\alpha} \sqrt{\varphi / K}\right). \quad (9)$$

For example, for  $\alpha = 0.025$  and  $d_{\infty} = 0.8$  we obtain  $\alpha' = 0.033$ .

---

## 2.3 Recalibration of the sceptical $p$ -value

The condition  $p_S \leq \alpha_S$  for replication success at the golden level is equivalent to  $z_S \geq z_\alpha / \sqrt{\varphi}$ , i. e.,  $z_S \sqrt{\varphi} \geq z_\alpha$ . In practice it may be preferable to recalibrate the sceptical  $p$ -value  $p_S = 1 - \Phi(z_S)$  to  $\tilde{p}_S = 1 - \Phi(z_S \sqrt{\varphi})$ , which then needs to be compared to  $\alpha$  (rather than  $\alpha_S$ ) to assess replication success and can thus be interpreted on the same scale as an ordinary  $p$ -value. For example, the recalibrated sceptical  $p$ -value for the replication of [Pyc and Rawson \(2010\)](#) turns out to be  $\tilde{p}_S = 0.061$  and does not lead to replication success at any level  $\alpha < 0.061$ , including the standard 0.025 level.

## 2.4 Comparison with the two-trials rule

A useful benchmark for comparison is the two-trials rule in drug development ([Kay, 2015](#), Section 9.4), which requires “at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness” ([FDA, 1998](#), p. 3). This is usually achieved by independently replicating the result of a first study in a second study, both significant at one-sided level  $\alpha = 0.025$ . It is worth noting that in practice the two trials are often run in parallel ([Senn, 2008](#)), so do not exactly resemble the replication setting.

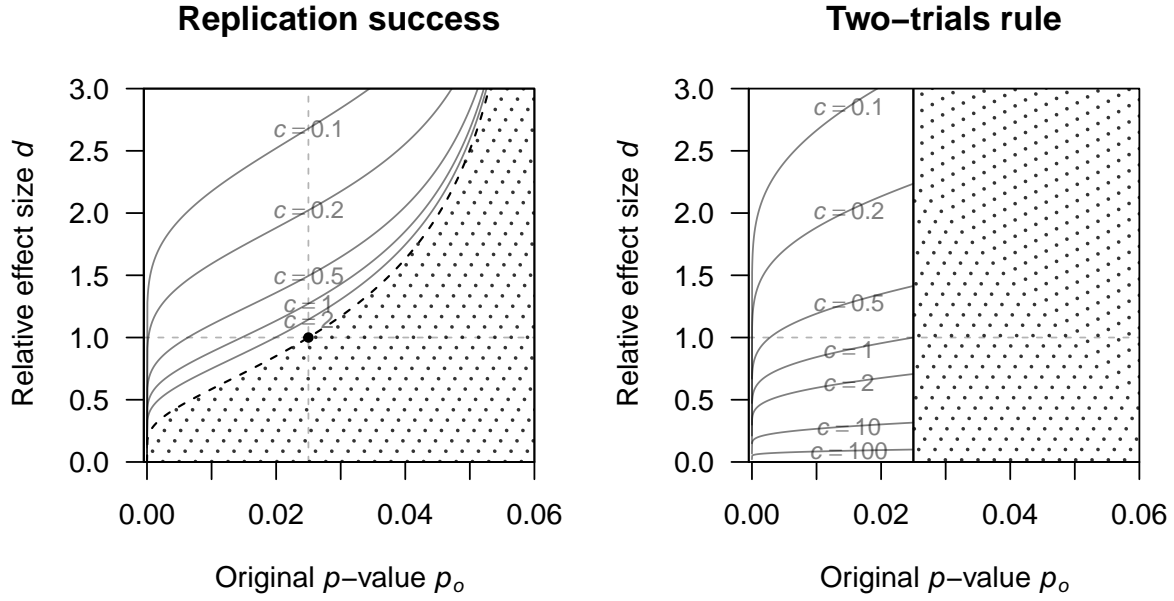
The main difference between the replication success and the two-trials rule approach concerns how shrinkage of the replication effect estimate is handled. Figure 2 illustrates that shrinkage is penalized in the assessment of replication success, i. e., the original  $p$ -value needs to be quite small to achieve replication success for a relative effect size  $d < 1$ . In contrast, significance of the replication study can be achieved even if there is substantial shrinkage, provided the replication sample size is large enough.

It is interesting to directly compare the two-trials rule and replication success at the golden level in terms of the required relative effect size  $d$  to fulfill the criteria (2) and (6), respectively, see Figure 2. If the original  $p$ -value is not significant at level  $\alpha$ , only replication success can be achieved, but will require a replication effect estimate larger than the original one.

For example, four studies with one-sided  $p_o \in (0.025, 0.03)$  have been included in the *Reproducibility Project: Psychology* ([Open Science Collaboration, 2015](#)) and one of them achieves replication success (see Section 4 for details). By definition, such non-significant original findings can never fulfill the two-trials rule.

If the original  $p$ -value is smaller than  $\alpha$ , then the situation depends on the relative sample size  $c$ . For example, when the replication sample size is chosen to be the same as in the original study ( $c = 1$ ) and  $\alpha = 0.025$ , original studies with a  $p$ -value larger than 0.006 will require a smaller relative effect size  $d$  with the two-trials rule, while  $p$ -values smaller than 0.006 will require a smaller relative effect size  $d$  with the replication success method. This illustrates that the latter method is less stringent than the two-trials rule if the original study is already sufficiently convincing.





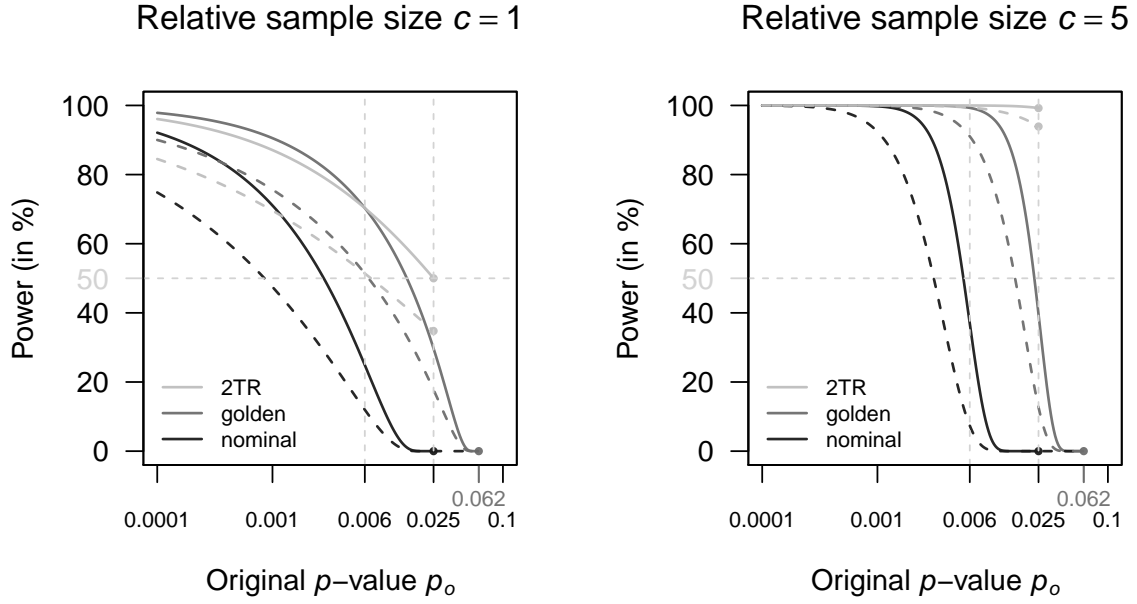
**Figure 2:** Comparison of replication success at the golden level ( $p_S \leq \alpha_S = 0.062$ ) and the two-trials rule ( $p_o \leq 0.025$  and  $p_r \leq 0.025$ ). The dotted areas indicate that success is impossible for original  $p$ -value  $p_o$  and relative effect size  $d$ . In the white areas success is possible and depends on the relative sample size  $c$  as indicated by the grey lines. The dashed black line in the left plot indicates the limiting relative effect size  $d_\infty$ .

### 3 Power and Type-I Error Rate

Although Bayesian methods do not rely on the frequentist paradigm of repeated testing, it is still useful to investigate their frequentist operating characteristics (Dawid, 1982; Rubin, 1984; Grieve, 2016) and this also holds for the proposed reverse-Bayes assessment of replication success. We first condition on the results from the original study and compare the power to achieve replication success with the two-trials rule in Section 3.1. We then assume that none of the two studies have been conducted and investigate the overall Type-I error rate (Section 3.2) and the project power (Section 3.3) (Maca et al., 2002) over both studies in combination for fixed relative sample size  $c$ .

#### 3.1 Conditional power

Figure 3 compares the power for replication success (see Held, 2020b, Section 4 for details) at the golden and at the nominal level with the power of the two-trials rule for relative sample size  $c = 1$  (left) and  $c = 5$  (right) as a function of the one-sided  $p$ -value  $p_o$  from the original study. Shown is the conditional power assuming the unknown parameter  $\theta$  is equal to the original effect estimate  $\hat{\theta}_o$ . Then  $\hat{\theta}_r | \hat{\theta}_o \sim N(\hat{\theta}_o, \kappa^2/n_r)$  and it follows that  $d | \hat{\theta}_o \sim N(1, 1/(cz_o^2))$ .



**Figure 3:** Conditional power as a function of the one-sided  $p$ -value of the original study with relative sample size  $c = 1$  (left) and  $c = 5$  (right). Shown is conditional power assuming the unknown parameter is equal to the original effect estimate (solid) and conditional power based on 20% shrinkage of the original effect estimate (dashed) for the two-trials rule (2TR) at level  $\alpha = 0.025$  and for replication success at the corresponding golden and nominal level. Power values of exactly zero are omitted.

The conditional power for replication success can therefore be calculated as

$$\Pr(d \geq d_{\min} | \hat{\theta}_o) = \Phi \left[ \sqrt{c} z_o (1 - d_{\min}) \right], \quad (10)$$

where  $d_{\min}$  is given in (6). Predictive power, which is conditional power averaged over a  $N(\hat{\theta}_o, \sigma_o^2)$  distribution for the effect size  $\theta$ , could also be calculated, then  $d | \hat{\theta}_o \sim N(1, (1 + 1/c)/z_o^2)$ . Conditional and predictive power of the two-trials rule also depend on  $z_o$ ,  $c$  and  $\alpha$  and are given in [Micheloud and Held \(2022\)](#).

The two-trials rule requires a significant original study and hence it is impossible to power a replication study when  $p_o > 0.025$ . The same applies for replication success at the nominal level, where the power is zero for any  $p_o > 0.025$ , regardless of the replication sample size. This is different for the golden level, where the conditional power of an original study with  $0.025 < p_o < 0.062$  is low, but not zero. However, if the original  $p$ -value  $p_o$  is slightly smaller than 0.025, the two-trials rule has a larger power, both for  $c = 1$  and  $c = 5$ . But if the original  $p$ -value is sufficiently small ( $p_o < 0.006$  for  $c = 1$ ), the power for replication success at the golden level is larger than the power of the two-trials rule.

Compared to  $c = 1$ , the conditional power for  $c = 5$  of both the two-trials rule and the replication success approach at the golden level increases if  $p_o \leq \alpha$ . A remarkable feature of the replication success approach at the golden level is that conditional power can be pushed

towards 100% for large enough  $c$  if  $p_o < \alpha$ , but not otherwise. This can be seen from (10) because  $d_{\min} < 1$  for  $p_o < \alpha$  and large enough relative sample size  $c$ . On the other hand, for  $p_o > \alpha$  conditional power for replication success will tend to 0% for increasing  $c$  because  $d_{\min} > 1$  for all  $c$ . Finally, for  $p_o = \alpha$  the limit is 50%. The same property can be observed at the nominal level, however at the smaller threshold  $1 - \Phi(z_\alpha \sqrt{\varphi})$  which is 0.006 for  $\alpha = 0.025$ . Only if  $p_o < 0.006$  will the conditional power for replication success attain 100% for  $c \rightarrow \infty$ . This further highlights the stringency of the nominal level.

The approach described so far takes the original study at face-value since it assumes that  $\hat{\theta}_o$  is equal to the unknown effect size  $\theta$ . In practice, however, there are often good reasons to believe that original effect estimates have a tendency to be inflated (e. g., due to publication bias). One way to address this issue is to base power calculations on a shrunken version of the original effect estimate, where the amount of shrinkage is guided by domain knowledge and a risk of bias assessment of the original study. For illustration, Figure 3 also shows conditional power based on 20% shrinkage of the original effect estimate which reduces the conditional power for all methods, especially for a relative sample size  $c = 1$ . Conditional power for replication success at the golden level can now be pushed towards 100% only for  $p_o < 0.018$ , which can be derived by solving (9) for  $\alpha$  with  $\alpha' = 0.025$  and  $d_\infty = 0.8$ . To be able to push conditional power based on 20% shrinkage towards 100% for all  $p_o < 0.025$ , equation (9) would have to be used directly to relax the level from  $\alpha = 0.025$  to  $\alpha' = 0.033$ .

### 3.2 Overall Type-I error rate

The two studies are assumed to be independent with Type-I error rate fixed at  $\alpha$  for each of them, so the Type-I error rate of the two-trials rule over the entire project is simply  $\alpha^2$  for any value of the relative effect size  $c$ . In contrast, the Type-I error rate of the proposed replication success assessment depends on the relative sample size  $c$ .

For  $c = 1$ , Held (2020b, Section 3) showed that  $z_S^2$  in (4) simplifies to half the harmonic mean of the squared test statistics  $z_o^2$  and  $z_r^2$ . The connection  $z_S^2 = z_H^2/4$  to the harmonic mean  $\chi^2$ -test statistic  $z_H^2$  (Held, 2020a), which has a  $\chi^2(1)$ -distribution under the null hypothesis, makes it straightforward to compute the Type-I error rate at level  $\alpha_S$  for  $c = 1$  as

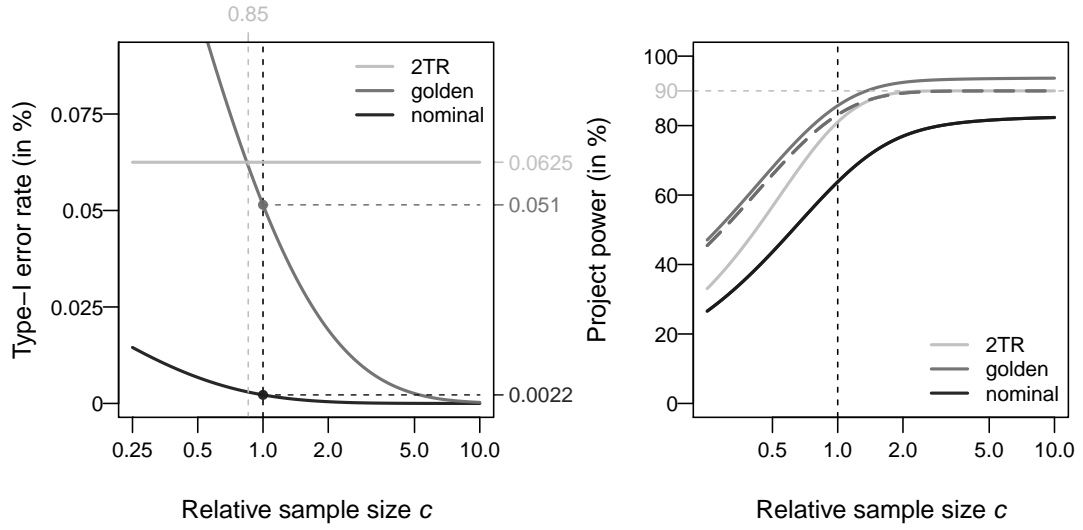
$$\text{T1E} = \left\{ 1 - \Phi \left[ 2 \Phi^{-1} (1 - \alpha_S) \right] \right\} / 2. \quad (11)$$

For the golden level  $\alpha_S = 0.062$  at  $\alpha = 0.025$ , the Type-I error rate (11) is 0.0515%, slightly less than the Type-I error rate  $\alpha^2 = 0.0625\%$  of the two-trials rule. For comparison, the Type-I error rate at the nominal level  $\alpha_S = 0.025$  is 0.0022%, much smaller than 0.0625%.

For  $c \neq 1$ , the Type-I error rate can be calculated through numerical integration:

$$\text{T1E} = \int_{z_{\alpha_S}}^{\infty} \Pr(z_r \geq z_r^{\min} \mid z_o, c, \alpha_S) \phi(z_o) dz_o, \quad (12)$$

where  $\phi(\cdot)$  denotes the standard normal density function. The first term in the integral of (12) is the probability of replication success at level  $\alpha_S$  conditional on a fixed original test statistic



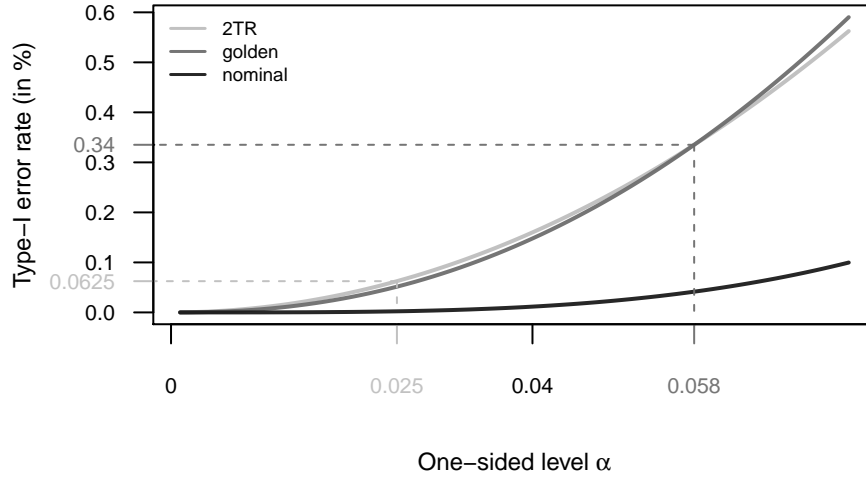
**Figure 4:** Overall Type-I error rate (left) and project power (right) for fixed relative sample size  $c$ . Results are given for replication success at the nominal and golden level and compared with the two-trials rule (2TR) at  $\alpha = 0.025$ . The dashed darkgrey line is the project power at the golden level based on significant original studies ( $p_o \leq 0.025$ ). The power of the original study is 90%.

$z_o$  and a relative sample size  $c$ . Now  $z_r \sim N(0,1)$  under the null hypothesis, so this term simplifies to  $\Pr(z_r \geq z_r^{\min} | z_o, c, \alpha_S) = 1 - \Phi(z_r^{\min})$  where  $z_r^{\min}$  in (5) depends on  $z_o$ ,  $c$ , and  $\alpha_S$ .

The left plot in Figure 4 displays the Type-I error rate for  $\alpha = 0.025$  as a function of the relative sample size  $c$ . It can be seen that the Type-I error of the replication success approach decreases with increasing relative sample size  $c$ . This also follows from (12) where  $\Pr(z_r \geq z_r^{\min} | z_o, c, \alpha_S) = 1 - \Phi(z_r^{\min})$  decreases with increasing  $c$ , because  $z_r^{\min}$  increases with increasing  $c$ , see equation (5).

The Type-I error rate of the nominal level is always below the target 0.0625%. Although the Type-I error will eventually attain  $\alpha^2$  in the limit  $c \downarrow 0$  (Held, 2020b, Section 3.4), the nominal level seems to be too stringent for realistic values of  $c$ . The Type-I error rate of the golden level is smaller than 0.0625% for  $c > 0.85$ . Appropriate Type-I error control is thus ensured even for replication studies where the sample size is slightly smaller than in the original study.

Figure 5 compares for  $c = 1$  the Type-I error rate (11) of replication success at the golden and at the nominal level with the two-trials rule for different values of  $\alpha$ . The Type-I error rate of the two-trials rule is  $\alpha^2$  and the replication success approach at the nominal level always has a much smaller Type-I error rate than  $\alpha^2$ . At the golden level the Type-I error rate of the replication success approach is much closer to  $\alpha^2$ , still slightly smaller if  $\alpha < 0.058$ . For  $\alpha = 0.058$  the Type-I error rate is equal to the Type-I error rate  $0.058^2 = 0.34\%$  of the two-trials rule and for  $\alpha > 0.058$  the Type-I error rate is slightly larger than  $\alpha^2$ . The Type-I error rate for replication success decreases with increasing  $c$ , so as long as the replication sample size is not smaller than the original sample size, Type-I error control at  $\alpha^2$  is guaranteed at the golden level for any one-sided level  $\alpha < 0.058$ .



**Figure 5:** Overall Type-I error rate if the replication sample size equal to the original study ( $c = 1$ ). The two-trials rule (2TR) is compared to replication success at the golden and nominal level for different values of  $\alpha$ .

### 3.3 Project power

Under the alternative we have  $z_o \sim N(\mu, 1)$  with  $\mu = z_\alpha + z_\beta$  where  $\alpha$  is the assumed significance level and  $1 - \beta = \Phi(\mu - z_\alpha)$  is the power to detect the assumed effect  $\theta = \mu\sigma_o$  in the original study (Matthews, 2006, Section 3.3). In the following  $\alpha = 0.025$  and  $\beta = 0.1$  are used. The power of a significant replication study with sample size  $n_r = cn_o$  is

$$\Phi(\theta/\sigma_r - z_\alpha) = \Phi(\sqrt{c}\mu - z_\alpha),$$

so depends on  $\mu$  and the relative sample size  $c$ . The project power of the two-trials rule is therefore  $(1 - \beta) \Phi(\sqrt{c}\mu - z_\alpha)$  and increases with increasing  $c$ .

The project power for replication success is computed as

$$PP = \int_{z_{\alpha_S}}^{\infty} \Pr(z_r \geq z_r^{\min} | z_o, c, \alpha_S) \phi(z_o - \mu) dz_o$$

and shown in the right plot of Figure 4 as a function of  $c$ . For the golden level, the project power quickly increases to values above 90%, whereas the nominal level only reaches around 80% project power. The project power based on the two-trials rule is shown for comparison, which is always smaller than for the golden level and converges to 90% for large  $c$ .

The advantage in power stems partly from replication success still being possible when the original  $p$ -value is larger than 0.025, but smaller than 0.062. If we assume that a replication

study is only conducted if the original study is significant (with  $p_o \leq 0.025$ ), then the project power based on the golden level (the dashed line in Figure 4) is slightly smaller and for  $c > 1$  barely different than for the two-trials rule. More substantial gains are still visible for  $c < 1$ . However, the restriction to original studies with  $p_o \leq 0.025$  may not reflect current practice in large-scale replication projects. For example, 5 out of 143 replication studies considered in Section 4 do have original  $p$ -values between 0.025 and 0.062.

## 4 Application

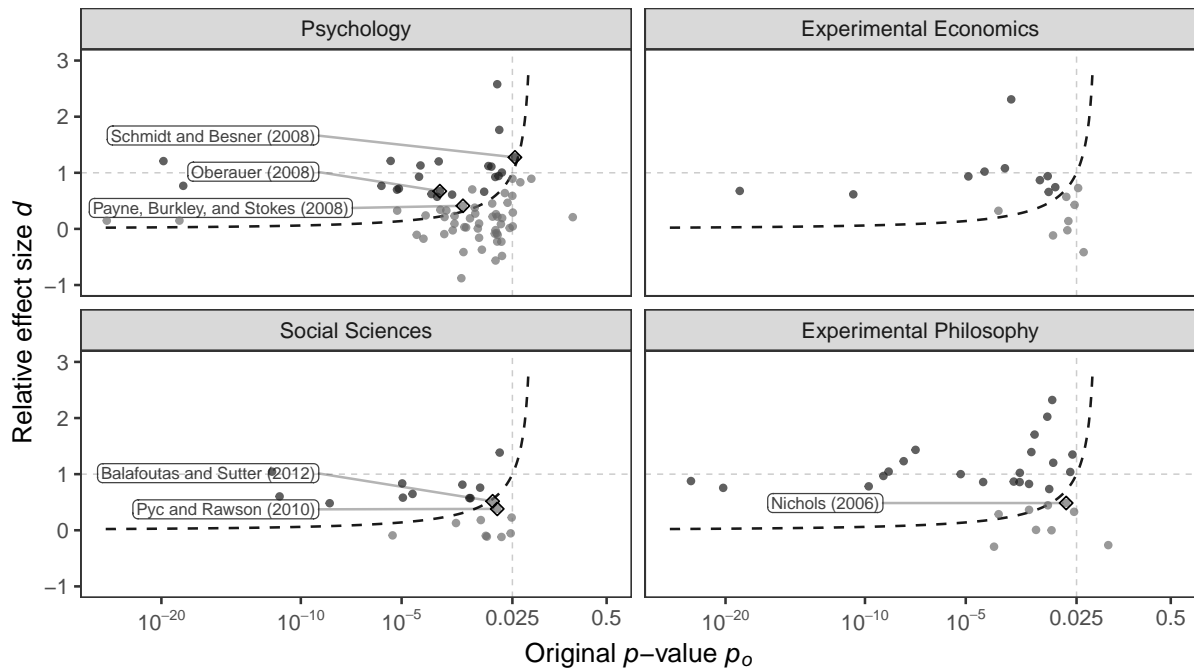
In this section, we illustrate the proposed methodology using data from four replication projects. All four projects reported effect estimates that were transformed to correlation coefficients ( $r$ ). This scale allows for easy comparison of effect estimates from studies that investigate different phenomena and is bounded to the interval between minus one and one. Moreover, the Fisher  $z$ -transformation  $\hat{\theta} = \tanh^{-1}(r)$  can be applied to the correlation coefficients, resulting in the transformed estimates being asymptotically normal with variance which is only a function of the study sample size  $n$ , i. e.,  $\text{Var}(\hat{\theta}) = 1/(n - 3)$  (Fisher, 1921).

The first data set comprises the results from the *Reproducibility Project: Psychology* (Open Science Collaboration, 2015), whose aim was to replicate 100 studies, all of which were published in three major Psychology journals in 2008. For our purpose only the 73 study pairs from the “meta-analytic” subset are considered, since only for these studies the standard error of the Fisher  $z$ -transformed effect estimates can be computed (Johnson et al., 2016). The second data set comes from the *Experimental Economics Replication Project* (Camerer et al., 2016) which attempted to replicate 18 experimental economics studies published in two high impact economics journals between 2011 and 2015. The third data stem from the *Social Sciences Replication Project* (Camerer et al., 2018) where 21 replications of studies on the social sciences were carried out, all of which were originally published in the journals *Nature* and *Science* between 2010 and 2015. The last data set originates from the *Experimental Philosophy Replicability Project* (Cova et al., 2018) which involved 40 replications of studies from the emerging field of experimental philosophy. Since only for 31 studies effective sample size for original and replication study were available simultaneously, only these pairs were included. For more information on the data sets see also Pawel and Held (2020).

**Table 1:** Results for each replication project: Relative effect size  $d$  (median with 25% and 75% quantiles on Fisher’s  $z$  scale), proportion of successful replications with the two-trials rule (2TR) and the replication success (RS) approach (at the golden level), and number of studies where the methods disagree.

Project	relative effect size $d$	2TR (%)	RS (%)	discrepant
Psychology	0.29 [0.03, 0.77]	28.8	30.1	3/73
Experimental Economics	0.67 [0.35, 0.92]	55.6	55.6	0/18
Social Sciences	0.52 [0.13, 0.65]	61.9	52.4	2/21
Experimental Philosophy	0.86 [0.47, 1.12]	74.2	71.0	1/31

Table 1 presents overall results for each of the replication projects. While the median relative effect size is below one for all of the four projects, there are still large differences. For example, the median relative effect size is only 0.29 in the Psychology project, whereas it is 0.86 in the Philosophy project. The degree of shrinkage is also reflected in the success rates (according to the two-trials rule and the replication success approach at the golden level), which are around 30% for the former and more than 70% for the latter. The proportion of successful replications is similar for the two-trials rule and the replication success approach. In the Experimental Economics project the methods perfectly agree, while in the other three projects the methods disagree for a few studies.



**Figure 6:** Relative effect size  $d$  versus original  $p$ -value  $p_o$ . Black indicates that replication success was achieved at the golden level while grey indicates that it was not. The diamonds mark studies where the replication success approach (at the golden level) and the two-trials rule disagree. The dashed black line indicates the limiting relative effect size at the golden level with  $\alpha = 0.025$ .

Figure 6 displays the relative effect size  $d$  versus the original  $p$ -value  $p_o$  for each study pair and stratified by project. Note that one study pair from the Philosophy project is not shown due to extremely small original  $p$ -value and another study pair from the Psychology project is not shown due to a very large relative effect size. We can see that for most of the study pairs, the replication success approach and the two-trials rule lead to the same conclusion, only six replications show conflicting results. They are highlighted with diamonds in Figure 6 and their characteristics are summarised in Table 2. Two studies from the Psychology project show replication success but fail the two-trials rule. These studies show  $p$ -values that are slightly above the significance threshold in either original or replication study, but do not exhibit much shrinkage; in the replication of Oberauer (2008), the replication  $p$ -value was  $p_r = 0.035$ , a little too large to pass the two-trials rule. However, as the replication effect estimate shrunk only about 30% compared to the original one, replication success is still achieved. Conversely, the



**Table 2:** Characteristics of studies for which the replication success approach (at the golden level) and the two-trials rule disagree (at one-sided  $\alpha = 0.025$ ). Shown are relative sample size  $c$ , relative effect size  $d$ , original, replication and recalibrated sceptical  $p$ -value  $p_o$ ,  $p_r$  and  $\tilde{p}_s$ .

Study	Project	$c$	$d$	$p_o$	$p_r$	$\tilde{p}_s$
Schmidt and Besner (2008)	Psychology	2.58	1.28	<b>0.028</b>	< 0.0001	0.024
Oberauer (2008)	Psychology	0.60	0.67	0.0003	<b>0.035</b>	0.017
Payne et al. (2008)	Psychology	2.65	0.41	0.001	0.023	<b>0.031</b>
Balafoutas and Sutter (2012)	Social Sciences	3.48	0.52	0.009	0.011	<b>0.04</b>
Pyc and Rawson (2010)	Social Sciences	9.18	0.38	0.011	0.004	<b>0.061</b>
Nichols (2006)	Experimental Philosophy	9.40	0.49	0.015	0.0006	<b>0.049</b>

original  $p$ -value  $p_o = 0.028$  in Schmidt and Besner (2008) was just above the significance level, yet the replication led to a highly significant result  $p_r < 0.0001$  with the effect estimate being even 30% larger than the original counterpart, which therefore also resulted in replication success.

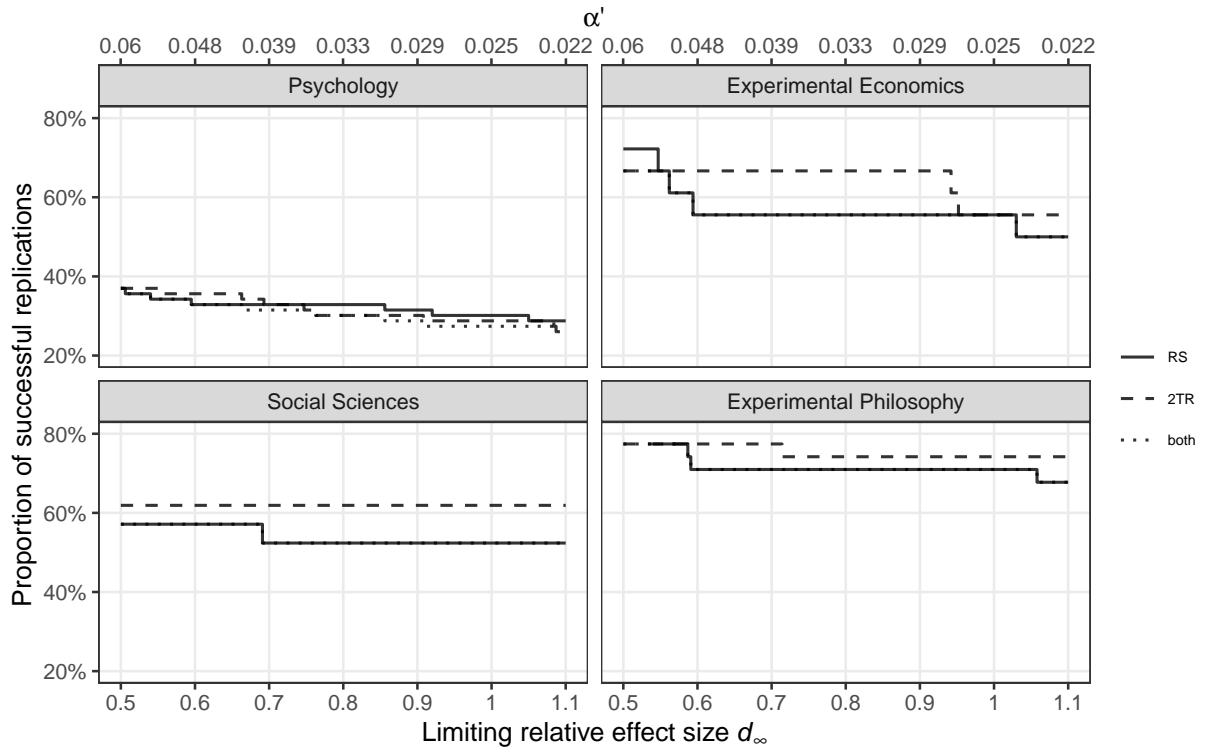
The remaining conflicting studies do not show replication success despite passing the two-trials rule. In all cases, there is substantial shrinkage of the replication effect estimate compared to the original one. For instance, in the replication study of Pyc and Rawson (2010), the estimate shrunk by 62% and the replication  $p$ -value was only significant because the sample size was increased by a factor of  $c = 9.2$ .

This analysis was based on the default choice  $d_\infty = 1$  at  $\alpha = 0.025$  for the golden level as described in Section 2.2. We may also choose a different value for the limiting relative effect size  $d_\infty$  at  $\alpha = 0.025$  which then corresponds to  $d_\infty = 1$  at a different level  $\alpha'$  as given in (9). Figure 7 compares the proportion of successful replications with the replication success approach for  $d_\infty \in (0.5, 1.1)$  with the two-trials rule at the corresponding levels  $\alpha' \in (0.06, 0.022)$  for all four replication projects. We can see that the two proportions agree fairly well for all values of  $\alpha'$  considered. The number of discrepant studies in each project varies between 0 and 3. Only in the Psychology project there are some studies which are successful with the replication success approach but not the two-trials rule and some studies successful with the two-trials rule but not the replication success approach. The proportion of studies where both methods are successful (also shown in Figure 7) is then smaller than the proportion of successful replications with either one of the two methods. The three discrepant studies from the Psychology project listed in the top three rows of Table 2 are an example of this particular feature.

## 5 Discussion

In this paper, we have expanded on the replication success approach introduced in Held (2020b) and demonstrated its advantages over alternative methods such as the two-trials rule. In particular, the method provides an attractive compromise between hypothesis testing and





**Figure 7:** Proportion of successful replications as a function of the limiting relative effect size  $d_\infty$  at  $\alpha = 0.025$ . The upper axis gives the equivalent level  $\alpha'$  where the corresponding limiting relative effect size is 1. The replication success (RS) approach is compared with the two-trials rule (2TR).

estimation, as it penalizes shrinkage of the replication effect estimate compared to the original one, while ensuring that both are statistically significant to some extent.

We further refined the method by proposing the golden level, a new threshold for replication success. It guarantees that borderline significant original studies can only be replicated successfully if the replication effect estimate is larger than the original one. Compared to the two-trials rule, the golden level offers uniform gains in project power and controls the Type-I error rate at any one-sided level  $\alpha < 0.058$  if the replication sample size is not smaller than the original one. Empirical evaluation of data from four replication projects highlights that in most cases the methods are in agreement, however, for the study pairs where the approaches disagree, the replication success approach seems to lead to more sensible conclusions. The good performance has been recently confirmed by a comparison of different replication success metrics through a simulation study in the presence of publication bias (Muradchanian et al., 2021).

Despite a lack of agreement as to which statistical method should be used to evaluate replication studies, conclusions based on different methods usually agree. Nevertheless, in some cases, classical methods such as the two-trials rule may produce anomalies. We argue that the replication success approach improves upon existing methods leading to more appropriate inferences and decisions that better reflect the available evidence. However, in extreme cases the performance of the sceptical  $p$ -value may be considered as strange or even counterintuitive.

---

Specifically, if the original study was only borderline significant, a highly significant replication study can only lead to success if the replication effect estimate is larger than the original one. To understand this behaviour it is important to realize that the proposed approach does not synthesize the evidence from the two studies (like a standard meta-analysis). The sceptical  $p$ -value is designed to confirm claims of new discoveries through replication, but will remain “stubborn” (Ly and Wagenmakers, 2020) if the original study was not particularly convincing, even if the replication study provides overwhelming evidence for an effect. It will lead to a different result if the order of studies was reversed, as long as original and replication study do not have the same sample size ( $c \neq 1$ ). The related harmonic mean  $\chi^2$ -test (Held, 2020a) for evidence synthesis of two or more studies also requires each study to be convincing on its own to a certain degree, but treats them as exchangeable.

With this paper we further advanced the reverse-Bayes methodology for the analysis and design of replication studies, yet certain limitations and opportunities for future research remain: First, assuming normality of the effect estimates may be questionable, especially for small sample sizes, and more robust distributional assumptions could be considered. Second, in some types of analyses (e.g., regression or ANOVA) the effect estimate is a vector and the approach would need suitable adaptations. Third, there is a recent trend to not only conduct one but several replications for one original study (e.g., Klein et al., 2014; Ebersole et al., 2016; Klein et al., 2018). Also for this situation, the method would need to be adapted, e.g., the replication estimates could be first synthesized and an analysis of replication success could be performed subsequently.

Throughout the paper we have assumed that the relative sample size is fixed in advance. In practice the sample size of the replication study is often chosen based on the result of the original study (Anderson and Maxwell, 2017). Power calculations as shown in Figure 3 can then be inverted to determine the appropriate sample size of the replication study. We can also invert equation (6) to obtain the required replication sample size based on the specification of the minimum relative effect size  $d_{\min}$  to achieve replication success. This novel way of calculating the sample size requires the specification of the minimum relative effect size which can still be considered as acceptable. Sample size calculations based on the two-trials rule can also be formulated in terms of the minimum relative effect size by inverting equation (2). We will report on a detailed comparison of the different approaches in future work.

## Data and Software

Code and data to reproduce our analyses are available at <https://github.com/SamCH93/RGolden/>. A snapshot of the Git repository at the time of writing is archived at <https://doi.org/10.5281/zenodo.XXXXXX>. The data which was used is available in the R package ReplicationSuccess (available on CRAN). Further information can be found on the corresponding help page (with the command `?RProjects`).

---

## Acknowledgments

Support by the Swiss National Science Foundation (Project # 189295) is gratefully acknowledged. We acknowledge helpful and constructive comments by the editor and a referee on an earlier version of this article.

## Bibliography

- Anderson, S. F. and Maxwell, S. E. (2017). Addressing the “replication crisis”: Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research*, 52(3):305–324. doi:[10.1080/00273171.2017.1289361](https://doi.org/10.1080/00273171.2017.1289361).
- 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](https://doi.org/10.1126/science.1211180).
- Begley, C. G. and Ioannidis, J. P. (2015). Reproducibility in science. *Circulation Research*, 116(1):116–126. doi:[10.1161/circresaha.114.303819](https://doi.org/10.1161/circresaha.114.303819).
- Box, G. E. P. (1980). Sampling and Bayes’ inference in scientific modelling and robustness (with discussion). *Journal of the Royal Statistical Society: Series A (General)*, 143:383–430. doi:[10.2307/2982063](https://doi.org/10.2307/2982063).
- 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](https://doi.org/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](https://doi.org/10.1038/s41562-018-0399-z).
- 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](https://doi.org/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](https://doi.org/10.1080/01621459.1982.10477856).
- Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B., Boucher, L., et al. (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67:68–82. doi:[10.1016/j.jesp.2015.10.012](https://doi.org/10.1016/j.jesp.2015.10.012).
- 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](https://doi.org/10.7554/elife.04333).
- FDA (1998). Providing clinical evidence of effectiveness for human drug and biological products. URL [www.fda.gov/regulatory-information/search-fda-guidance-documents/providing-clinical-evidence-effectiveness-human-drug-and-biological-products](http://www.fda.gov/regulatory-information/search-fda-guidance-documents/providing-clinical-evidence-effectiveness-human-drug-and-biological-products).

- 
- Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32.
- 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](https://doi.org/10.1002/pst.1736).
- Held, L. (2020a). The harmonic mean  $\chi^2$ -test to substantiate scientific findings. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(3):697–708. doi:[10.1111/rssc.12410](https://doi.org/10.1111/rssc.12410).
- Held, L. (2020b). A new standard for the analysis and design of replication studies (with discussion). *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(2):431–448. doi:[10.1111/rssa.12493](https://doi.org/10.1111/rssa.12493).
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8):e124. doi:[10.1371/journal.pmed.0020124](https://doi.org/10.1371/journal.pmed.0020124).
- 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](https://doi.org/10.1080/01621459.2016.1240079).
- Kay, R. (2015). *Statistical Thinking for Non-Statisticians in Drug Regulation*. John Wiley & Sons, Chichester, U.K., second edition. doi:[10.1002/9781118451885](https://doi.org/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](https://doi.org/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](https://doi.org/10.1177/2515245918810225).
- Ly, A. and Wagenmakers, E.-J. (2020). Discussion of “A new standard for the analysis and design of replication studies” by Leonhard Held. *Journal of the Royal Statistical Society, Series A*, 183:460–461. doi:[10.1111/rssa.12544](https://doi.org/10.1111/rssa.12544).
- Maca, J., Gallo, P., Branson, M., and Maurer, W. (2002). Reconsidering some aspects of the two-trials paradigm. *Journal of Biopharmaceutical Statistics*, 12(2):107–119. doi:[10.1081/bip-120006450](https://doi.org/10.1081/bip-120006450).
- Matthews, J. N. (2006). *Introduction to Randomized Controlled Clinical Trials*. Chapman and Hall/CRC, New York. doi:[10.1201/9781420011302](https://doi.org/10.1201/9781420011302).
- Matthews, R. A. J. (2001a). Methods for assessing the credibility of clinical trial outcomes. *Drug Information Journal*, 35:1469–1478. doi:[10.1177/009286150103500442](https://doi.org/10.1177/009286150103500442).
- Matthews, R. A. J. (2001b). Why should clinicians care about Bayesian methods? *Journal of Statistical Planning and Inference*, 94:43–71. doi:[10.1016/S0378-3758\(00\)00232-9](https://doi.org/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](https://doi.org/10.1214/21-sts828).
-

- 
- Muradchianian, J., Hoekstra, R., Kiers, H., and van Ravenzwaaij, D. (2021). How best to quantify replication success? A simulation study on the comparison of replication success metrics. *Royal Society Open Science*, 8:201697. doi:[10.1098/rsos.201697](https://doi.org/10.1098/rsos.201697).
- Nichols, S. (2006). Folk intuitions on free will. *Journal of Cognition and Culture*, 6(1-2):57–86. doi:[10.1163/156853706776931385](https://doi.org/10.1163/156853706776931385).
- Oberauer, K. (2008). How to say no: Single- and dual-process theories of short-term recognition tested on negative probes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(3):439–459. doi:[10.1037/0278-7393.34.3.439](https://doi.org/10.1037/0278-7393.34.3.439). URL <https://doi.org/10.1037/0278-7393.34.3.439>. <https://doi.org/10.1037/0278-7393.34.3.439>.
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716. doi:[10.1126/science.aac4716](https://doi.org/10.1126/science.aac4716).
- Pawel, S. and Held, L. (2020). Probabilistic forecasting of replication studies. *PLOS ONE*, 15(4):e0231416. doi:[10.1371/journal.pone.0231416](https://doi.org/10.1371/journal.pone.0231416).
- Payne, B. K., Burkley, M. A., and Stokes, M. B. (2008). Why do implicit and explicit attitude tests diverge? the role of structural fit. *Journal of Personality and Social Psychology*, 94(1):16–31. doi:[10.1037/0022-3514.94.1.16](https://doi.org/10.1037/0022-3514.94.1.16).
- Pyc, M. A. and Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, 330(6002):335–335. doi:[10.1126/science.1191465](https://doi.org/10.1126/science.1191465).
- Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172. doi:[10.1214/aos/1176346785](https://doi.org/10.1214/aos/1176346785).
- Schmidt, J. R. and Besner, D. (2008). The Stroop effect: Why proportion congruent has nothing to do with congruency and everything to do with contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(3):514–523. doi:[10.1037/0278-7393.34.3.514](https://doi.org/10.1037/0278-7393.34.3.514).
- Senn, S. (2008). *Statistical issues in drug development*, volume 69. John Wiley & Sons.



---

**Bayesian approaches to designing replication studies**

*Samuel Pawel, Guido Consonni, Leonhard Held*

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

---





---

## Abstract

Replication studies are essential for assessing the credibility of claims from original studies. A critical aspect of designing replication studies is determining their sample size; a too small sample size may lead to inconclusive studies whereas a too large sample size may waste resources that could be allocated better in other studies. Here we show how Bayesian approaches can be used for tackling this problem. The Bayesian framework allows researchers to combine the original data and external knowledge in a design prior distribution for the underlying parameters. Based on a design prior, predictions about the replication data can be made, and the replication sample size can be chosen to ensure a sufficiently high probability of replication success. Replication success may be defined through Bayesian or non-Bayesian criteria, and different criteria may also be combined to meet distinct stakeholders and allow conclusive inferences based on multiple analysis approaches. We investigate sample size determination in the normal-normal hierarchical model where analytical results are available and traditional sample size determination is a special case where the uncertainty on parameter values is not accounted for. An application to data from a multisite replication project of social-behavioral experiments illustrates how Bayesian approaches help to design informative and cost-effective replication studies. Our methods can be used through the R package BayesRepDesign.

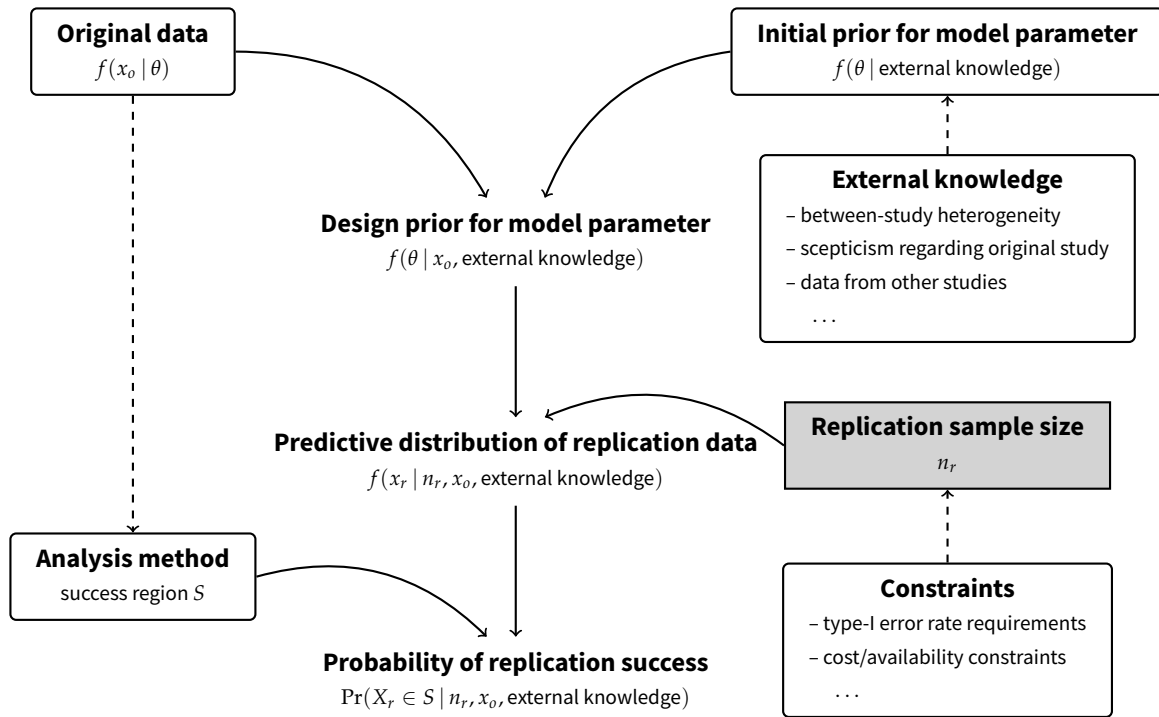
**Key words:** Bayesian design, design prior, multisite replication, sample size determination

## 1 Introduction

The replicability of research findings is a cornerstone for the credibility of science. However, there is growing evidence that the replicability of many scientific findings is lower than expected (Ioannidis, 2005; Open Science Collaboration, 2015; Camerer et al., 2018; Errington et al., 2021). This “replication crisis” has led to methodological reforms in various fields of science, one of which is an increased conduct of replication studies (Munafò et al., 2017). Statistical methodology plays a key role in the evaluation of replication studies, and various methods have been proposed for quantifying how “successful” a replication study was in replicating the original finding (Bayarri and Mayoral, 2002; 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; Held et al., 2022b; Pawel and Held, 2022, among others). Yet, as with ordinary studies, statistical methodology is not only important for analyzing replication studies but also for designing them, in particular for their *sample size determination* (SSD). Optimal SSD is important since too small sample sizes may lead to inconclusive studies, whereas too large sample sizes may waste resources which could have been allocated better in other research projects.

SSD for replication studies comes with unique opportunities and challenges; the data from the original study can be used to inform SSD, at the same time the analysis of replication success based on original and replication study is typically different from an analysis of a single

study for which traditional SSD methodology was developed. Since analysis and design of replication studies should be in accordance, a relatively small literature has emerged which specifically deals with replication study power calculations and SSD (Bayarri and Mayoral, 2002; Goodman, 1992; Senn, 2002; Anderson and Maxwell, 2017; Micheloud and Held, 2022; van Zwet and Goodman, 2022; Held, 2020; Pawel and Held, 2022; Hedges and Schauer, 2021; Anderson and Kelley, 2022). However, most of these articles only deal with selected analysis methods and data models. An exception is the excellent article by Anderson and Kelley (2022) which discusses more general principles of replication SSD in the context of psychological research, mostly from a frequentist perspective. As they state “the literature on Bayesian sample size planning is still nascent, particularly with respect to Bayes Factors (Schönbrodt and Wagenmakers, 2017), and has not yet been clearly optimized for the context of most replication goals” (Anderson and Kelley, 2022, p. 18). Our goal is therefore to complement their article by developing a unified framework of replication SSD (schematically illustrated in Figure 1) based on principles from Bayesian design approaches (Spiegelhalter et al., 1986; Spiegelhalter and Freedman, 1986; Weiss, 1997; O’Hagan and Stevens, 2001; Gelfand and Wang, 2002; De Santis, 2004; Spiegelhalter et al., 2004; Schönbrodt and Wagenmakers, 2017; Pek and Park, 2019; Kunzmann et al., 2021; Park and Pek, 2022; Grieve, 2022). We aim to provide both a theoretical basis for methodologists developing new methods for design and analysis methods of replication studies, and also to illustrate how Bayesian design approaches can practically be used by researchers planning a replication study.



**Figure 1:** Schematic illustration of Bayesian sample size determination for replication studies. The original and replication data are denoted by  $x_o$  and  $x_r$ , respectively. Both are assumed to come from a distribution with density/probability mass function denoted by  $f(x_i | \theta)$  for  $i \in \{o, r\}$ . An initial prior with density function  $f(\theta | \text{external knowledge})$  is assigned to the model parameter  $\theta$ .

---

The design of replication studies is a natural candidate for Bayesian knowledge updating as it allows to combine uncertain information from different sources –for instance, the data from the original study and/or expert knowledge– in a so-called *design prior* distribution for the underlying model parameters. If the analysis of the replication data is also Bayesian, the design prior may be different from the so-called *analysis prior* which, unlike the design prior, is usually desired to be objective or “uninformative” (O’Hagan and Stevens, 2001). Based on the design prior, predictions about the replication data can then be made, and the sample size can be chosen such that the probability of replication success becomes sufficiently high. Importantly, Bayesian design approaches can also be used if the planned analysis of the replication study is non-Bayesian, which is the more common situation in practice. Bayesian design based on a frequentist analysis is known under various names, such as “hybrid classical-Bayesian design” (Spiegelhalter et al., 2004) or “Bayesian assurance” (O’Hagan et al., 2005), and has also been used before for psychological applications (Pek and Park, 2019; Park and Pek, 2022) and replication studies (Anderson and Maxwell, 2017; Micheloud and Held, 2022).

This paper is structured as follows: We start with presenting a general framework for Bayesian SSD of replication studies which applies to any kind of data model and analysis method. We then investigate design priors and SSD in the normal-normal hierarchical model framework which provides sufficient flexibility for incorporating the original data and external knowledge in replication design. No advanced computational methods, such as (Markov Chain) Monte Carlo sampling, are required for conducting Bayesian SSD in this framework, and in many cases there are even simple formulae which generalize classical power and sample size calculations. We illustrate the methodology for several Bayesian and non-Bayesian analysis methods, and for both singlesite and multisite replication studies. Since multisite replication studies are becoming increasingly popular in psychology (e.g., Klein et al., 2018), we also discuss how to choose the optimum allocation of samples within and between sites from a Bayesian design point of view. As a running example we use data from a multisite replication project of social-behavioral experiments (Protzko et al., 2020). Finally, we close with concluding remarks, limitations, and open questions.

## 2 General framework

Suppose an original study has been conducted and resulted in a data set  $x_o$ . These data are assumed to come from a distribution characterized by an unknown parameter  $\theta$  and with density function  $f(x_o | \theta)$ . To assess the replicability of a claim from the original study, an independent and identically designed (apart from the sample size) replication study is conducted, and the goal of the design stage is to determine its sample size  $n_r$ .

As the observed original data  $x_o$ , the yet unobserved replication data  $X_r$  are assumed to come from a distribution depending on the parameter  $\theta$ . The parameter  $\theta$  thus provides a link between the two studies, and the knowledge obtained from the original study can be used to make predictions about the replication. The central quantity for doing so is the so-called *design prior* of the parameter  $\theta$ , which we write as the posterior distribution of  $\theta$  based on the

---

original data and an *initial prior* for  $\theta$

$$f(\theta | x_o, \text{external knowledge}) = \frac{f(x_o | \theta) f(\theta | \text{external knowledge})}{f(x_o | \text{external knowledge})}. \quad (1)$$

The initial prior of  $\theta$  may depend on external knowledge (e. g., data from other studies) and it represents the uncertainty about  $\theta$  before observing the original data. We will discuss common types of external knowledge in the replication setting in the next Section. The design prior (1) hence represents the state of knowledge and uncertainty about the parameter  $\theta$  before the replication is conducted, and, along with an assumed replication sample size  $n_r$ , it can be used to compute a predictive distribution for the replication data

$$f(x_r | n_r, x_o, \text{external knowledge}) = \int f(x_r | n_r, \theta) f(\theta | x_o, \text{external knowledge}) d\theta. \quad (2)$$

After completion of the replication, the observed data  $x_r$  will be analyzed in some way to quantify to what extent the original result could be replicated. The analysis may involve the original data (for example, a meta-analysis of the two data sets) or it may only use the replication data. Typically, there is a *success region*  $S$  which implies that if the replication data fall within it ( $x_r \in S$ ), the replication is successful. The *probability of replication success* can thus be computed by integrating the predictive density (2) over  $S$ . To ensure a sufficiently conclusive replication design, the sample size  $n_r$  is determined such that the probability of replication success is at least as large as a desired target probability of success, here and henceforth denoted by  $1 - \beta$ . The required sample size  $n_r^*$  is then the smallest sample size which leads to a probability of replication success of at least  $1 - \beta$ , i. e.,

$$n_r^* = \inf \{n_r : \Pr(X_r \in S | n_r, x_o, \text{external knowledge}) \geq 1 - \beta\}. \quad (3)$$

Often, replication studies are analyzed using several methods which quantify different aspects of replicability, and which have different success regions (e. g., one method for quantifying parameter compatibility and another for quantifying evidence against a null hypothesis). In this case, the sample size may be chosen such that the probability of replication success is as large as desired for all planned analysis methods.

There may sometimes be certain constraints which the replication sample size needs to satisfy. For instance, in most cases there is an upper limit on the possible sample size due to limited resources and/or availability of samples. Moreover, funders and regulators may also require methods to be *calibrated* (Grieve, 2016), that is, to have appropriate type I error rate control. The sample size  $n_r^*$  may thus also need to satisfy a type I error rate not larger than some required level.

### 3 Sample size determination in the normal-normal hierarchical model

We will now illustrate the general methodology from the previous section in the *normal-normal hierarchical model* where predictive distributions and the probability of replication success can

often be expressed in closed-form, permitting further insight. It is pragmatic to adopt a meta-analytic perspective and use only study level summary statistics instead of the raw study data since the raw data from the original study are not always available to the replicators. Typically, the underlying parameter  $\theta$  is a univariate effect size quantifying the effect on the outcome variable (e. g., a mean difference, a log odds ratio, or a log hazard ratio). The original and replication study can then be summarized through an effect estimate  $\hat{\theta}$ , possibly the maximum likelihood estimate, and a corresponding standard error  $\sigma$ , i. e.,  $x_o = \{\hat{\theta}_o, \sigma_o\}$  and  $x_r = \{\hat{\theta}_r, \sigma_r\}$ . Effect estimates and standard errors are routinely reported in research articles or can, under some assumptions, be computed from  $p$ -values and confidence intervals. As in the conventional meta-analytic framework (Sutton and Abrams, 2001), we further assume that for study  $k \in \{o, r\}$  the (suitably transformed) effect estimate  $\hat{\theta}_k$  is approximately normally distributed around a study specific effect size  $\theta_k$  and with (known) variance equal to its squared standard error  $\sigma_k^2$ , here and henceforth denoted by  $\hat{\theta}_k | \theta_k \sim N(\theta_k, \sigma_k^2)$ . The standard error  $\sigma_k$  is typically of the form  $\sigma_k = \lambda / \sqrt{n_k}$  with  $\lambda^2$  some unit variance and  $n_k$  the sample size. The ratio of the original to the replication variance is thus the ratio of the replication to the original sample size

$$c = \sigma_o^2 / \sigma_r^2 = n_r / n_o,$$

which is often the main focus of SSD as it quantifies how much the replication sample  $n_r$  size needs to be changed compared to the original sample size  $n_o$ . Depending on the effect size type, this framework might require slight modifications (see e. g., Spiegelhalter et al., 2004, ch. 2.4).

Assuming a normal sampling model for the effect estimates (4a), as described previously, and specifying an initial hierarchical normal prior for the study specific effect sizes (4b) and the effect size (4c), leads then to the normal-normal hierarchical model

$$\hat{\theta}_k | \theta_k \sim N(\theta_k, \sigma_k^2) \tag{4a}$$

$$\theta_k | \theta \sim N(\theta, \tau^2) \tag{4b}$$

$$\theta \sim N(\mu_\theta, \sigma_\theta^2). \tag{4c}$$

By marginalizing over the study specific effects sizes, the model (4) can alternatively be expressed as

$$\hat{\theta}_k | \theta \sim N(\theta, \sigma_k^2 + \tau^2) \tag{5a}$$

$$\theta \sim N(\mu_\theta, \sigma_\theta^2) \tag{5b}$$

which is often more useful for derivations and computations. In the following we will explain how the normal-normal hierarchical model can be used for SSD of the replication study.

### 3.1 Design prior and predictive distribution

The observed original data  $x_o = \{\hat{\theta}_o, \sigma_o\}$  can be combined with the initial prior (5b) by standard Bayesian theory for normal prior and likelihood (Spiegelhalter et al., 2004, ch. 3.7) to

---

obtain a posterior distribution for the effect size  $\theta$

$$\theta | \hat{\theta}_o, \sigma_o^2 \sim N \left( \frac{\hat{\theta}_o}{1 + 1/g} + \frac{\mu_\theta}{1 + g}, \frac{\sigma_o^2 + \tau^2}{1 + 1/g} \right) \quad (6)$$

where  $g = \sigma_\theta^2 / (\sigma_o^2 + \tau^2)$  is the *relative prior variance*. This posterior serves then as the design prior for predicting the replication data.

It is interesting to contrast the design prior (6) to the “conditional” design prior (Micheloud and Held, 2022), that is, to assume that the unknown effect size  $\theta$  corresponds to the original effect estimate  $\hat{\theta}_o$ . This is a standard approach in practice, for instance, Open Science Collaboration (2015) determined the sample sizes of its 100 replications under this assumption. In our framework it implies that the normal design prior (6) becomes a point mass at the original effect estimate  $\hat{\theta}_o$ , which can either be achieved through overwhelmingly informative original data ( $\sigma_o^2 \downarrow 0$ ) along with no heterogeneity ( $\tau^2 = 0$ ), or through an overwhelmingly informative initial prior ( $g \downarrow 0$ ) centered around the original effect estimate ( $\mu_\theta = \hat{\theta}_o$ ). Both cases show that from a Bayesian perspective the standard approach is unnatural as it either corresponds to making the standard error  $\sigma_o$  smaller than it actually was, or to cherry-picking the prior based on the data.

Based on the design prior (6), the predictive distribution of the replication effect estimate  $\hat{\theta}_r$  can then be computed by assuming a replication standard error  $\sigma_r$  and integrating the marginal density of the replication effect estimate (5a) with respect to the prior density, leading to

$$\hat{\theta}_r | \hat{\theta}_o, \sigma_o^2, \sigma_r^2 \sim N \left( \mu_{\hat{\theta}_r} = \frac{\hat{\theta}_o}{1 + 1/g} + \frac{\mu_\theta}{1 + g}, \sigma_{\hat{\theta}_r}^2 = \sigma_r^2 + \tau^2 + \frac{\sigma_o^2 + \tau^2}{1 + 1/g} \right) \quad (7)$$

which can again be shown using standard Bayesian theory (Spiegelhalter et al., 2004, ch. 3.13.3). The design prior (6) and the resulting predictive distribution (7) depend on the parameters of the initial prior ( $\tau^2, \mu_\theta, \sigma_\theta^2$ ). We will now explain how these parameters can be specified based on external knowledge.

### 3.2 Incorporating external knowledge in the initial prior

At least three common types of external knowledge can be distinguished in the replication setting: (i) expected heterogeneity between original and replication study due to differences in study design, execution, and population, (ii) prior knowledge about the effect size either from theory or from related studies, (iii) scepticism regarding the original study due to the possibility of exaggerated results.

#### Between-study heterogeneity

The expected degree of between-study heterogeneity can be incorporated via the variance  $\tau^2$  in (4b). As  $\tau^2$  decreases, the study specific effect sizes become more similar, whereas for increasing  $\tau^2$  they become more unrelated. If the replicators do not expect any heterogeneity they can thus set  $\tau^2 = 0$  which will lead to the model collapsing to a fixed effects model.



---

If heterogeneity is expected, there are different approaches for specifying  $\tau^2$ . A domain expert may subjectively assess how much heterogeneity is to be expected due to the change in laboratory, study population, and other factors. An alternative is to take an estimate from the literature, e. g., from multisite replication projects or from systematic reviews. Finally, one can also specify an upper limit of “tolerable heterogeneity”. This approach is similar to specifying a minimal clinically relevant difference in classical power analysis in the sense that a true replication effect size which is intolerably heterogeneous from the original effect size is not relevant to be detected. An absolute (Spiegelhalter et al., 2004, ch. 5.7.3) and a relative approach (Held and Pawel, 2020) can be considered. In the absolute approach, a value of  $\tau^2$  is chosen such that a suitable range of study-specific effect sizes is not larger than an effect size difference considered negligible. For example, when 95% of the study specific effect sizes should not vary more than a small effect size e. g.,  $d = 0.2$  on standardized mean difference scale based on the Cohen (1992) effect size classification, this would lead to  $\tau = d / (2 \cdot 1.96) \approx 0.05$ . In the relative approach,  $\tau^2$  is specified relative to the variance of the original estimate  $\sigma_o^2$  using field conventions for tolerable relative heterogeneity. For example, in the Cochrane guidelines for systematic reviews (Deeks et al., 2019) a value of  $I^2 = \tau^2 / (\tau^2 + \sigma_o^2) = 40\%$  is classified as “negligible”, which translates to  $\tau^2 = \sigma_o^2 / (1/I^2 - 1) = (2\sigma_o^2)/3$ .

We note that one can also assign a prior distribution to  $\tau^2$  (for an overview of prior distributions for heterogeneity variances in the normal-normal hierarchical model see Röver et al., 2021). In this case there is no closed-form expression for the predictive distribution of the replication effect estimate but numerical or Monte Carlo integration need to be used. We illustrate in the supplement how the probability of replication success can in this case be computed. The derived closed-form expressions conditional on  $\tau^2$  are still useful as they enable computation of the predictive distribution up to a single one-dimensional integral which can be computed numerically.

### Knowledge about the effect size

Prior knowledge about the effect size  $\theta$  can be incorporated via the prior mean  $\mu_\theta$  and the prior variance  $\sigma_\theta^2$  in (4c). For instance, the parameters may be specified based on a meta-analysis of related studies (McKinney et al., 2021) or based on expert elicitation (O’Hagan, 2019). The resulting design prior will then contain more information than what was provided by the original data alone, leading to potentially more efficient designs. If there is no prior knowledge available, a standard approach is to specify an (improper) flat prior by letting the variance go to infinity ( $\sigma_\theta^2 \rightarrow \infty$ ). The resulting design prior will then only contain the information from the original study.

### Exaggerated original results

Potentially exaggerated original results can be counteracted by setting  $\mu_\theta = 0$  which shrinks the design prior towards smaller effect sizes (in absolute value) than the observed effect estimate  $\hat{\theta}_o$ . For instance, replicators could believe that the results from the original study are exaggerated because there is no pre-registered study protocol available. Even without such

beliefs, weakly informative shrinkage priors may also be motivated from a “regularization” point of view as they can correct for statistical biases (Copas, 1983; Firth, 1993) or prevent unreasonable parameter values from taking over the posterior in settings with uninformative data (Gelman, 2009).

The amount of shrinkage is determined via the prior variance  $\sigma_\theta^2$ . A diffuse prior ( $\sigma_\theta^2 \rightarrow \infty$ ) will lead to no shrinkage, while a highly concentrated prior ( $\sigma_\theta^2 \downarrow 0$ ) will completely shrink the design prior to a point mass on zero. One option for specifying  $\sigma_\theta^2$  is to use an estimate from a corpus of related studies. For instance, Zwet et al. (2021) used the Cochrane library of systematic reviews to specify design priors for hypothetical replication studies of RCTs. If no corpus is available, a pragmatic alternative is to use the empirical Bayes estimate based on the original data

$$\hat{\sigma}_\theta^2 = \max\{(\hat{\theta}_o - \mu_\theta)^2 - \tau^2 - \sigma_o^2, 0\}. \quad (8)$$

The estimate (8) will lead to adaptive shrinkage (Pawel and Held, 2020) in the sense that shrinkage is large for unconvincing original studies (those with small effect estimates in absolute value  $|\hat{\theta}_o|$  and/or large standard errors  $\sigma_o$ ), but disappears as the data become more convincing (through larger effect estimates in absolute value  $|\hat{\theta}_o|$  and/or smaller standard errors  $\sigma_o$ ).

### 3.3 Example: Cross-laboratory replication project

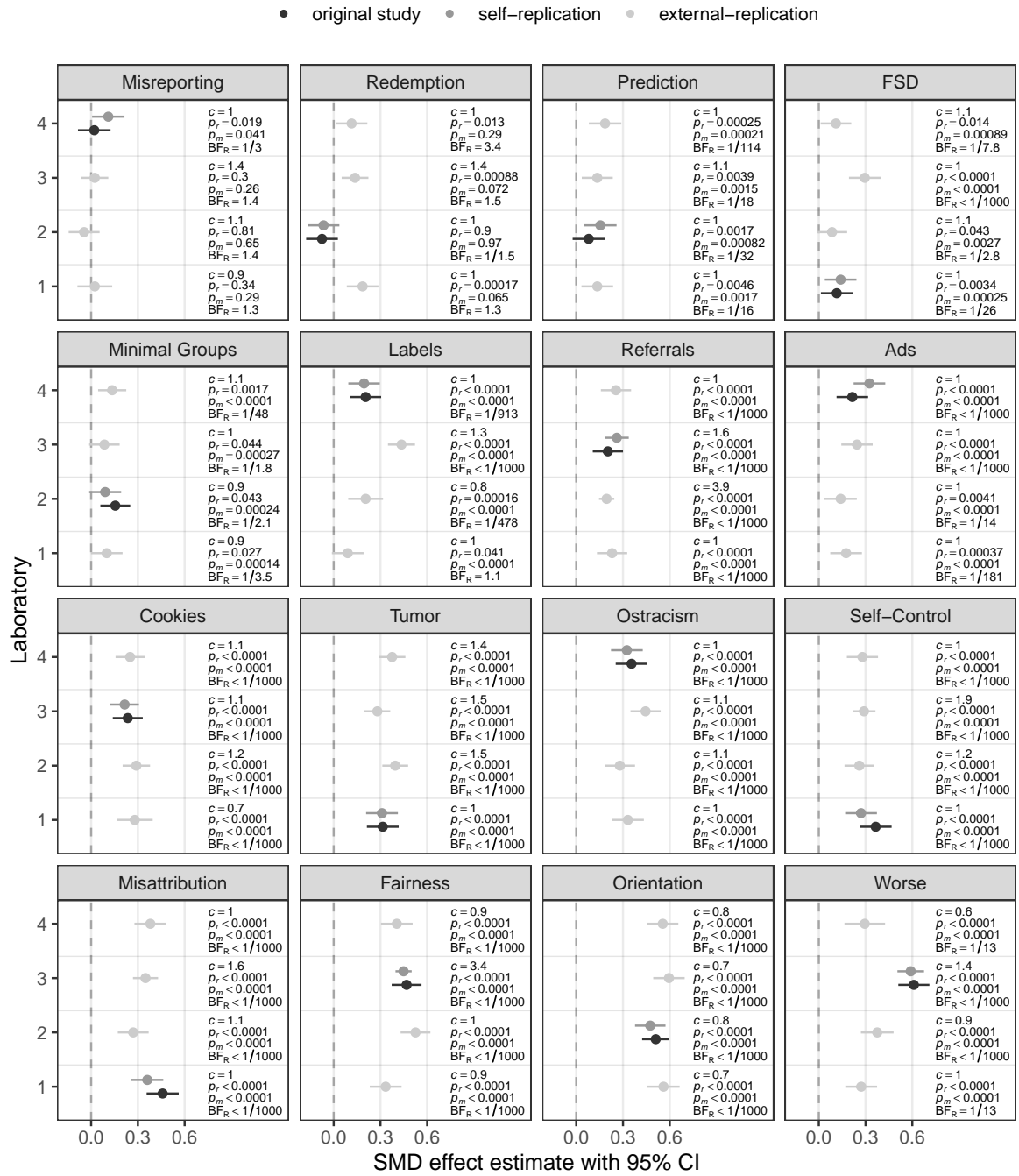
We will now illustrate the construction of design priors based on data from a recently conducted replication project (Protzko et al., 2020), see Figure 2 for a summary of the data. The data were collected in four laboratories which, over the course of five years, conducted their typical social-behavioral experiments on topics such as psychology, communication, or political science. From the experiments conducted in this period, each lab submitted four original findings to be replicated. For instance, the original finding from the “Labels” experiment was: “When a researcher uses a label to describe people who hold a certain opinion, he or she is interpreted as disagreeing with those attributes when a negative label is used and agreeing with those attributes when a positive label is used” (Protzko et al., 2020, p. 17), which was based on an effect estimate  $\hat{\theta}_o = 0.205$  with 95% confidence interval from 0.11 to 0.3. For each submitted original finding, four replication studies were then carried out, one by the same lab (a *self-replication*) and three by the other three labs (three *external-replications*).

Most studies used simple between-subject designs with two groups and a continuous outcome so that for a study  $i \in \{o, r\}$  the standardized mean difference (SMD) effect estimate  $\hat{\theta}_i$  can be computed from the group means  $\bar{y}_{i1}, \bar{y}_{i2}$ , group standard deviations  $s_{i1}, s_{i2}$ , and group sample sizes  $n_{i1}, n_{i2}$  by

$$\hat{\theta}_i = \frac{\bar{y}_{i1} - \bar{y}_{i2}}{s_i}$$

with  $s_i^2 = \{(n_{i1} - 1)s_{i1}^2 + (n_{i2} - 1)s_{i2}^2\} / (n_{i1} + n_{i2} - 2)$  the pooled sample variance. Under a normal sampling model and assuming equal variances in both groups, the approximate





**Figure 2:** Data from cross-laboratory replication project by Protzko et al. (2020). Shown are standardized mean difference (SMD) effect estimates with 95% confidence intervals stratified by experiment and laboratory. For each replication study, the relative sample size  $c = n_r/n_o$ , the one-sided replication  $p$ -value  $p_r$ , the one-sided meta-analytic  $p$ -value  $p_m$ , and the replication Bayes factor  $BF_R$  are shown. Experiments are ordered (left to right, top to bottom) by their original one-sided  $p$ -value  $p_o = 1 - \Phi(|\hat{\theta}_o|/\sigma_o)$

variance of  $\hat{\theta}_i$  is

$$\sigma_i^2 = \frac{n_{i1} + n_{i2}}{n_{i1}n_{i2}} + \frac{\hat{\theta}_i^2}{2(n_{i1} + n_{i2})} \quad (9)$$

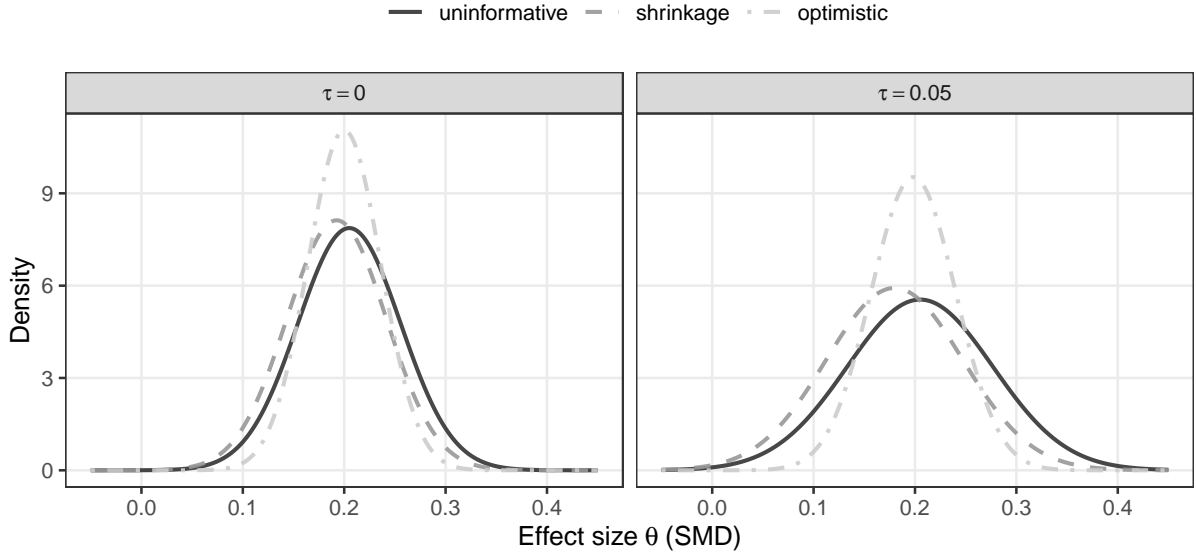
(Hedges, 1981). A cruder, but for SSD more useful, approximation  $\sigma_i^2 \approx 4/n_i$  is obtained by assuming the same sample size in both groups  $n_{i1} = n_{i2} = n_i/2$ , with  $n_i$  the total sample size, and neglecting the second term in (9) which will be close to zero for small effect estimates and/or large sample sizes (Hedges and Schauer, 2021). We thus have the approximate unit variance  $\lambda^2 = 4$  and the relative variance  $c = \sigma_o^2/\sigma_r^2 = n_r/n_o$ , which can be interpreted as the ratio of the replication to the original sample size.

Suppose now the original studies have been finished, and we want to conduct SSD for the not yet conducted replication studies. We start by specifying the design priors (one for each replication). Since the original studies have been preregistered, we do not expect an exaggeration of their effect estimates due to selective reporting or other questionable research practices. Therefore, we choose an uninformative initial prior ( $\sigma_\theta^2 \rightarrow \infty$ ), which leads to design prior and predictive distribution both centered around the original effect estimate  $\hat{\theta}_o$ .

For specifying the between-study heterogeneity variance  $\tau^2$ , a distinction needs to be made between self-replications and external-replications. For self-replications it is reasonable to set  $\tau^2 = 0$  because we would expect no between-study heterogeneity as the experimental conditions will be nearly identical in both studies. In contrast, one would expect some between-study heterogeneity for external-replications as the experimental conditions may slightly differ between the labs. In the following, we will use  $\tau^2 = 0.05^2$  elicited via the “absolute” approach as discussed previously, so that the range between the 2.5% and the 97.5% quantile of the study specific effect size distribution is equal to a small effect size  $d = 0.2$ .

Taken together, we obtain the design prior  $\theta | \hat{\theta}_o, \sigma_o^2 \sim N(\hat{\theta}_o, \sigma_o^2)$  for self-replications and the design prior  $\theta | \hat{\theta}_o, \sigma_o^2 \sim N(\hat{\theta}_o, \sigma_o^2 + \tau^2)$  with  $\tau^2 = 0.05^2$  for external-replications. For example, for the experiment “Labels”, the design prior would be centered around the original effect estimate  $\hat{\theta}_o = 0.205$  with variance  $\sigma_o^2 = 0.05^2$  for a self-replication, and with variance  $\sigma_o^2 + \tau^2 = 0.05^2 + 0.05^2 \approx 0.07^2$  for an external-replication. Figure 3 (dark-gray solid lines) shows the densities of the two priors.

While these two priors seem sensible for the Protzko et al. (2020) data, it is interesting to think about alternative scenarios. If there had been reasons to believe that the original result might be exaggerated, we could have specified an initial shrinkage prior ( $\mu_\theta = 0$  and  $\sigma_\theta^2 < \infty$ ). For instance, the empirical Bayes estimate for the prior variance  $\sigma_\theta^2$  from (8) leads to a prior whose mean and variance are shrunken towards zero by 12% (medium-gray dashed lines in Figure 3). In contrast, if we had prior knowledge about the effect size  $\theta$  from another study, we could have specified an initial “optimistic” prior. For example, if the self-replication of the “Labels” experiment had been a pilot study and we used its effect estimate  $\hat{\theta}_p = 0.195$  and standard error  $\sigma_p = 0.05$  to specify the initial prior, this would lead to a design prior centered around the weighted mean of original and pilot study, and a prior precision equal to the sum of the precision of both estimates (light-gray dot-dashed lines in Figure 3). Due to the inclusion of the external data, this design prior is much more concentrated than the other two.



**Figure 3:** Design priors for the effect size  $\theta$  (SMD) in the experiment “Labels” based on the original effect estimate  $\hat{\theta}_o = 0.205$  with standard error  $\sigma_o = 0.051$ . Shown are different choices for the between-study heterogeneity  $\tau$  and the initial prior for the effect size  $\theta$ , “uninformative” corresponds to a flat prior, “shrinkage” corresponds to a zero-mean normal prior with empirical Bayes variance estimate (8), and “optimistic” corresponds to a flat prior updated by the data from a pilot study with effect estimate  $\hat{\theta}_p = 0.195$  and standard error  $\sigma_p = 0.052$ .

### 3.4 Probability of replication success and required sample size

To compute the probability of replication success one needs to select an analysis method and integrate the predictive distribution (7) over the associated success region  $S$ . There is no universally accepted method for quantifying replicability and here we do not intend to contribute to the debate about the most appropriate method. We will simply show the success regions of different methods, and how the replication sample size can be computed from them. Some methods depend on the direction of the original effect estimate  $\hat{\theta}_o$  and throughout we will assume that it was positive ( $\hat{\theta}_o > 0$ ). Functions for computing the probability of replication success and the required sample size are implemented in the R package **BayesRepDesign** (see Appendix) for all analysis methods discussed in the following.

#### The two-trials rule

The most common approach for the analysis of replication studies is to declare replication success when both the original and replication study lead to a  $p$ -value for testing the null hypothesis  $H_0: \theta = 0$  smaller than a pre-specified threshold  $\alpha$ , usually  $\alpha = 5\%$  for two-sided tests and  $\alpha = 2.5\%$  for one-sided tests. This procedure is known as the *two-trials rule* in drug regulation (Senn, 2008).

We now assume that the one-sided original  $p$ -value was significant at some level  $\alpha$ , i. e.,  $p_o = 1 - \Phi(\hat{\theta}_o/\sigma_o) \leq \alpha$ . Replication success at level  $\alpha$  is then achieved if the replication  $p$ -value is also significant, i. e.,  $p_r = 1 - \Phi(\hat{\theta}_r/\sigma_r) \leq \alpha$ , which implies a success region

$$S_{2TR} = [z_\alpha \sigma_r, \infty),$$

where  $z_\alpha$  is the  $1 - \alpha$  quantile of the standard normal distribution. The probability of replication success is thus given by

$$\Pr(\hat{\theta}_r \in S_{2TR} | \hat{\theta}_o, \sigma_o, \sigma_r) = \Phi\left(\frac{\mu_{\hat{\theta}_r} - z_\alpha \sigma_r}{\sigma_{\hat{\theta}_r}}\right) \quad (10)$$

with  $\Phi(\cdot)$  the standard normal cumulative distribution function and  $\mu_{\hat{\theta}_r}$  and  $\sigma_{\hat{\theta}_r}$  the mean and standard deviation of the predictive distribution (7). Importantly, by decreasing the standard error  $\sigma_r$  (through increasing the sample size  $n_r$ ), the probability of replication success (10) cannot become arbitrarily large but is bounded from above by

$$\lim \Pr_{2TR} = \Phi\left(\frac{\mu_{\hat{\theta}_r}}{\sqrt{\tau^2 + (\sigma_o^2 + \tau^2)/(1 + 1/g)}}\right). \quad (11)$$

The required replication standard error  $\sigma_r^*$  to achieve a target probability of replication success  $1 - \beta < \lim \Pr_{2TR}$  can now be obtained by equating (10) to  $1 - \beta$  and solving for  $\sigma_r$ . This leads to

$$\sigma_r^* = \frac{\mu_{\hat{\theta}_r} z_\alpha - z_\beta \sqrt{(z_\alpha^2 - z_\beta^2) \{ \tau^2 + (\sigma_o^2 + \tau^2)/(1 + 1/g) \} + \mu_{\hat{\theta}_r}^2}}{z_\alpha^2 - z_\beta^2} \quad (12)$$

for  $\alpha < \beta$ . The standard error  $\sigma_r^*$  can subsequently be translated in a sample size. The translation depends on the type of effect size, for instance, for SMD effect sizes we can use the approximation  $n_r^* \approx 4/(\sigma_r^*)^2$  from earlier. Moreover, by assuming a standard error of the form  $\sigma_r = \lambda/\sqrt{n_r}$  and plugging in the parameters of the “conditional” design prior ( $\tau^2 = 0$ ,  $\mu_\theta = \hat{\theta}_o$ ,  $g \downarrow 0$ ), we obtain the well-known sample size formula (Matthews, 2006, ch. 3.3)

$$n_r^* = \frac{(z_\alpha + z_\beta)^2}{(\hat{\theta}_o/\lambda)^2}$$

for a one-sided significance test at level  $\alpha$  with power  $1 - \beta$  to detect the original effect estimate  $\hat{\theta}_o$ . The formula (12) thus generalizes standard sample size calculation to take into account the uncertainty of the original estimate, between-study heterogeneity and other types of external knowledge.

### Fixed effects meta-analysis

The data from the original and replication studies are sometimes pooled via fixed-effects meta-analysis. The pooled effect estimate  $\hat{\theta}_m$  and standard error  $\sigma_m$  are then given by

$$\hat{\theta}_m = (\hat{\theta}_o/\sigma_o^2 + \hat{\theta}_r/\sigma_r^2) \sigma_m^2 \quad \text{and} \quad \sigma_m = (1/\sigma_o^2 + 1/\sigma_r^2)^{-1/2},$$

and they are also equivalent to the mean and standard deviation of a posterior distribution for the effect size  $\theta$  based on the data from both studies and an initial flat prior for  $\theta$ . The success region

$$S_{MA} = \left[ \sigma_r z_\alpha \sqrt{1 + \sigma_r^2 / \sigma_o^2} - (\hat{\theta}_o \sigma_r^2) / \sigma_o^2, \infty \right) \quad (13)$$

then corresponds to both replication success defined via a one-sided meta-analytic  $p$ -value being smaller than level  $\alpha$ , i.e.,  $p_m = 1 - \Phi(\hat{\theta}_m / \sigma_m) \leq \alpha$ , or to replication success defined via a Bayesian posterior probability  $\Pr(\theta > 0 | \hat{\theta}_o, \hat{\theta}_r, \sigma_o, \sigma_r) \geq 1 - \alpha$ . Based on the success region (13) and an assumed standard error  $\sigma_r$ , the probability of replication success can be computed by

$$\Pr(\hat{\theta}_r \in S_{MA} | \hat{\theta}_o, \sigma_o, \sigma_r) = \Phi \left( \frac{\mu_{\hat{\theta}_r} - \sigma_r z_\alpha \sqrt{1 + \sigma_r^2 / \sigma_o^2} + (\hat{\theta}_o \sigma_r^2) / \sigma_o^2}{\sigma_{\hat{\theta}_r}} \right). \quad (14)$$

As for the two-trials rule, the probability (14) cannot be made arbitrarily large by decreasing the standard error  $\sigma_r$  but is bounded from above by  $\lim \Pr_{2TR}$  defined in (11). The required standard error  $\sigma_r^*$  to achieve a target probability of replication success  $1 - \beta < \lim \Pr_{2TR}$  can be computed numerically using root finding algorithms.

### Effect size equivalence test

Anderson and Maxwell (2016) proposed a method for quantifying replicability based on effect size equivalence. Under normality, replication success at level  $\alpha$  is achieved if the  $(1 - \alpha)$  confidence interval for the effect size difference  $\theta_r - \theta_o$

$$\hat{\theta}_r - \hat{\theta}_o \pm z_{\alpha/2} \sqrt{\sigma_r^2 + \sigma_o^2}$$

is fully inside an equivalence region  $[-\Delta, \Delta]$  defined via the pre-specified margin  $\Delta > 0$ . This procedure corresponds to rejecting the null hypothesis  $H_0: |\theta_r - \theta_o| > \Delta$  in an equivalence test, and it implies a success region for the replication effect estimate  $\hat{\theta}_r$  given by

$$S_E = \left[ \hat{\theta}_o - \Delta + z_{\alpha/2} \sqrt{\sigma_o^2 + \sigma_r^2}, \hat{\theta}_o + \Delta - z_{\alpha/2} \sqrt{\sigma_o^2 + \sigma_r^2} \right] \quad (15)$$

for  $\Delta \geq z_{\alpha/2} \sqrt{\sigma_o^2 + \sigma_r^2}$ . For too small margins ( $\Delta < z_{\alpha/2} \sqrt{\sigma_o^2 + \sigma_r^2}$ ), the success region (15) becomes the empty set meaning that replication success is impossible. Assuming now that the margin is large enough, the probability of replication success can be computed by

$$\begin{aligned} \Pr(\hat{\theta}_r \in S_E | \hat{\theta}_o, \sigma_o, \sigma_r) &= \Phi \left( \frac{\hat{\theta}_o + \Delta - z_{\alpha/2} \sqrt{\sigma_o^2 + \sigma_r^2} - \mu_{\hat{\theta}_r}}{\sigma_{\hat{\theta}_r}} \right) \\ &\quad - \Phi \left( \frac{\hat{\theta}_o - \Delta + z_{\alpha/2} \sqrt{\sigma_o^2 + \sigma_r^2} - \mu_{\hat{\theta}_r}}{\sigma_{\hat{\theta}_r}} \right). \end{aligned} \quad (16)$$

As with the previous methods, the probability (16) cannot be made arbitrarily large by decreasing the replication standard error  $\sigma_r$ , but is bounded by

$$\lim \Pr_E = \Phi \left( \frac{\hat{\theta}_o + \Delta - z_{\alpha/2} \sigma_o - \mu_{\hat{\theta}_r}}{\sqrt{\tau^2 + (\sigma_o^2 + \tau^2)/(1 + 1/g)}} \right) - \Phi \left( \frac{\hat{\theta}_o - \Delta + z_{\alpha/2} \sigma_o - \mu_{\hat{\theta}_r}}{\sqrt{\tau^2 + (\sigma_o^2 + \tau^2)/(1 + 1/g)}} \right).$$

The required replication standard error  $\sigma_r^*$  to achieve a target probability of replication success  $1 - \beta < \lim \Pr_E$  can again be computed numerically.

### The replication Bayes factor

A Bayesian hypothesis testing approach for assessing replication success was proposed by [Verhagen and Wagenmakers \(2014\)](#) and further developed by [Ly et al. \(2018\)](#). They define a “replication Bayes factor”

$$\text{BF}_R = \frac{f(x_r | H_0)}{f(x_r | H_1)}$$

which is the ratio of the marginal likelihoods of the replication data  $x_r$  under the null hypothesis  $H_0: \theta = 0$  to the marginal likelihood of  $x_r$  under the alternative hypothesis  $H_1: \theta \sim f(\theta | x_o)$ , that is the posterior of the effect size  $\theta$  based on the original data  $x_o$ . If the original study provides evidence against the null hypothesis, replication Bayes factor values  $\text{BF}_R < 1$  indicate replication success, and the smaller the value the higher the degree of success.

Under normality and assuming no heterogeneity, the success region for achieving  $\text{BF}_R \leq \gamma$  is given by

$$S_{\text{BF}_R} = \left( -\infty, -\sqrt{A} - (\hat{\theta}_o \sigma_r^2) / \sigma_o^2 \right] \cup \left[ \sqrt{A} - (\hat{\theta}_o \sigma_r^2) / \sigma_o^2, \infty \right) \quad (17)$$

with  $A = \sigma_r^2(1 + \sigma_r^2/\sigma_o^2)\{\hat{\theta}_o^2/\sigma_o^2 - 2\log \gamma + \log(1 + \sigma_o^2/\sigma_r^2)\}$ . Details of this calculation are given in the supplement. The fact that the success region (17) is defined on both sides around zero shows that replication success is also possible if the replication effect estimate goes in opposite direction of the original one, which is known as the “replication paradox” ([Ly et al., 2018](#)). The paradox can be avoided using a modified version of the replication Bayes factor but the success region is no longer available in closed-form ([Pawel and Held, 2022](#), Appendix D). Based on the success region (17), the probability of replication success can be computed by

$$\Pr(\hat{\theta}_r \in S_{\text{BF}_R} | \hat{\theta}_o, \sigma_o, \sigma_r) = \Phi \left( \frac{\mu_{\hat{\theta}_r} - \sqrt{A} + (\hat{\theta}_o \sigma_r^2) / \sigma_o^2}{\sigma_{\hat{\theta}_r}} \right) + \Phi \left( \frac{-\sqrt{A} - (\hat{\theta}_o \sigma_r^2) / \sigma_o^2 - \mu_{\hat{\theta}_r}}{\sigma_{\hat{\theta}_r}} \right). \quad (18)$$

One may want to compute the probability of replication success only for the part of the success region with the same sign as the original effect estimate to avoid the replication paradox. As for the other methods, the probability (18) is bounded from above by a constant  $\lim \Pr_{\text{BF}_R} = \lim_{\sigma_r \downarrow 0} \Pr(\hat{\theta}_r \in S_{\text{BF}_R} | \hat{\theta}_o, \sigma_o, \sigma_r)$ , and root finding algorithms can be used to numerically determine the required standard error  $\sigma_r^*$  for achieving a target probability of replication success  $1 - \beta < \lim \Pr_{\text{BF}_R}$ .

---

## The sceptical $p$ -value

Held (2020) proposed a reverse-Bayes approach for quantifying replication success. The main idea is to determine the variance of a “sceptical” zero-mean normal prior for the effect size  $\theta$  such that its posterior distribution based on the original study is no longer credible. Replication success is then achieved if the replication data are in conflict with the sceptical prior. The procedure can be summarized by a “sceptical  $p$ -value”  $p_S$ , and the lower the  $p$ -value the higher the degree of replication success. Held et al. (2022b, sec. 2.1) showed that the success region for replication success defined by  $p_S \leq \alpha$  is given by

$$S_{p_S} = \left[ z_\alpha \sqrt{\sigma_r^2 + \frac{\sigma_o^2}{(z_o^2/z_\alpha^2) - 1}}, \infty \right). \quad (19)$$

From the success region (19) the probability of replication success at level  $\alpha$  is

$$\Pr(\hat{\theta}_r \in S_{p_S} \mid \hat{\theta}_o, \sigma_o, \sigma_r) = \Phi \left( \frac{\mu_{\hat{\theta}_r} - z_\alpha \sqrt{\sigma_r^2 + \sigma_o^2 / \{(z_o^2/z_\alpha^2) - 1\}}}{\sigma_{\hat{\theta}_r}} \right),$$

and also bounded from above by a constant  $\lim_{p_S} \Pr(\hat{\theta}_r \in S_{p_S} \mid \hat{\theta}_o, \sigma_o, \sigma_r) = \lim_{\sigma_r \downarrow 0} \Pr(\hat{\theta}_r \in S_{p_S} \mid \hat{\theta}_o, \sigma_o, \sigma_r)$ . As for the two-trials rule, the required standard error  $\sigma_r^*$  to achieve a probability of replication success  $1 - \beta < \lim_{p_S}$  can be computed analytically for  $\alpha < \beta$ :

$$\sigma_r^* = \sqrt{x^2 - \frac{\sigma_o^2}{(z_o/z_\alpha)^2 - 1}}$$

with

$$x = \frac{z_\alpha \mu_{\hat{\theta}_r} - z_\beta \sqrt{\mu_{\hat{\theta}_r}^2 - (z_\alpha^2 - z_\beta^2)[\tau^2 + (\sigma_o^2 + \tau^2)/(1 + 1/g) - \sigma_o^2 / \{(z_o/z_\alpha)^2 - 1\}]}}{z_\alpha^2 - z_\beta^2}.$$

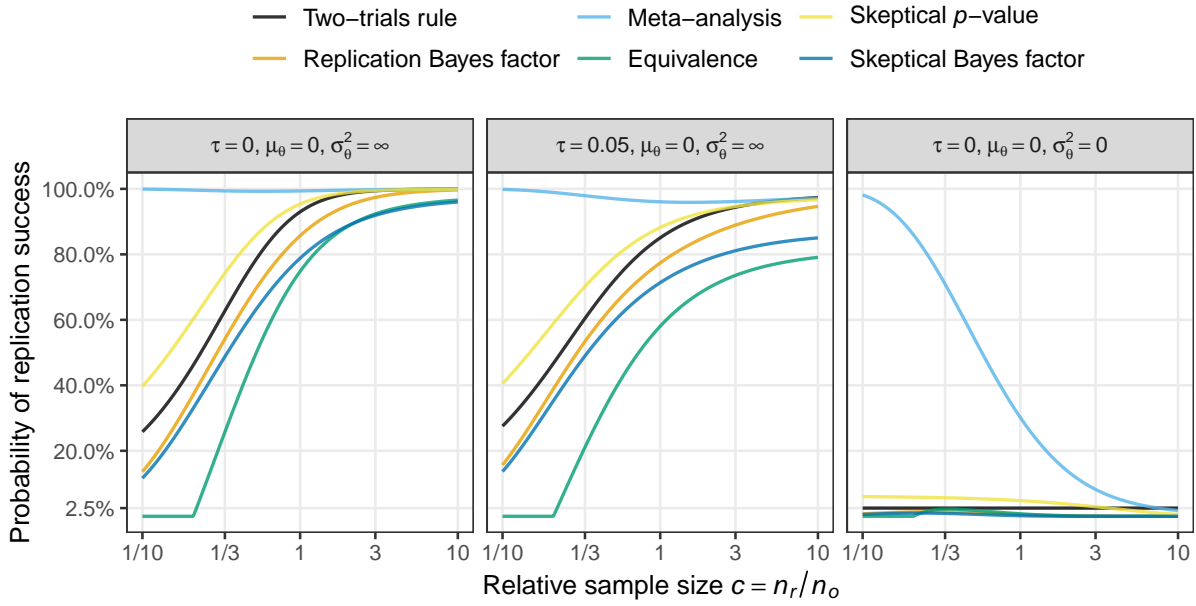
## The sceptical Bayes factor

Pawel and Held (2022) modified the previously described reverse-Bayes assessment of replication success from Held (2020) to work with Bayes factors instead of tail probabilities as measures of evidence and prior data conflict. Again, the procedure can be summarized in a single measure termed the “sceptical Bayes factor”  $BF_S$ , with lower values of  $BF_S$  pointing to higher degrees of replication success. Also for this method, the success region and the probability of replication success can be expressed in closed-form but the derivations are more involved than for the other methods. For this reason, they are only given in the supplement.

### 3.5 Example: Cross-laboratory replication project (continued)

We will now revisit the experiment “Labels” and compute the probability of replication success. The parameters of the analysis methods are specified as follows: For the two-trials rule

we use the conventional one-sided significance level  $\alpha = 0.025$ , while for meta-analysis we use the more stringent level  $\alpha = 0.025^2$  as the method is based on two data sets rather than one. We use a  $1 - \alpha = 90\%$  confidence interval which is conventionally used in equivalence testing, along with a margin  $\Delta = 0.2$  corresponding to a small SMD effect size according to the classification from [Cohen \(1992\)](#). For the sceptical  $p$ -value we use the recommended “golden” level  $\alpha = 0.062$  as it guarantees that for original studies which were just significant at  $\alpha = 0.025$  replication success is only possible if the replication effect estimate is larger than the original one ([Held et al., 2022b](#)). Finally, for the replication Bayes factor and the sceptical Bayes factor we use the “strong evidence” level  $\gamma = 1/10$  from [Jeffreys \(1961\)](#).



**Figure 4:** Probability of replication success as a function of relative sample size  $c = n_r/n_o$  for experiment “Labels” with original effect estimate  $\hat{\theta}_o = 0.205$  and standard error  $\sigma_o = 0.051$  for different initial prior parameters  $(\tau, \mu_\theta, \sigma_\theta^2)$ . The probability of replication success with the design prior  $\tau = 0, \mu_\theta = 0$ , and  $\sigma_\theta^2 = 0$  (right plot) corresponds to the type I error rate under the fixed effects null hypothesis ( $H_0: \theta = 0, \tau^2 = 0$ ). Replication success is defined by the two-trials rule at level  $\alpha = 0.025$ , the replication Bayes factor at level  $\gamma = 1/10$ , fixed effects-meta analysis at level  $\alpha = 0.025^2$ , effect size equivalence based on 90% confidence interval and with margin  $\Delta = 0.2$ , sceptical  $p$ -value at level  $\alpha = 0.062$ , and sceptical Bayes factor at level  $\gamma = 1/10$ .

Figure 4 shows the probability of replication success as a function of the relative sample size  $c = n_r/n_o$  and for different initial priors. The left and middle plot are based on an uninformative prior for the effect size ( $\sigma_\theta^2 \rightarrow \infty$ ) without heterogeneity ( $\tau^2 = 0$ ) and with heterogeneity ( $\tau^2 = 0.05^2$ ), respectively. The right plot shows the prior corresponding to the “fixed effects null hypothesis”  $H_0: \theta = 0, \tau^2 = 0$ , so that the probability of replication success is the type I error rate which some stakeholders might require to be “controlled” at some adequate level.

We see from the left and middle plots that increasing the relative sample size monotonically

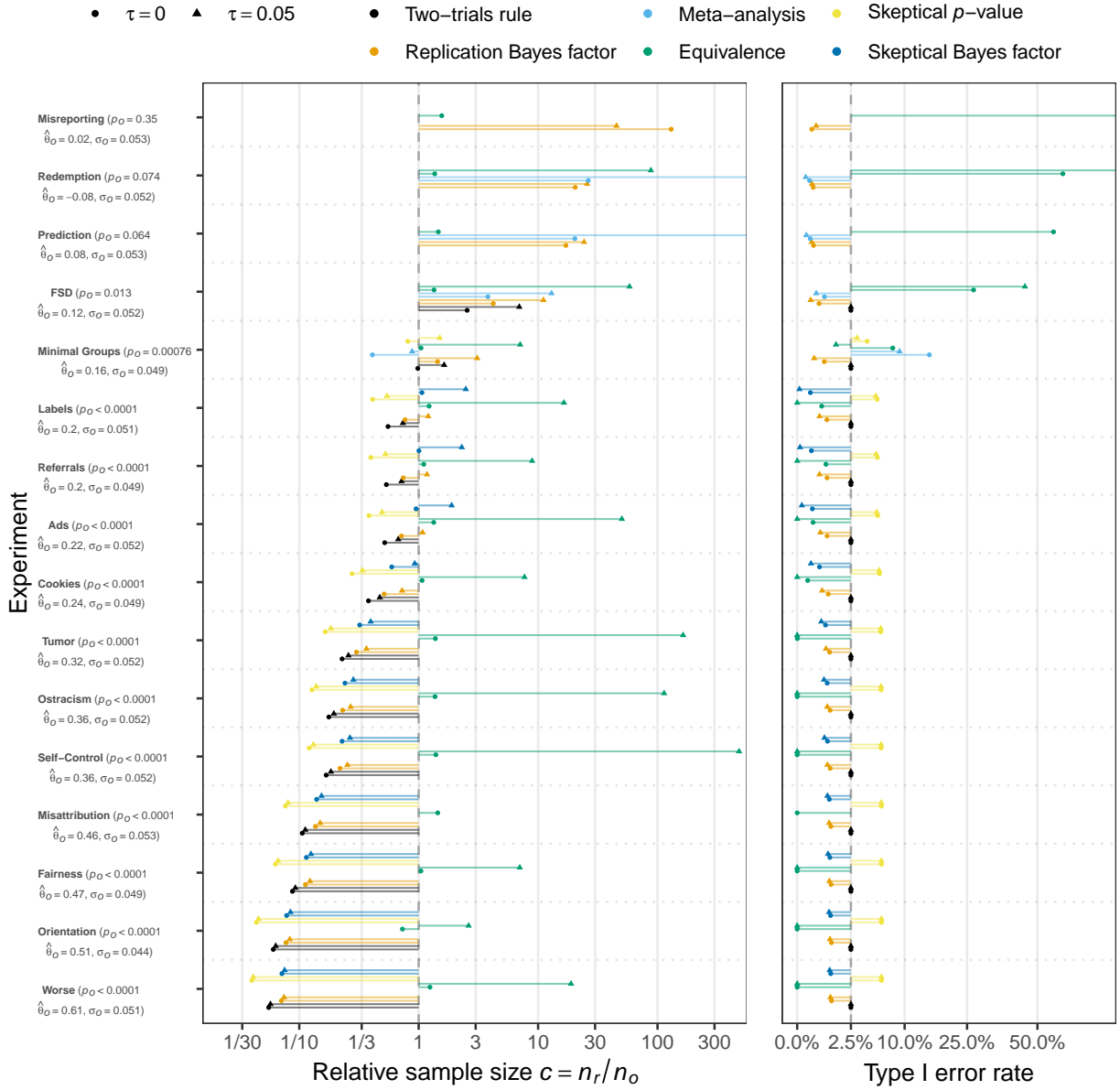


increases the probability of replication success for all methods but meta-analysis (light blue). Meta-analysis shows a non-monotone behavior because the original study was already highly significant so that the pooled effect estimate is significant even for replication studies with very small sample size (Micheloud and Held, 2022). The uncertainty regarding the replication effect estimate  $\hat{\theta}_r$  may therefore even reduce the probability of replication success for meta-analysis if the sample size is increased. If heterogeneity is taken into account (middle plot) the probability of replication success becomes closer to 50% for all methods but the equivalence test, reflecting the larger uncertainty about the effect size  $\theta$ . To achieve 80% probability of replication success the fewest samples are required with meta-analysis, followed by the sceptical  $p$ -value, the two-trials rule, the replication Bayes factor, the sceptical Bayes factor, and lastly the equivalence test. If the chosen sample size should guarantee a sufficiently conclusive replication study with all these methods, the replication sample size has to be slightly larger than the original one in the situation of no heterogeneity ( $\tau^2 = 0$ ), while it has to be increased more than ten-fold if there is heterogeneity ( $\tau^2 = 0.05^2$ ). However, this is mostly due to the equivalence test which requires by far the most samples. If the equivalence test sample size is ignored, the relative sample size  $c = 2.5$  ensures at least 80% probability of replication success with the remaining methods.

The right plot in Figure 4 shows that the type I error rate of the two-trials rule (black) stays constant at  $\alpha = 0.025$ , as expected by definition of the method. In contrast, the type I error rates of the other methods vary with the relative sample size  $c$  but most of them stay below  $\alpha = 0.025$  for all  $c$  with the exception of meta-analysis and the sceptical  $p$ -value. Meta-analysis (light blue) has an extremely high type I error rate as the pooling with the highly significant original data leads to replication success if the replication sample size is not drastically increased. The type I error rate of the sceptical  $p$ -value (yellow) is only slightly larger than  $\alpha = 0.025$  which is expected since the level  $\alpha = 0.062$  is used for declaring replication success with the sceptical  $p$ -value, and its type I error rate is always smaller than the level for thresholding it (Held, 2020). The type I error rate of the sceptical  $p$ -value decreases to values smaller than  $\alpha = 0.025$  of the two-trials rule at approximately  $c = 3$ .

We now conduct SSD for all studies from the replication project of Protzko et al. (2020). Figure 5 shows the required relative sample size and the associated type I error rates if a sample size can be computed for a probability of replication success of  $1 - \beta = 80\%$ . If there is no sample size for which a probability of 80% can be achieved the space is left blank. This is, for instance, the case for the meta-analysis method for all studies below the “Labels” experiment as the probability stays above 80% for any relative sample size.

We see that for all methods but the equivalence test the required relative sample size  $c$  decreases with decreasing original  $p$ -value  $p_o$ , and original studies with very small  $p$ -values require much fewer samples in the replication study. For the equivalence test, the required sample size depends instead on the size of the original standard error  $\sigma_o$ , and smaller standard errors leads to smaller required sample sizes in the replication. We also see that taking into account heterogeneity (triangle) increases the required sample size for all methods compared to not taking it into account (dot). At the same time, a larger required sample size reduces the type I error rate for most methods. We see again the pattern that the type I error rate of the equivalence test and the sceptical  $p$ -value is larger than the type I error rate 2.5% of the



**Figure 5:** The left plot shows the required relative sample size  $c = n_r/n_o$  to achieve a target probability of replication success of  $1 - \beta = 80\%$  (if possible). Replication success is defined through the two-trials rule at level  $\alpha = 0.025$ , replication Bayes factor at level  $\gamma = 1/10$ , fixed effects-meta analysis at level  $\alpha = 0.025^2$ , effect size equivalence at level  $\alpha = 0.1$  with margin  $\Delta = 0.2$ , sceptical  $p$ -value at level  $\alpha = 0.062$ , and sceptical Bayes factor at level  $\gamma = 1/10$  for data from the replication project by Protzko et al. (2020). A flat initial prior ( $\mu_\theta = 0, \sigma_\theta^2 \rightarrow \infty$ ) is used for the effect size  $\theta$  is used either without ( $\tau = 0$ ) or with heterogeneity ( $\tau = 0.05$ ). The right plot shows the type I error rate associated with the required sample size. Experiments are ordered (top to bottom) by their original one-sided  $p$ -value  $p_o = 1 - \Phi(|\hat{\theta}_o|/\sigma_o)$ .

two-trials rule. However, while the type I error rate of the sceptical  $p$ -value decreases when replication studies require larger sample sizes, the type I error rate of the equivalence test

may also be large if the replication requires very large sample sizes (e. g., for the experiment “FSD”) since it depends on whether the original effect estimate  $\hat{\theta}_o$  is sufficiently different from zero. If the original effect estimate  $\hat{\theta}_o$  is close to zero, the type I error rate of the equivalence test is drastically increased as equivalence can also be established if the effect estimates from original and replication studies are close to zero (as under the null hypothesis).

Taken together, for most of the experiments all methods but the equivalence test require fewer samples in the replication than in the original study to achieve a target probability of replication success  $1 - \beta = 80\%$ . This is still the case if heterogeneity is taken into account in the design prior (triangles), which generally increases the required sample size compared to when heterogeneity is not taken into account (dots), especially for studies with large original  $p$ -value. Larger replication sample sizes are required for some original studies. In some cases these are unrealistically large (e. g., in the experiment “Prediction” an almost 30 times increase in sample size for the replication Bayes factor), but in other cases they seem more realistic and could be reallocated from the other studies which require fewer samples (e. g., in the experiment “Referrals” an almost three times increase for the sceptical Bayes factor). On the other hand, the equivalence test typically requires larger sample sizes in the replication because the original standard errors are large relative to the specified margin. If one anticipates to analyze the original and replication pair with an equivalence test, this should therefore be taken into account already at the design stage of the original study.

### 3.6 Sample size determination for multisite replication projects

So far we considered the situation where a pair of a single original and a single replication study are analyzed in isolation. However, if multiple replications per single original study are conducted (so-called *multisite* replication studies) the ensemble of replications can also be analyzed jointly. In this case, some adaptations of the SSD methodology are required.

The replication effect estimate and its standard error are now vectors  $\hat{\theta}_r = (\hat{\theta}_{r1}, \dots, \hat{\theta}_{rm})^\top$  and  $\sigma_r^2 = (\sigma_{r1}^2, \dots, \sigma_{rm}^2)^\top$  consisting of  $m$  replication effect estimates, respectively, their standard errors. The normal hierarchical model for the replication estimates  $\hat{\theta}_r$  then becomes

$$\hat{\theta}_r | \theta_r \sim N_m \{ \theta_r, \text{diag}(\sigma_r^2) \} \quad (20a)$$

$$\theta_r | \theta \sim N_m \{ \theta \mathbf{1}_m, \tau^2 \text{diag}(\mathbf{1}_m) \}, \quad (20b)$$

where  $\theta_r$  is a vector of  $m$  study specific effect sizes,  $\mathbf{1}_m$  is a vector of  $m$  ones, and  $N_m(\mu, \Sigma)$  denotes the  $m$ -variate normal distribution with mean vector  $\mu$  and covariance matrix  $\Sigma$ . By marginalizing over the study specific effect size  $\theta_k$ , the model can alternatively be expressed by

$$\hat{\theta}_r | \theta \sim N_m \{ \theta \mathbf{1}_m, \text{diag}(\sigma_r^2 + \tau^2 \mathbf{1}_m) \}, \quad (21)$$

so the predictive distribution of  $\hat{\theta}_r$  based on the design prior (6) is given by

$$\hat{\theta}_r | \hat{\theta}_o, \sigma_o^2, \sigma_r^2 \sim N_m \left\{ \mu_{\hat{\theta}_r} \mathbf{1}_m, \text{diag}(\sigma_r^2 + \tau^2 \mathbf{1}_m) + \left( \frac{\tau^2 + \sigma_o^2}{1 + 1/g} \right) \mathbf{1}_m \mathbf{1}_m^\top \right\} \quad (22)$$

with  $\mu_{\hat{\theta}_r}$  the mean of the predictive distribution of a single replication effect estimate from (7). Importantly, the replication effect estimates are correlated as the covariance matrix in (22) has  $(\tau^2 + \sigma_o^2)/(1 + 1/g)$  in the off-diagonal entries.

Often the assessment of replication success can be formulated in terms of a weighted average of the replication effect estimates  $\hat{\theta}_{r*} = (\sum_{i=1}^m w_i \hat{\theta}_{ri}) / (\sum_{i=1}^m w_i)$  with  $w_i$  the weight of replication  $i$ . For instance, several multisite replication projects (e. g., [Klein et al., 2018](#)) have defined replication success by the fixed or random effects meta-analytic effect estimate of the replication effect estimates achieving statistical significance. Based on the predictive distribution of the replication effect estimate vector (22), the predictive distribution of the weighted average  $\hat{\theta}_{r*}$  is given by

$$\hat{\theta}_{r*} | \hat{\theta}_o, \sigma_o^2, \sigma_r^2 \sim N \left\{ \mu_{\hat{\theta}_r}, \sigma_{\hat{\theta}_{r*}}^2 = \left( \sum_{i=1}^m w_i^2 \sigma_{\hat{\theta}_{ri}}^2 + \sum_{i=1}^m \sum_{j=1, j \neq i}^m w_i w_j \frac{\tau^2 + \sigma_o^2}{1 + 1/g} \right) / \left( \sum_{i=1}^m w_i \right)^2 \right\} \quad (23)$$

with  $\sigma_{\hat{\theta}_{ri}}^2$  the predictive variance of a single replication effect estimate with standard error  $\sigma_{ri}$  as in (7). In particular when the studies receive equal weights ( $w_i = w$  for  $i = 1, \dots, m$ ) and the standard errors of the replication effect estimates are equal ( $\sigma_{ri} = \sigma_r$  for  $i = 1, \dots, m$ ), the predictive variance becomes

$$\sigma_{\hat{\theta}_{r*}}^2 = \frac{\sigma_r^2 + \tau^2}{m} + \frac{\tau^2 + \sigma_o^2}{1 + 1/g}. \quad (24)$$

The probability of replication success can now be obtained by integrating (22), respectively (23), over the corresponding success region  $S$ . This may be more involved if the success region is defined in terms of the replication effect estimate vector  $\hat{\theta}_r$ , whereas it is as simple as in the singlesite replication case if the success region is formulated in terms of the weighted average  $\hat{\theta}_{r*}$ .

### Optimal allocation within and between sites

A key challenge in SSD for multisite replication studies is the optimal allocation of samples within and between sites, that is, how many sites  $m$  and how many samples  $n_{ri}$  per site  $i$  should be used. A similar problem exists in SSD for cluster randomized trials and we can adapt the common solution based on cost functions ([Raudenbush, 1997](#)). That is, the optimal configuration is determined so that the probability of replication success is maximized subject to a constrained cost function which accounts for the (typically different) costs of additional samples and sites.

For example, assume a balanced design ( $n_{ri} = n_r$  for  $i = 1, \dots, m$ ) and that the standard errors of the replication effect estimates are inversely proportional to the square-root of the sample size  $\sigma_{ri} = \lambda / \sqrt{n_r}$  for some unit variance  $\lambda^2$ . Further assume that maximizing the probability of replication success corresponds to minimizing the variance of the weighted average  $\sigma_{\hat{\theta}_{r*}}^2$  in (24). Let  $K_s$  denote the cost of an additional site, and  $K_c$  the cost of an additional sample/case.

---

The total cost of the project is then  $K = m(K_c n_r + K_s)$ , and constrained minimization of the predictive variance (24) leads to the optimal sample size per site

$$n_r^* = \frac{\lambda}{\tau} \sqrt{\frac{K_s}{K_c}}$$

which is equivalent to the optimal cluster sample size known from cluster randomized trials (Raudenbush and Liu, 2000). Note that the optimal sample size per site may be different for other analysis approaches where maximizing the probability of replication success does not correspond to minimizing the variance of the weighted average. Moreover, there are also practical considerations which affect the choice of how many sites should be included in a project. For instance, there may simply not be enough labs available with the required expertise to perform the replication experiments.

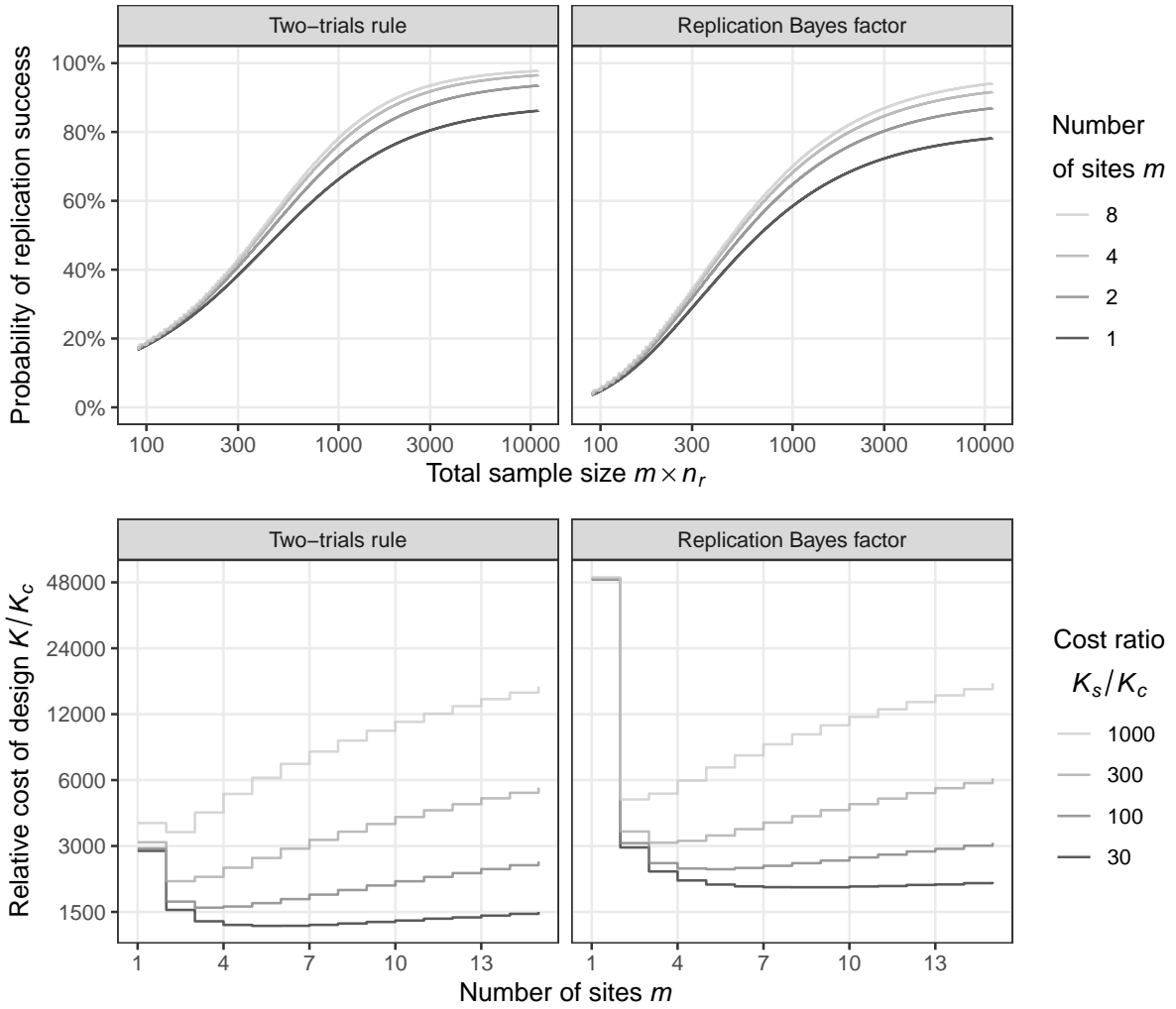
### 3.7 Example: Cross-laboratory replication project (continued)

Figure 6 illustrates multisite SSD for the experiment “Labels” from Protzko et al. (2020) for planned analyses based on the two-trials rule and the replication Bayes factor (see the supplement for details on the multisite extension of these two methods). As for singlesite SSD, we use the design prior based on an initial flat prior for the effect size and taking into account heterogeneity ( $\tau = 0.05$ ). The top plots show the probability of replication success as a function of the total sample size  $m \times n_r$  for different number of sites  $m$ . We see that for the same total sample size a larger number of sites increases the probability of replication success. For instance, a total sample size of roughly 3000 is required to achieve an 80% target probability with one site for the two-trials rule, whereas only approximately half as many samples are required for two sites.

However, focusing only on the total sample size ignores the fact that the cost of an additional site is usually larger than the cost of an additional observation. The bottom plot shows the total cost  $K$  of a design (relative to the cost of one sample  $K_c$ ) whose sample size is determined for a target probability of replication success  $1 - \beta = 80\%$ . We see that if the cost of an additional site  $K_s$  is not much larger than the cost of an additional sample  $K_c$ , e. g.,  $K_s/K_c = 30$  the optimal number of sites is  $m = 5$  for the two-trials rule and  $m = 8$  for the replication Bayes factor. If an additional site is more costly the optimal number of sites is lower, e. g., if the cost ratio is  $K_s/K_c = 300$ , the optimal number of sites is  $m = 2$  for the two trials rule and  $m = 3$  for the replication Bayes factor. This is similar to the actually used number of sites  $m = 3$  (counting only external-replications), respectively,  $m = 4$  (counting also the internal-replication) from Protzko et al. (2020).

## 4 Discussion

We showed how Bayesian approaches can be used to determine the sample size of replication studies based on all the available information and the associated uncertainty. A key strength of the approach is that it can be applied to any kind of replication analysis method, Bayes or



**Figure 6:** Top plots show the probability of replication success based on the replication Bayes factor at level  $\gamma = 1/10$  (left) and the two-trials rule at level  $\alpha = 0.025$  (right) as a function of the total sample size and for different number of sites  $m$  for data from the experiment “Labels”. A design prior with heterogeneity  $\tau = 0.05$  and flat initial prior for the effect size  $\theta$  is used. The same heterogeneity value is assumed in the analysis of the replications. design (relative to the cost of a single sample  $K_c$ ) as a function of the number of sites  $m$  and for different site costs  $K_s$ . The sample size of each design corresponds to a target probability of replication success  $1 - \beta = 80\%$ .

non-Bayes, as long as there is a well-defined success region for the replication effect estimate. Methods for assessing replication success which have not yet been adapted to Bayesian design approaches in the normal-normal hierarchical model (or which have not even been proposed) can thus benefit from our methodology. For instance, our methods could be straightforwardly applied to the “dual-criterion” from [Rosenkranz \(2021\)](#) which defines replication success via simultaneous statistical significance and practical relevance of the effect estimates from original and replication study.

There are some limitations and possible extensions: we have developed the methodology for

---

“direct” replication studies (Simons, 2014) which attempt to replicate the conditions from the original study as closely as possible; yet SSD methodology is also needed for “conceptual” replication and for “generalization” studies which may show systematic deviations from the original study. While the heterogeneity variance in the design prior allows to take effect size heterogeneity into account for SSD, to some extent, further research is needed for investigating how systematic study deviations and external knowledge can be incorporated. Furthermore, as is standard in meta-analysis we assumed that the variances of the effect estimates are known, which can sometimes be inadequate (Jackson and White, 2018). Specifying priors also for the variances could better reflect the available uncertainty but would come at the price of lower interpretability and higher computational complexity. We did also not consider designs where the replication data are analyzed in a sequential manner. Ideas from the Bayesian sequential design (Schönbrodt and Wagenmakers, 2017; Stefan et al., 2022) or from the adaptive trials literature (Bretz et al., 2009) could be adapted to the replication setting as in Micheloud and Held (2022). A sequential analysis of the replication data could possibly increase the efficiency of the replication. An additional point is that we assumed that the original study has been completed when planning the replication study. One could also consider a scenario where both the original and replication study are planned simultaneously and adopt a “project” perspective (Maca et al., 2002; Held et al., 2022b). However, in this case no information from the original study is available and the design prior needs to be specified entirely based on external knowledge. Finally, researchers have only limited resources and it may happen that they cannot afford a large enough sample size to obtain their desired probability of replication success. In this situation a reverse-Bayes approach (Held et al., 2022a) could be applied in order to determine the prior for the effect size which is required to achieve the desired probability of replication success based on the maximally possible sample size. Researchers can then judge whether or not such prior beliefs are scientifically sensible, and decide whether they should conduct the replication study with their limited resources.

## Software and data

All our analyses were conducted in the R programming language version 4.2.2 (R Core Team, 2022). Code to reproduce this manuscript is available at <https://github.com/SamCH93/BAtDRS>. A snapshot of the Git repository at the time of writing is archived at <https://doi.org/10.5281/zenodo.7291076>. Methods for Bayesian design of replication studies are implemented in the R package BayesRepDesign which is available at <https://github.com/SamCH93/BayesRepDesign>. The CC-BY 4.0 licensed data were downloaded from <https://osf.io/42ef9/>. The R markdown script “Decline effects main analysis.Rmd” was executed and the relevant variables from the objects “ES\_experiments” and “decline\_effects” were saved.

## Acknowledgments

This work was supported by the Swiss National Science Foundation (#189295). The funder had no role in study design, data collection, data analysis, data interpretation, decision to publish,

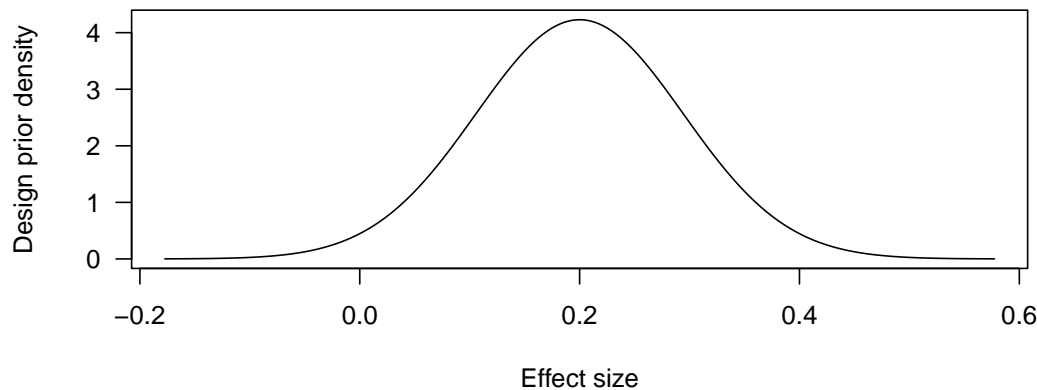
---

or preparation of the manuscript. We thank [Protzko et al. \(2020\)](#) for publicly sharing their data. We thank Charlotte Micheloud and Angelika Stefan for helpful comments on drafts of the manuscript. Our acknowledgment of these individuals does not imply their endorsement of this article.

## A The BayesRepDesign R package

```
library("BayesRepDesign")

## design prior (flat initial prior for effect size + heterogeneity)
dp <- designPrior(to = 0.2, so = 0.05, tau = 0.08)
plot(dp)
```



```
## compute replication standard error for achieving significance at 2.5%
ssdSig(level = 0.025, dprior = dp, power = 0.8)

##      Bayesian sample size calculation for replication studies
##      =====
##
## success criterion and computation
## -----
## replication p-value <= 0.025 (exact computation)
##
## original data and initial prior for effect size
## -----
## to = 0.2 : original effect estimate
## so = 0.05 : standard error of original effect estimate
## tau = 0.08 : assumed heterogeneity standard deviation
## N(mean = 0, sd = Inf) : initial normal prior
##
## design prior for effect size
## -----
```



---

```

##      N(mean = 0.2, sd = 0.094) : normal design prior
##
## probability of replication success
## -----
##      PoRS = 0.8 : specified
##      PoRS = 0.8 : recomputed with sr
##
## required sample size
## -----
##      sr = 0.045 : required standard error of replication effect estimate
##      c = so^2/sr^2 ~= nr/no = 1.2 : required relative variance / sample size

## compute numerically via success region and method agnostic function
sregFun <- function(sr) {
  ## success region is [1.96*sr, Inf)
  successRegion(intervals = cbind(qnorm(p = 0.975)*sr, Inf))
}
ssd(sregionfun = sregFun, dprior = dp, power = 0.8)

##      Bayesian sample size calculation for replication studies
##      =====
##
## success criterion and computation
## -----
##      method agnostic success region (numerical computation)
##
## original data and initial prior for effect size
## -----
##      to = 0.2 : original effect estimate
##      so = 0.05 : standard error of original effect estimate
##      tau = 0.08 : assumed heterogeneity standard deviation
##      N(mean = 0, sd = Inf) : initial normal prior
##
## design prior for effect size
## -----
##      N(mean = 0.2, sd = 0.094) : normal design prior
##
## probability of replication success
## -----
##      PoRS = 0.8 : specified
##      PoRS = 0.8 : recomputed with sr
##
## required sample size
## -----
##      sr = 0.045 : required standard error of replication effect estimate
##      c = so^2/sr^2 ~= nr/no = 1.2 : required relative variance / sample size

```

## B Supplementary material

Here, we provide additional information on computing the predictive distribution of the replication effect estimate when a prior is assigned to the heterogeneity variance  $\tau^2$  (Section B.1). We also provide additional information on methods for analyzing replication data. For each method we derive the *success region* in terms of the effect estimate of the replication study  $\hat{\theta}_r$ , which is required for sample size determination as illustrated in the main manuscript (Section B.2 to B.8). For the two-trials rule and the replication Bayes factor methods we additionally provide derivations on how these methods can be generalized to the multisite replication setting. We show then how the optimal number of samples per site can be derived for multisite SSD (Section B.9). Finally, we show SSD for the [Protzko et al. \(2020\)](#) project but using an adaptive shrinkage prior instead of the flat prior as in the main manuscript (Section B.10).

### B.1 Prior on the heterogeneity variance

When also a prior is assigned to the heterogeneity variance  $\tau^2$ , the predictive distribution of the replication effect estimate  $\hat{\theta}_r$  is given by

$$f(\hat{\theta}_r | \hat{\theta}_o, \sigma_o, \sigma_r) = \int_0^{+\infty} f(\hat{\theta}_r | \sigma_r, \hat{\theta}_o, \sigma_o, \tau^2) f(\tau^2 | \hat{\theta}_o, \sigma_o) d\tau^2.$$

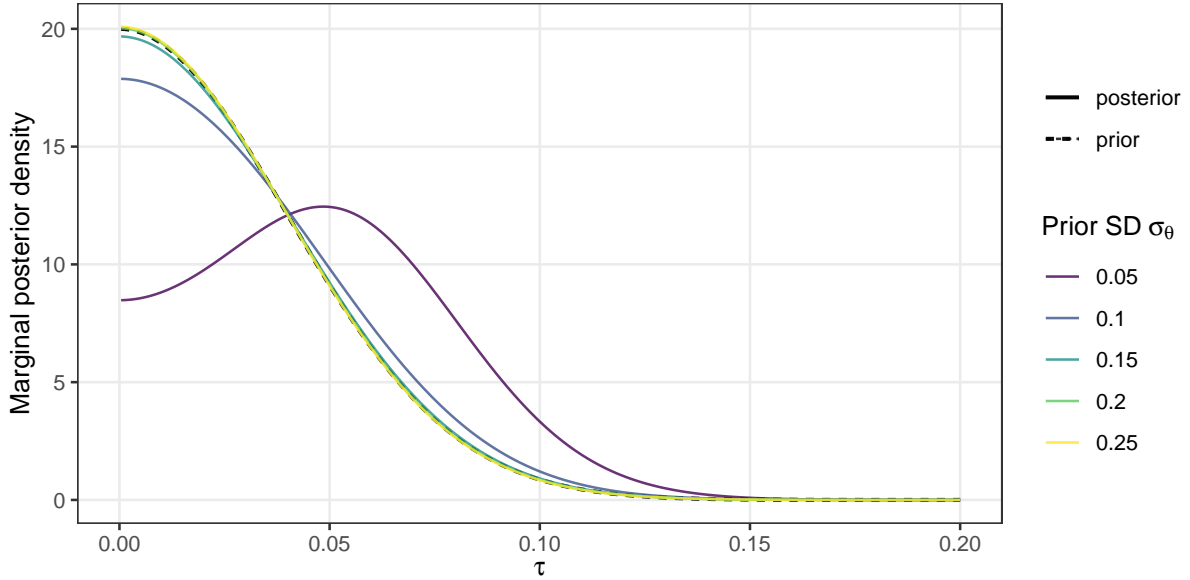
That is, it is the predictive distribution of the replication effect estimate  $\hat{\theta}_r$  integrated with respect to the marginal posterior of  $\tau^2$  based on the original data  $x_o = \{\hat{\theta}_o, \sigma_o^2\}$ . If the initial prior for  $\theta$  is normal  $\theta \sim N(\mu_\theta, \sigma_\theta^2)$ , and the initial prior for  $\tau^2$  has density  $f(\tau^2)$ , we have

$$\begin{aligned} f(\tau^2 | \hat{\theta}_o, \sigma_o) &= \int_{-\infty}^{+\infty} f(\theta, \tau^2 | \hat{\theta}_o, \sigma_o) d\theta \\ &= \frac{\int_{-\infty}^{+\infty} f(\hat{\theta}_o | \theta, \tau^2, \sigma_o^2) f(\theta | \tau^2) f(\tau^2) d\theta}{\int_0^{+\infty} \int_{-\infty}^{+\infty} f(\hat{\theta}_o | \theta_*, \tau_*^2, \sigma_o^2) f(\theta_* | \tau_*^2) f(\tau_*^2) d\theta_* d\tau_*^2} \\ &= \frac{f(\tau^2) \int_{-\infty}^{+\infty} N(\hat{\theta}_o | \theta, \tau^2 + \sigma_o^2) N(\theta | \mu_\theta, \sigma_\theta^2) d\theta}{\int_0^{+\infty} f(\tau_*^2) \int_{-\infty}^{+\infty} N(\hat{\theta}_o | \theta_*, \tau_*^2 + \sigma_o^2) N(\theta_* | \mu_\theta, \sigma_\theta^2) d\theta_* d\tau_*^2} \\ &= \frac{f(\tau^2) N(\hat{\theta}_o | \mu_\theta, \tau^2 + \sigma_o^2 + \sigma_\theta^2)}{\int_0^{+\infty} f(\tau_*^2) N(\hat{\theta}_o | \mu_\theta, \tau_*^2 + \sigma_o^2 + \sigma_\theta^2) d\tau_*^2}. \end{aligned}$$

To compute the marginal posterior density of  $\tau^2$  one numerical integration is hence required. The updating of the prior depends on the distance between prior mean  $\mu_\theta$  and the original effect estimate  $\hat{\theta}_o$  relative to the prior variance  $\sigma_\theta^2$  and the squared standard error  $\sigma_o^2$ . If an improper uniform prior is assigned to  $\theta$  ( $\sigma_\theta^2 \rightarrow \infty$ ), the posterior reduces to the prior

$$\begin{aligned} \lim_{\sigma_\theta^2 \rightarrow \infty} f(\tau^2 | \hat{\theta}_o, \sigma_o) &= \lim_{\sigma_\theta^2 \rightarrow \infty} \frac{f(\tau^2) N(\hat{\theta}_o | \mu_\theta, \tau^2 + \sigma_o^2 + \sigma_\theta^2)}{\int_0^{+\infty} f(\tau_*^2) N(\hat{\theta}_o | \mu_\theta, \tau_*^2 + \sigma_o^2 + \sigma_\theta^2) d\tau_*^2} \\ &= \lim_{\sigma_\theta^2 \rightarrow \infty} \int_0^{+\infty} \frac{f(\tau^2)}{f(\tau_*^2)} \underbrace{\sqrt{\frac{\tau_*^2 + \sigma_o^2 + \sigma_\theta^2}{\tau^2 + \sigma_o^2 + \sigma_\theta^2}}}_{\rightarrow 1} \exp \left[ -\frac{1}{2} \left\{ \underbrace{\frac{(\hat{\theta}_o - \mu_\theta)^2}{\tau^2 + \sigma_o^2 + \sigma_\theta^2}}_{\downarrow 0} - \underbrace{\frac{(\hat{\theta}_o - \mu_\theta)^2}{\tau_*^2 + \sigma_o^2 + \sigma_\theta^2}}_{\downarrow 0} \right\} \right] d\tau_*^2 \\ &= f(\tau^2), \end{aligned}$$

the limit can be interchanged with the integral because of the monotone convergence theorem. This means that with a uniform prior nothing can be learned about the variance  $\tau^2$  which intuitively makes sense as estimation of a variance requires at least two observations. The phenomenon is illustrated in Figure 7 for the data from the experiment “Labels” (Protzko et al., 2020) as also used in the main manuscript. We see that as the prior standard deviation increases (making the prior more uniform), the marginal posterior density becomes closer to the prior density.



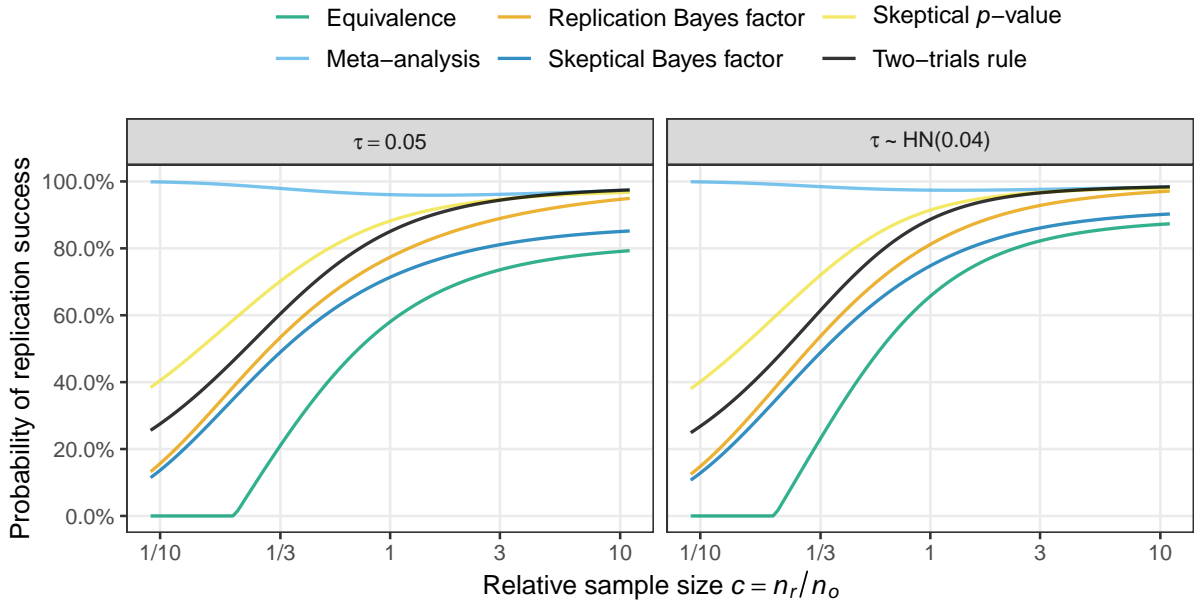
**Figure 7:** Marginal posterior distribution of heterogeneity variance  $\tau^2$  based on data from experiment “Labels” from Protzko et al. (2020) with original effect estimate  $\hat{\theta}_o = 0.205$  and standard error  $\sigma_o = 0.051$ . A  $\theta \sim N(0, \sigma_\theta^2)$  prior is assigned to the effect size  $\theta$  and a half normal prior with standard deviation 0.04 is assigned to  $\tau$ .

Combining all the previous results, we obtain the probability of replication success as

$$\begin{aligned} \Pr(\hat{\theta}_r \in S \mid \hat{\theta}_o, \sigma_o, \sigma_r) &= \int_S \int_0^{+\infty} f(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_o, \sigma_r, \tau^2) f(\tau^2 \mid \hat{\theta}_o, \sigma_o) d\hat{\theta}_r d\tau^2 \\ &= \int_0^{+\infty} \Pr(\hat{\theta}_r \in S \mid \hat{\theta}_o, \sigma_o, \sigma_r, \tau^2) f(\tau^2 \mid \hat{\theta}_o, \sigma_o) d\tau^2. \end{aligned}$$

This mean computing the probability of replication success with a prior on  $\tau^2$  requires two-dimensional numerical integration. However, in the common case when a uniform prior is assigned to  $\theta$ , the marginal posterior distribution of  $\tau^2$  reduces to the prior, and only one numerical integration is required.

Figure 8 shows the probability of replication success based on data from the “Labels” experiment, as in the main manuscript. A half normal prior with is assigned to the heterogeneity  $\tau$  which is a typical prior distribution used for heterogeneity modeling in meta-analysis (Röver



**Figure 8:** Probability of replication success as a function of relative sample size  $c = n_r/n_o$  for experiment “Labels” with original effect estimate  $\hat{\theta}_o = 0.205$  and standard error  $\sigma_o = 0.051$  for uniform initial prior for effect size  $\theta$  and either fixed  $\tau = 0.05$  (as in main manuscript) or half normal prior with standard deviation 0.04 assigned to  $\tau$ . Replication success is defined by the two-trials rule at level  $\alpha = 0.025$ , the replication Bayes factor at level  $\gamma = 1/10$ , fixed effects-meta analysis at level  $\alpha = 0.025^2$ , effect size equivalence based on 90% confidence interval and with margin  $\Delta = 0.2$ , sceptical  $p$ -value at level  $\alpha = 0.062$ , and sceptical Bayes factor at level  $\gamma = 1/10$ .

et al., 2021). The standard deviation of the prior is set to 0.04 so that the mean of the prior equals the value of the fixed heterogeneity  $\tau = 0.05$  elicited in the main manuscript. We see that the probability of replication success is only slightly higher compared to the fixed  $\tau = 0.05$  from the main manuscript.

## B.2 The two-trials rule

The two-trials rule is the most common analysis approach for replication studies. Replication success is declared if both original and replication study achieve statistical significance at some level  $\alpha$  (and both estimates go in the same direction which can be taken into account by using one-sided  $p$ -values). We will study the two-trial under normality using the data model  $\hat{\theta}_i | \theta \sim N(\theta, \sigma_i^2)$  with  $\hat{\theta}_i$  the estimate of the unknown effect size  $\theta$  from study  $i$  and  $\sigma_i$  is the corresponding standard error (assumed to be known). The  $p$ -values for testing  $H_0: \theta = 0$  versus  $H_1: \theta > 0$  are then  $p_i = 1 - \Phi(\hat{\theta}_i/\sigma_i)$  whereas for the alternative  $H_1: \theta < 0$  they are  $p_i = \Phi(\hat{\theta}_i/\sigma_i)$ . Suppose the original effect estimate was statistically significant at level  $\alpha$ , i. e.,  $p_o \leq \alpha$ . Replication success at level  $\alpha$  is then established if the replication effect estimate  $\hat{\theta}_r$  is also statistically significant at level  $\alpha$ , i. e.,  $p_r \leq \alpha$ . By applying some algebraic manipulations to the success condition, one can show that this implies that replication success is achieved if

the replication effect estimate  $\hat{\theta}_r$  is contained in the success region

$$S_{2TR} = \begin{cases} [z_\alpha \sigma_r, \infty) & \text{for } \hat{\theta}_o > 0 \\ [-\infty, -z_\alpha \sigma_r) & \text{for } \hat{\theta}_o < 0. \end{cases}$$

### The multisite two-trials rule

If multiple replication studies are conducted for one original study (a *multisite* replication), the two-trials rule is typically modified by meta-analyzing the effect estimates from all replications and then using the combined estimate as usual in the two-trials rule (see e.g., the “Many labs” projects from Klein et al., 2014, 2018). Suppose  $m$  replication studies are conducted and produce  $m$  effect estimates  $\hat{\theta}_{r1}, \dots, \hat{\theta}_{rm}$  with standard errors  $\sigma_{r1}, \dots, \sigma_{rm}$ . Subsequently, a weighted average  $\hat{\theta}_{r*} = \{\sum_{i=1}^m \hat{\theta}_{ri} / (\sigma_{ri}^2 + \tau_r^2)\} / \sigma_{r*}^2$  with standard error  $\sigma_{r*} = 1 / \sqrt{\{\sum_{i=1}^m 1 / (\sigma_{ri}^2 + \tau_r^2)\}}$  can be computed. If the between-replication heterogeneity variance  $\tau_r^2$  is set to zero this corresponds to the fixed effects estimate of  $\theta$ , while estimating  $\tau_r^2$  from the data corresponds to the random effects estimate. Replication success at level  $\alpha$  is then established if the replication  $p$ -value is smaller than  $\alpha$ , i.e.,  $p_{r*} = 1 - \Phi(\hat{\theta}_{r*} / \sigma_{r*}) \leq \alpha$ . With some algebra one can show that this implies a success region for the weighted average replication effect estimate  $\hat{\theta}_{r*}$  given by

$$S_{2TR} = \begin{cases} [z_\alpha \sigma_{r*}, \infty) & \text{for } \hat{\theta}_o > 0 \\ [-\infty, -z_\alpha \sigma_{r*}) & \text{for } \hat{\theta}_o < 0. \end{cases}$$

### B.3 Fixed effects meta-analysis

Assume again the data model  $\hat{\theta}_i | \theta \sim N(\theta, \sigma_i^2)$  where  $\hat{\theta}_i$  is an estimate of the effect size  $\theta$  from study  $i \in \{o, r\}$  and  $\sigma_i$  is the corresponding standard error (assumed to be known). In the fixed effects meta-analysis approach replicability is assessed in terms of the pooled effect estimate  $\hat{\theta}_m$  and standard error  $\sigma_m$  which are

$$\hat{\theta}_m = (\hat{\theta}_o / \sigma_o^2 + \hat{\theta}_r / \sigma_r^2) / \sigma_m^2 \quad \text{and} \quad \sigma_m = (1 / \sigma_o^2 + 1 / \sigma_r^2)^{-1/2},$$

which are also equivalent to the mean and standard deviation of a posterior distribution for the effect size  $\theta$  based on the data from original and replication study and an initial flat prior for  $\theta$ . Fixed effects meta-analysis is typically used because estimating a heterogeneity variance from two studies is highly unstable. Replication success at level  $\alpha$  is established if the one-sided meta-analytic  $p$ -value (in the direction of the original effect estimate  $\hat{\theta}$ ) is significant at level  $\alpha$ , i.e.,  $p_m = 1 - \Phi(\hat{\theta}_m / \sigma_m) \leq \alpha$  for  $\hat{\theta}_o > 0$  and  $p_m = \Phi(\hat{\theta}_m / \sigma_m) \leq \alpha$  for  $\hat{\theta}_o < 0$ . With some algebraic manipulations one can show that this criterion implies a success region  $S_{MA}$  for the replication effect estimate  $\hat{\theta}_r$  given by

$$S_{MA} = \begin{cases} [\sigma_r z_\alpha \sqrt{1 + \sigma_r^2 / \sigma_o^2} - (\hat{\theta}_o \sigma_r^2) / \sigma_o^2, \infty) & \text{for } \hat{\theta}_o > 0 \\ (-\infty, -\sigma_r z_\alpha \sqrt{1 + \sigma_r^2 / \sigma_o^2} - (\hat{\theta}_o \sigma_r^2) / \sigma_o^2] & \text{for } \hat{\theta}_o < 0. \end{cases}$$

## B.4 Effect size equivalence

The effect size equivalence approach (Anderson and Maxwell, 2016) defines replication success via compatibility of the effect estimates from both studies. Under normality we may assume the data model  $\hat{\theta}_i | \theta_i \sim N(\theta_i, \sigma_i^2)$  for study  $i \in \{o, r\}$ , and we are interested in the true effect size difference  $\delta = \theta_r - \theta_o$ . A  $(1 - \alpha)$  confidence interval for  $\delta$  is then given by

$$C_\alpha = \left[ \hat{\theta}_r - \hat{\theta}_o - z_{\alpha/2} \sqrt{\sigma_r^2 + \sigma_o^2}, \hat{\theta}_r - \hat{\theta}_o + z_{\alpha/2} \sqrt{\sigma_r^2 + \sigma_o^2} \right]$$

Effect size equivalence is established if the confidence interval is fully included in an equivalence region  $C_\alpha \subseteq [-\Delta, \Delta]$  with  $\Delta > 0$  a pre-specified margin. Applying some algebraic manipulations to the success conditions one can show that the equivalence test replication success criterion implies a success region  $S_E$  for the replication estimate  $\hat{\theta}_r$  given by

$$S_E = \left[ \hat{\theta}_o - \Delta + z_{\alpha/2} \sqrt{\sigma_o^2 + \sigma_r^2}, \hat{\theta}_o + \Delta - z_{\alpha/2} \sqrt{\sigma_o^2 + \sigma_r^2} \right].$$

## B.5 The replication Bayes factor

The replication Bayes factor approach uses the replication data  $x_r$  to quantify the evidence for the null hypothesis  $H_0: \theta = 0$  relative to the alternative hypothesis  $H_1: \theta \sim f(\theta | x_o)$ , which postulates that the effect size  $\theta$  is distributed according to its posterior distribution based on the original data  $x_o$ . Assume again a normal model  $\hat{\theta}_i | \theta \sim N(\theta, \sigma_i^2)$  with  $\hat{\theta}_i$  an estimate of the effect size  $\theta$  from study  $i \in \{o, r\}$  and  $\sigma_i$  the corresponding standard error (assumed to be known), and that we use the alternative  $H_1: N(\hat{\theta}_o, \sigma_o^2)$  which arises from updating an initial flat prior for  $\theta$  the original data  $x_o = \{\hat{\theta}_o, \sigma_o\}$ . The replication Bayes factor is then

$$\text{BF}_R = \frac{f(\hat{\theta}_r | H_0)}{f(\hat{\theta}_r | H_1)} = \sqrt{1 + \sigma_o^2 / \sigma_r^2} \exp \left[ -\frac{1}{2} \left\{ \frac{\hat{\theta}_r^2}{\sigma_r^2} - \frac{(\hat{\theta}_r - \hat{\theta}_o)^2}{\sigma_o^2 + \sigma_r^2} \right\} \right]. \quad (25)$$

Replication success at level  $\gamma \in (0, 1)$  is achieved if  $\text{BF}_R \leq \gamma$ . By applying some algebra to  $\text{BF}_R \leq \gamma$ , one can show that it is equivalent to the replication effect estimate  $\hat{\theta}_r$  falling in the success region

$$S_{\text{BF}_R} = \left( -\infty, -\sqrt{A} - (\hat{\theta}_o \sigma_r^2) / \sigma_o^2 \right] \cup \left[ \sqrt{A} - (\hat{\theta}_o \sigma_r^2) / \sigma_o^2, \infty \right)$$

where  $A = \sigma_r^2 (1 + \sigma_r^2 / \sigma_o^2) \{ \hat{\theta}_o^2 / \sigma_o^2 - 2 \log \gamma + \log(1 + \sigma_o^2 / \sigma_r^2) \}$ .

## B.6 The multisite replication Bayes factor

The generalization of the replication Bayes factor to the multisite setting is straightforward. The data are represented by vector of replication effect estimates  $\hat{\theta}_r = (\hat{\theta}_{r1}, \dots, \hat{\theta}_{rm})^\top$  with corresponding standard error vector  $\sigma_r = (\sigma_{r1}, \dots, \sigma_{rm})^\top$ , and we assume the data model  $\hat{\theta}_r | \theta \sim N_m\{\theta \mathbf{1}_m, \text{diag}(\sigma^2 + \tau_r^2 \mathbf{1}_m)\}$  where  $\mathbf{1}_m$  is a vector of  $m$  ones and  $\tau_r^2$  is a heterogeneity

variance for the replication effect sizes (not to be confused with the heterogeneity variance  $\tau^2$  used in the design prior).

As in the singlesite case, the replication Bayes factor quantifies the evidence that the data provide for the null hypothesis  $H_0: \theta = 0$  relative to the alternative hypothesis  $H_1: \theta \sim N(\hat{\theta}_o, \sigma_o^2)$ . The marginal density of the replication data under the null hypothesis is simply  $\hat{\theta}_r | H_0 \sim N_m\{0 \mathbf{1}_m, \text{diag}(\sigma^2 + \tau_r^2 \mathbf{1}_m)\}$ , whereas the marginal likelihood under the alternative  $H_1$  is obtained from integrating the likelihood with respect to the prior distribution of  $\theta$  under the alternative  $H_1$ . Let  $N(x; m, v)$  denote the normal density function mean  $m$  and variance  $v$  evaluated at  $x$ . Define also  $\hat{\theta}_{r*} = \{\sum_{i=1}^n \hat{\theta}_{ri} / (\sigma_{ri}^2 + \tau_r^2)\} \sigma_{r*}^2$  and  $\sigma_{r*}^2 = 1 / \{\sum_{i=1}^n 1 / (\sigma_{ri}^2 + \tau_r^2)\}$ , i.e., the weighted average of the replication effect estimates based on the heterogeneity  $\tau_r^2$  and its variance. The marginal density is then

$$\begin{aligned}
f(\hat{\theta}_r | H_1) &= \int f(\hat{\theta}_r | \theta) f(\theta | H_1) d\theta \\
&= \int \frac{\exp \left[ -\frac{1}{2} \left\{ \sum_{i=1}^n \frac{(\hat{\theta}_{ri} - \theta)^2}{\sigma_{ri}^2 + \tau_r^2} + \frac{(\theta - \hat{\theta}_o)^2}{\sigma_o^2} \right\} \right]}{\{2\pi\sigma_o^2 \prod_{i=1}^n 2\pi(\sigma_{ri}^2 + \tau_r^2)\}^{1/2}} d\theta \\
&= \int \frac{\exp \left[ -\frac{1}{2} \left\{ \sum_{i=1}^n \frac{(\hat{\theta}_{ri} - \hat{\theta}_{r*})^2}{\sigma_{ri}^2 + \tau_r^2} + \frac{(\hat{\theta}_{r*} - \theta)^2}{\sigma_{r*}^2} + \frac{(\theta - \hat{\theta}_o)^2}{\sigma_o^2} \right\} \right]}{\{2\pi\sigma_o^2 \prod_{i=1}^n 2\pi(\sigma_{ri}^2 + \tau_r^2)\}^{1/2}} d\theta \\
&= \frac{\exp \left[ -\frac{1}{2} \left\{ \sum_{i=1}^n \frac{(\hat{\theta}_{ri} - \hat{\theta}_{r*})^2}{\sigma_{ri}^2 + \tau_r^2} \right\} \right]}{\{2\pi\sigma_o^2 \prod_{i=1}^n 2\pi(\sigma_{ri}^2 + \tau_r^2)\}^{1/2}} \underbrace{\int \exp \left[ -\frac{1}{2} \left\{ \frac{(\hat{\theta}_{r*} - \theta)^2}{\sigma_{r*}^2} + \frac{(\theta - \hat{\theta}_o)^2}{\sigma_o^2} \right\} \right] d\theta}_{=N(\hat{\theta}_{r*}; m, \sigma_o^2 + \sigma_{r*}^2) 2\pi\sigma_o\sigma_{r*}} \\
&= \left\{ (1 + \sigma_o^2 / \sigma_{r*}^2) \prod_{i=1}^n 2\pi(\sigma_{ri}^2 + \tau_r^2) \right\}^{-1/2} \exp \left[ -\frac{1}{2} \left\{ \sum_{i=1}^n \frac{(\hat{\theta}_{ri} - \hat{\theta}_{r*})^2}{\sigma_{ri}^2 + \tau_r^2} + \frac{(\hat{\theta}_{r*} - \hat{\theta}_o)^2}{\sigma_{r*}^2 + \sigma_o^2} \right\} \right].
\end{aligned}$$

Dividing the marginal density of  $\hat{\theta}_r$  under  $H_0$  by the marginal density of  $\hat{\theta}_r$  under  $H_1$  leads to cancellation of several terms, and produces the replication Bayes factor

$$\text{BF}_{01}(\hat{\theta}_r) = \frac{f(\hat{\theta}_r | H_0)}{f(\hat{\theta}_r | H_1)} = \sqrt{1 + \sigma_o^2 / \sigma_{r*}^2} \exp \left[ -\frac{1}{2} \left\{ \frac{\hat{\theta}_{r*}^2}{\sigma_{r*}^2} - \frac{(\hat{\theta}_{r*} - \hat{\theta}_o)^2}{\sigma_{r*}^2 + \sigma_o^2} \right\} \right].$$

The multisite replication Bayes factor is therefore equivalent to the singlesite replication Bayes factor from (25) but using the weighted average  $\hat{\theta}_{r*}$  and its standard error  $\sigma_{r*}$  as the replication effect estimate  $\hat{\theta}_r$  and standard error  $\sigma_r$ .

## B.7 The sceptical $p$ -value

Held (2020) proposed a reverse-Bayes approach for assessing replicability. One assumes again the data model  $\hat{\theta}_i | \theta \sim N(\theta, \sigma_i^2)$  with  $i \in \{o, r\}$ , along with a zero-mean “sceptical” prior

$\theta \sim N(0, \sigma_s^2)$  for the effect size. In a first step, a level  $\alpha \geq p_o = 1 - \Phi(|\hat{\theta}_o|/\sigma_o)$  is fixed and the “sufficiently sceptical” prior variance  $\sigma_s^2$  is computed

$$\sigma_s^2 = \frac{\sigma_o^2}{(z_o^2/z_\alpha^2) - 1}$$

where  $z_o = \hat{\theta}_o/\sigma_o$ . The sufficiently sceptical prior variance  $\sigma_s^2$  has the property that it renders the resulting posterior of  $\theta$  no longer “credible” at level  $\alpha$ , that is, the posterior tail probability is fixed to  $\Pr(\theta \geq 0 | \hat{\theta}_o, \sigma_o, \sigma_s) = 1 - \alpha$  for positive estimates and  $\Pr(\theta \leq 0 | \hat{\theta}_o, \sigma_o, \sigma_s) = 1 - \alpha$  for negative estimates. In a second step, the conflict between the sceptical prior and the observed replication data is quantified, larger conflict indicating a higher degree of replication success. For doing so, a prior predictive tail probability

$$p_{\text{Box}} = \begin{cases} 1 - \Phi\{\hat{\theta}_r/(\sigma_r^2 + \sigma_s^2)\} & \text{if } \hat{\theta}_o > 0 \\ \Phi\{\hat{\theta}_r/(\sigma_r^2 + \sigma_s^2)\} & \text{if } \hat{\theta}_o < 0 \end{cases}$$

is computed and replication success at level  $\alpha$  is declared if  $p_{\text{Box}} \leq \alpha$ . The smallest level  $\alpha$  at which replication success is achieved is called the *sceptical p-value*  $p_s$  and replication success at level  $\alpha$  is equivalent with  $p_s \leq \alpha$  (see Held, 2020; Held et al., 2022b, for more details on  $p_s$ ). By applying some algebraic manipulations to the condition  $p_{\text{Box}} \leq \alpha$ , one can show that it is equivalent to the replication effect estimate  $\hat{\theta}_r$  falling in the success region

$$S_{p_s} = \begin{cases} [z_\alpha \sqrt{\{\sigma_r^2 + \frac{\sigma_o^2}{(z_o^2/z_\alpha^2) - 1}\}}, \infty) & \text{if } \hat{\theta}_o > 0 \\ (-\infty, -z_\alpha \sqrt{\{\sigma_r^2 + \frac{\sigma_o^2}{(z_o^2/z_\alpha^2) - 1}\}}] & \text{if } \hat{\theta}_o < 0. \end{cases}$$

## B.8 The sceptical Bayes factor

Pawel and Held (2022) modified the reverse-Bayes assessment of replication success from Held (2020) to use Bayes factors (Jeffreys, 1961; Kass and Raftery, 1995) instead of tail probabilities as measures of evidence and prior data conflict. The procedure assumes again the data model  $\hat{\theta}_i | \theta \sim N(\theta, \sigma_i^2)$  for study  $i \in \{o, r\}$ . In the first step the original data are used to contrast the evidence for the point null hypothesis  $H_0: \theta = 0$  relative to the “sceptical” alternative  $H_S: \theta \sim N(0, \sigma_s^2)$  with the Bayes factor

$$\text{BF}_{0S} = \frac{f(\hat{\theta}_o | H_0)}{f(\hat{\theta}_o | H_S)} = \sqrt{1 + \sigma_s^2/\sigma_o^2} \exp \left\{ -\frac{z_o^2}{2(1 + \sigma_o^2/\sigma_s^2)} \right\}.$$

where  $z_o = \hat{\theta}_o/\sigma_o$ . One then determines the sufficiently sceptical prior variance  $\sigma_s^2$  so that the Bayes factor is fixed to a level  $\gamma \in (0, 1)$  meaning that there is no longer evidence against the null hypothesis at level  $\gamma$ . The sufficiently sceptical prior variance can be computed by

$$\sigma_s^2 = \begin{cases} -\frac{\hat{\theta}_o^2}{q} - \sigma_o^2 & \text{if } -\frac{\hat{\theta}_o^2}{q} \geq \sigma_o^2 \\ \text{undefined} & \text{else} \end{cases} \quad (26)$$

$$\text{where } q = W_{-1} \left\{ -\frac{z_o^2}{\gamma^2} \exp(-z_o^2) \right\} \quad (27)$$



with  $W_{-1}(\cdot)$  the branch of the Lambert W function with  $W(y) \leq -1$  for  $y \in [-1/e, 0)$ .

In a second step the conflict between the sceptical prior and the replication data is quantified. To do so, the sceptic is contrasted to the “advocacy” alternative  $H_A: \theta \sim N(\hat{\theta}_o, \sigma_o^2)$  which represents the position of an advocate as the prior corresponds to the posterior distribution based on the original data  $\{\hat{\theta}_o, \sigma_o\}$  and a flat prior for the effect size  $\theta$ . This is done by computing the Bayes factor

$$\text{BF}_{SA} = \frac{f(\hat{\theta}_r | H_S)}{f(\hat{\theta}_r | H_A)} = \sqrt{\frac{\sigma_o^2 + \sigma_r^2}{\sigma_s^2 + \sigma_r^2}} \exp \left[ -\frac{1}{2} \left\{ \frac{\hat{\theta}_r^2}{\sigma_s^2 + \sigma_r^2} - \frac{(\hat{\theta}_r - \hat{\theta}_o)^2}{\sigma_o^2 + \sigma_r^2} \right\} \right]$$

and replication success at level  $\gamma$  is defined by  $\text{BF}_{SA} \leq \gamma$  as the data favor the advocate over the sceptic at a higher level than the sceptic’s initial objection to the null hypothesis. The smallest level  $\gamma$  at which replication success is achievable is then called *the sceptical Bayes factor*  $\text{BF}_S$ , and replication success at level  $\gamma$  is equivalent to  $\text{BF}_S \leq \gamma$  (see [Pawel and Held, 2022](#), for details on how to compute  $\text{BF}_S$ ). To derive the success region of the sceptical Bayes factor one can apply algebraic manipulations to  $\text{BF}_{SA} \leq \gamma$ , the condition for replication success at level  $\gamma$ , which leads to

$$S_{\text{BF}_S} = \begin{cases} (-\infty, -\sqrt{B} - M] \cup [\sqrt{B} - M, \infty) & \text{for } \sigma_s^2 < \sigma_o^2 \\ [\hat{\theta}_o - \{(\sigma_o^2 + \sigma_r^2) \log \gamma\} / \hat{\theta}_o, \infty) & \text{for } \sigma_s^2 = \sigma_o^2 \\ [-\sqrt{B} - M, \sqrt{B} - M] & \text{for } \sigma_s^2 > \sigma_o^2 \end{cases} \quad (28)$$

with

$$B = \left\{ \frac{\hat{\theta}_o^2}{\sigma_o^2 - \sigma_s^2} + 2 \log \left( \frac{\sigma_o^2 + \sigma_r^2}{\sigma_s^2 + \sigma_r^2} \right) - 2 \log \gamma \right\} \frac{(\sigma_s^2 + \sigma_r^2)(\sigma_o^2 + \sigma_r^2)}{\sigma_o^2 - \sigma_s^2}$$

$$M = \frac{\hat{\theta}_o(\sigma_s^2 + \sigma_r^2)}{\sigma_o^2 - \sigma_s^2}$$

and the sufficiently sceptical prior variance  $\sigma_s^2$  computed by (26).

## B.9 Optimal number of sites

The total cost of the design are  $K = m(K_c n_r + K_s)$  so that we can write the number of sites  $m$  for a given total cost as

$$m = K(K_c n_r + K_s)^{-1}. \quad (29)$$

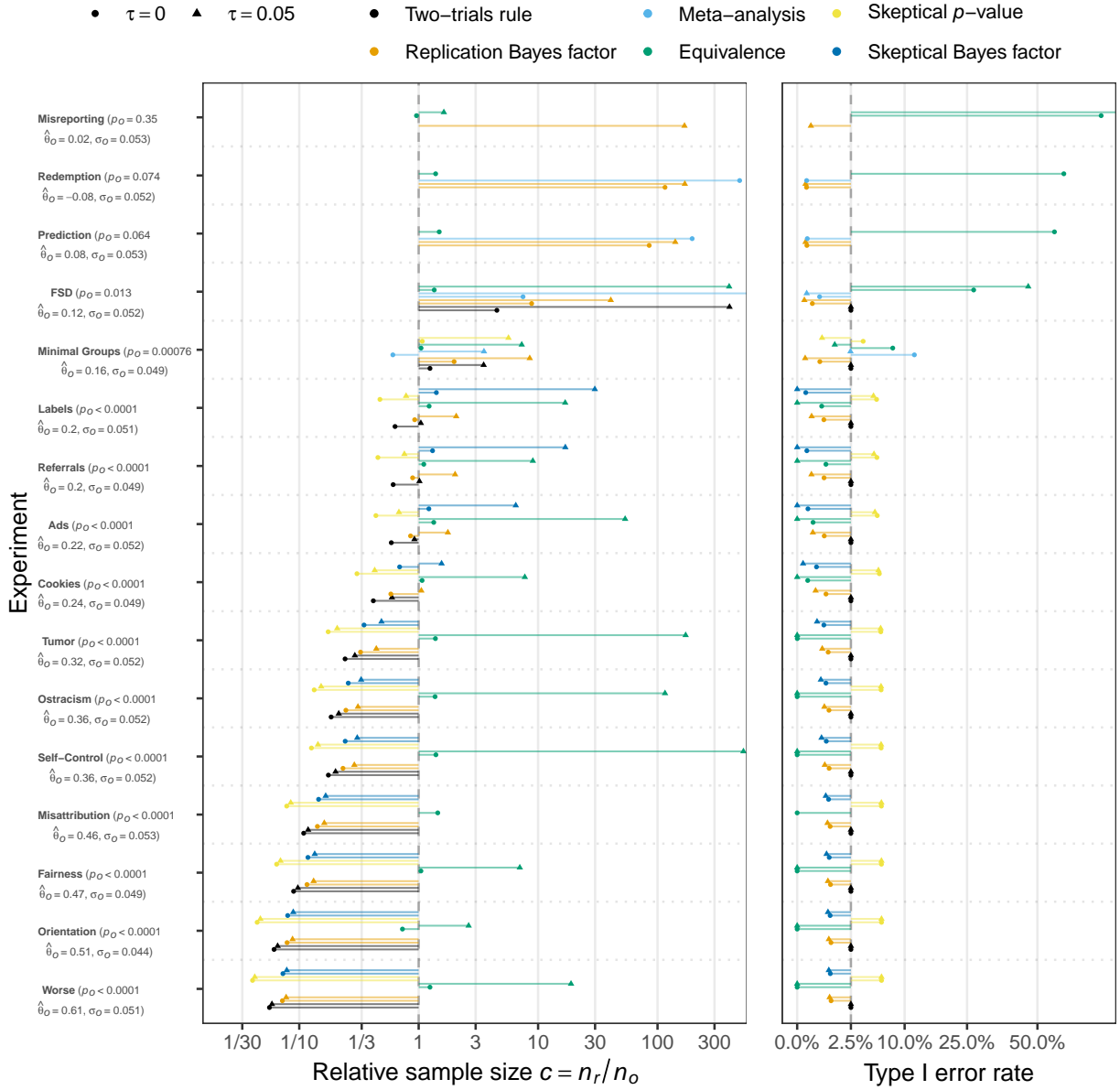
We now want to minimize the predictive variance of the weighted average  $\hat{\theta}_{r*}$  which, for a balanced design, is given by

$$\sigma_{\hat{\theta}_{r*}}^2 = \frac{\sigma_r^2 + \tau^2}{m} + \frac{\tau^2 + \sigma_o^2}{1 + 1/g}. \quad (30)$$

Plugging in (29) into (30) and minimizing it with respect to  $n_r$ , leads to the optimal sample size

$$n_r^* = \frac{\lambda}{\tau} \sqrt{\frac{K_s}{K_c}}$$

for a given cost ratio  $K_s/K_c$ .



**Figure 9:** The left plot shows the required relative sample size  $c = n_r/n_o$  to achieve a target probability of replication success of  $1 - \beta = 80\%$  (if possible). Replication success is defined through the two-trials rule at level  $\alpha = 0.025$ , replication Bayes factor at level  $\gamma = 1/10$ , fixed effects-meta analysis at level  $\alpha = 0.025^2$ , effect size equivalence at level  $\alpha = 0.1$  with margin  $\Delta = 0.2$ , sceptical  $p$ -value at level  $\alpha = 0.062$ , and sceptical Bayes factor at level  $\gamma = 1/10$  for data from the replication project by Protzko et al. (2020). An adaptive shrinkage prior is used for the effect size  $\theta$  either without ( $\tau = 0$ ) or with between-study heterogeneity ( $\tau = 0.05$ ). The right plot shows the type I error rate associated with the required sample size. Experiments are ordered (top to bottom) by their original one-sided  $p$ -value  $p_o = 1 - \Phi(|\hat{\theta}_o|/\sigma_o)$ .

---

## B.10 Sample size determination with adaptive shrinkage prior

Figure 9 shows sample size determination for all studies from the Protzko et al. (2020) project as in the main manuscript but using an “adaptive shrinkage prior” for the effect size  $\theta$  where the variance of the shrinkage prior is estimated by empirical Bayes. We see that the required sample size increases for studies with large  $p$ -values compared to the analysis based on a flat prior for  $\theta$  as in the main manuscript, whereas it stays about the same for studies with small  $p$ -values. This is because studies with large  $p$ -values receive more shrinkage.

## Bibliography

- Anderson, S. F. and Kelley, K. (2022). Sample size planning for replication studies: The devil is in the design. *Psychological Methods*. doi:[10.1037/met0000520](https://doi.org/10.1037/met0000520).
- 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](https://doi.org/10.1037/met0000051).
- Anderson, S. F. and Maxwell, S. E. (2017). Addressing the “replication crisis”: Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research*, 52(3):305–324. doi:[10.1080/00273171.2017.1289361](https://doi.org/10.1080/00273171.2017.1289361).
- Bayarri, M. J. and Mayoral, A. M. (2002). Bayesian design of “successful” replications. *The American Statistician*, 56:207–214. doi:[10.1198/000313002155](https://doi.org/10.1198/000313002155).
- Bonett, D. G. (2020). Design and analysis of replication studies. *Organizational Research Methods*, 24(3):513–529. doi:[10.1177/1094428120911088](https://doi.org/10.1177/1094428120911088).
- Bretz, F., Koenig, F., Brannath, W., Glimm, E., and Posch, M. (2009). Adaptive designs for confirmatory clinical trials. *Statistics in Medicine*, 28(8):1181–1217. doi:[10.1002/sim.3538](https://doi.org/10.1002/sim.3538).
- 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](https://doi.org/10.1038/s41562-018-0399-z).
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1):155–159. doi:[10.1037/0033-2909.112.1.155](https://doi.org/10.1037/0033-2909.112.1.155).
- Copas, J. B. (1983). Regression, prediction and shrinkage (with discussion). *Journal of the Royal Statistical Society*, 45:311–354. doi:[10.1111/j.2517-6161.1983.tb01258.x](https://doi.org/10.1111/j.2517-6161.1983.tb01258.x).
- De Santis, F. (2004). Statistical evidence and sample size determination for Bayesian hypothesis testing. *Journal of Statistical Planning and Inference*, 124(1):121–144. doi:[10.1016/s0378-3758\(03\)00198-8](https://doi.org/10.1016/s0378-3758(03)00198-8).
- Deeks, J. J., Higgins, J. P., and Altman, D. G. (2019). Analysing data and undertaking meta-analyses. In *Cochrane Handbook for Systematic Reviews of Interventions*, chapter 10, pages 241–284. John Wiley & Sons, Ltd.

- 
- 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.
- Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38. doi:10.1093/biomet/80.1.27.
- Gelfand, A. E. and Wang, F. (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*, 17(2). doi:10.1214/ss/1030550861.
- Gelman, A. (2009). Bayes, Jeffreys, prior distributions and the philosophy of statistics. *Statistical Science*, 24(2). doi:10.1214/09-sts284d.
- Goodman, S. N. (1992). A comment on replication, *p*-values and evidence. *Statistics in Medicine*, 11(7):875–879. doi:10.1002/sim.4780110705.
- 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.
- Grieve, A. P. (2022). *Hybrid frequentist/Bayesian power and Bayesian power in planning clinical trials*. Chapman & Hall/CRC Biostatistics Series. Taylor & Francis, London, England.
- 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. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2):107–128. doi:10.3102/10769986006002107.
- 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.
- Hedges, L. V. and Schauer, J. M. (2021). The design of replication studies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(3):868–886. doi:https://doi.org/10.1111/rssa.12688.
- 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*, 13(3):295–314. doi:10.1002/jrsm.1538.
- Held, L., Micheloud, C., and Pawel, S. (2022b). The assessment of replication success based on relative effect size. *The Annals of Applied Statistics*, 16(2):706–720. doi:10.1214/21-aos1502.
- Held, L. and Pawel, S. (2020). Comment on “the role of *p*-values in judging the strength of evidence and realistic replication expectations”. *Statistics in Biopharmaceutical Research*, 13(1):46–48. doi:10.1080/19466315.2020.1828161.

- 
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8):e124. doi:[10.1371/journal.pmed.0020124](https://doi.org/10.1371/journal.pmed.0020124).
- Jackson, D. and White, I. R. (2018). When should meta-analysis avoid making hidden normality assumptions? *Biometrical Journal*, 60(6):1040–1058. doi:[10.1002/bimj.201800071](https://doi.org/10.1002/bimj.201800071).
- Jeffreys, H. (1961). *Theory of Probability*. Oxford: Clarendon Press, third edition.
- 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](https://doi.org/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](https://doi.org/10.1080/01621459.1995.10476572).
- 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](https://doi.org/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](https://doi.org/10.1177/2515245918810225).
- Kunzmann, K., Grayling, M. J., Lee, K. M., Robertson, D. S., Rufibach, K., and Wason, J. M. S. (2021). A review of Bayesian perspectives on sample size derivation for confirmatory trials. *The American Statistician*, 75(4):424–432. doi:[10.1080/00031305.2021.1901782](https://doi.org/10.1080/00031305.2021.1901782).
- 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](https://doi.org/10.3758/s13428-018-1092-x).
- Maca, J., Gallo, P., Branson, M., and Maurer, W. (2002). Reconsidering some aspects of the two-trials paradigm. *Journal of Biopharmaceutical Statistics*, 12(2):107–119. doi:[10.1081/bip-120006450](https://doi.org/10.1081/bip-120006450).
- 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](https://doi.org/10.1111/rssa.12572).
- Matthews, J. N. (2006). *Introduction to Randomized Controlled Clinical Trials*. Chapman and Hall/CRC, New York. doi:[10.1201/9781420011302](https://doi.org/10.1201/9781420011302).
- McKinney, K., Stefan, A., and Gronau, Q. F. (2021). Developing prior distributions for Bayesian meta-analyses. doi:[10.31234/osf.io/2v5bz](https://doi.org/10.31234/osf.io/2v5bz).
- Micheloud, C. and Held, L. (2022). Power calculations for replication studies. *Statistical Science*, 37(3):369–379. doi:[10.1214/21-sts828](https://doi.org/10.1214/21-sts828).
- 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](https://doi.org/10.1038/s41562-016-0021).
-

- 
- O'Hagan, A. (2019). Expert knowledge elicitation: Subjective but scientific. *The American Statistician*, 73(sup1):69–81. doi:[10.1080/00031305.2018.1518265](https://doi.org/10.1080/00031305.2018.1518265).
- O'Hagan, A. and Stevens, J. (2001). Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Medical Decision Making*, 21(3):219–230. doi:[10.1177/02729890122062514](https://doi.org/10.1177/02729890122062514).
- O'Hagan, A., Stevens, J. W., and Campbell, M. J. (2005). Assurance in clinical trial design. *Pharmaceutical Statistics*, 4(3):187–201. doi:[10.1002/pst.175](https://doi.org/10.1002/pst.175).
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716. doi:[10.1126/science.aac4716](https://doi.org/10.1126/science.aac4716).
- Park, J. and Pek, J. (2022). Conducting Bayesian-classical hybrid power analysis with R package hybridpower. *Multivariate Behavioral Research*, pages 1–17. doi:[10.1080/00273171.2022.2038056](https://doi.org/10.1080/00273171.2022.2038056).
- 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](https://doi.org/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](https://doi.org/10.1371/journal.pone.0231416).
- Pawel, S. and Held, L. (2022). The sceptical Bayes factor for the assessment of replication success. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(3):879–911. doi:[10.1111/rssb.12491](https://doi.org/10.1111/rssb.12491).
- Pek, J. and Park, J. (2019). Complexities in power analysis: Quantifying uncertainties with a Bayesian-classical hybrid approach. *Psychological Methods*, 24(5):590–605. doi:[10.1037/met0000208](https://doi.org/10.1037/met0000208).
- 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., Waliczek, J., and Schooler, J. (2020). High replicability of newly-discovered social-behavioral findings is achievable. doi:[10.31234/osf.io/n2a9x](https://doi.org/10.31234/osf.io/n2a9x).
- 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/>.
- Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2(2):173–185. doi:[10.1037/1082-989x.2.2.173](https://doi.org/10.1037/1082-989x.2.2.173).
- Raudenbush, S. W. and Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, 5(2):199–213. doi:[10.1037/1082-989x.5.2.199](https://doi.org/10.1037/1082-989x.5.2.199).
- Rosenkranz, G. K. (2021). Replicability of studies following a dual-criterion design. *Statistics in Medicine*, 40(18):4068–4076. doi:[10.1002/sim.9014](https://doi.org/10.1002/sim.9014).
- Röver, C., Bender, R., Dias, S., Schmid, C. H., Schmidli, H., Sturtz, S., Weber, S., and Friede, T. (2021). On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis. *Research Synthesis Methods*, 12(4):448–474. doi:[10.1002/jrsm.1475](https://doi.org/10.1002/jrsm.1475).



- 
- Schönbrodt, F. D. and Wagenmakers, E.-J. (2017). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, 25(1):128–142. doi:[10.3758/s13423-017-1230-y](https://doi.org/10.3758/s13423-017-1230-y).
- Senn, S. (2002). Letter to the editor: A comment on replication,  $p$ -values and evidence by S. N. Goodman, *Statistics in Medicine* 1992; 11:875–879. *Statistics in Medicine*, 21(16):2437–2444. doi:[10.1002/sim.1072](https://doi.org/10.1002/sim.1072).
- Senn, S. (2008). *Statistical issues in drug development*, volume 69. John Wiley & Sons.
- Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, 9(1):76–80. doi:[10.1177/1745691613514755](https://doi.org/10.1177/1745691613514755).
- Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, 26:559–569. doi:[10.1177/0956797614567341](https://doi.org/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. and Freedman, L. S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine*, 5(1):1–13. doi:[10.1002/sim.4780050103](https://doi.org/10.1002/sim.4780050103).
- 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](https://doi.org/10.1016/0197-2456(86)90003-6).
- Stefan, A., Gronau, Q. F., and Wagenmakers, E.-J. (2022). Interim design analysis using Bayes factor forecasts. doi:[10.31234/osf.io/9sazk](https://doi.org/10.31234/osf.io/9sazk).
- Sutton, A. J. and Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, 10(4):277–303. doi:[10.1177/096228020101000404](https://doi.org/10.1177/096228020101000404).
- 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](https://doi.org/10.1371/journal.pone.0175302).
- van Zwet, E. W. and Goodman, S. N. (2022). How large should the next study be? predictive power and sample size requirements for replication studies. *Statistics in Medicine*, 41(16):3090–3101. doi:[10.1002/sim.9406](https://doi.org/10.1002/sim.9406).
- 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](https://doi.org/10.1037/a0036731).
- Weiss, R. (1997). Bayesian sample size calculations for hypothesis testing. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(2):185–191. doi:[10.1111/1467-9884.00075](https://doi.org/10.1111/1467-9884.00075).
- Zwet, E., Schwab, S., and Senn, S. (2021). The statistical properties of RCTs and a proposal for shrinkage. *Statistics in Medicine*, 40(27):6107–6117. doi:[10.1002/sim.9173](https://doi.org/10.1002/sim.9173).





## 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>.

---



---

## Abstract

It is now widely accepted that the standard inferential toolkit used by the scientific research community – null-hypothesis significance testing (NHST) – is not fit for purpose. Yet despite the threat posed to the scientific enterprise, there is no agreement concerning alternative approaches for evidence assessment. This lack of consensus reflects long-standing issues concerning Bayesian methods, the principal alternative to NHST. We report on recent work that builds on an approach to inference put forward over 70 years ago to address the well-known “Problem of Priors” in Bayesian analysis, by reversing the conventional prior-likelihood-posterior (“forward”) use of Bayes’s Theorem. Such Reverse-Bayes analysis allows priors to be deduced from the likelihood by requiring that the posterior achieve a specified level of credibility. We summarise the technical underpinning of this approach, and show how it opens up new approaches to common inferential challenges, such as assessing the credibility of scientific findings, setting them in appropriate context, estimating the probability of successful replications, and extracting more insight from NHST while reducing the risk of misinterpretation. We argue that Reverse-Bayes methods have a key role to play in making Bayesian methods more accessible and attractive for evidence assessment and research synthesis. As a running example we consider a recently published meta-analysis from several randomized controlled trials (RCTs) investigating the association between corticosteroids and mortality in hospitalized patients with COVID-19.

**Key words:** Analysis of Credibility, Bayes factor, false positive risk, meta-analysis, prior-data conflict, reverse-Bayes

## 1 Introduction: the origin of Reverse-Bayes methods

“We can make judgments of initial probabilities and infer final ones, or we can equally make judgments of final ones and infer initial ones by *Bayes’s theorem in reverse*.”

Good (1983, p. 29)

There is now a common consensus that the most widely-used methods of statistical inference have led to a crisis in both the interpretation of research findings and their replication (Gelman and Loken, 2014; Wasserstein and Lazar, 2016). At the same time, there is a lack of consensus on how to address the challenge (Matthews, 2017), as highlighted by the plethora of alternative techniques to null-hypothesis significance testing now being put forward, see for example Wasserstein et al. (2019) and the references therein. Especially striking is the relative dearth of alternatives based on Bayesian concepts. Given their intuitive inferential basis and output (Wagenmakers et al., 2008; McElreath, 2018), these would seem obvious candidates to supplant the prevailing frequentist methodology. However, it is well-known that the adoption of Bayesian methods continues to be hampered by several factors, such as the belief that advanced computational tools are required to make Bayesian statistics practical (Green et al., 2015). The most persistent of these is that the full benefit of Bayesian methods demands

---

specification of a prior level of belief, even in the absence of any appropriate insight. This “Problem of Priors” has cast a shadow over Bayesian methods since their emergence over 250 years ago (McGrayne, 2011), and has led to a variety of approaches, such as prior elicitation, prior sensitivity analysis, and objective Bayesian methodology; all have their supporters and critics.

One of the least well-known was suggested over 70 years ago (Good, 1950) by one of the best-known proponents of Bayesian methods during the 20<sup>th</sup> century, I.J. Good. It involves reversing the conventional direction of Bayes’s Theorem and determining the level of prior belief required to reach a specified level of posterior belief, given the evidence observed. This reversal of Bayes’s Theorem allows the assessment of new findings on the basis of whether the resulting prior is reasonable in the light of existing knowledge. Whether a prior is plausible in the light of existing knowledge can be assessed informally or more formally using techniques for comparing priors with existing data as suggested by Box (1980) and further refined by Evans and Moshonov (2006), see also Nott et al. (2020, 2021) for related approaches. Good stressed that despite the routine use of the adjectives “prior” and “posterior” in applications of Bayes’s Theorem, the validity of any resulting inference does not require a specific temporal ordering, as the theorem is simply a constraint ensuring consistency with the axioms of probability. While reversing Bayes’s Theorem is still regarded as unacceptable by some on the grounds it allows “cheating” in the sense of choosing priors to achieve a desired posterior inference (O’Hagan and Forster, 2004, p. 143), others point out this is not an ineluctable consequence of the reversal (Cox, 2006, pp. 78–79). As we shall show, recent technical advances further weaken this criticism.

Good’s belief in the value of Reverse-Bayes methods won support from E.T. Jaynes in his well-known treatise on probability. Explaining a specific manifestation of the approach (to be discussed shortly) Jaynes remarked: “We shall find it helpful in many cases where our prior information seems at first too vague to lead to any definite prior probabilities; it stimulates our thinking and tells us how to assign them after all” (Jaynes, 2003, p. 126). Yet despite the advocacy of two leading figures in the foundations of Bayesian methodology, the potential of Reverse-Bayes methods has remained largely unexplored. Most published work has focused on their use in putting new research claims in context, with Reverse-Bayes methods being used to assess whether the prior evidence needed to make a claim credible is consistent with existing insight (Carlin and Louis, 1996; Matthews, 2001a,b; Spiegelhalter et al., 2004; Greenland, 2006, 2011; Held, 2013; Colquhoun, 2017, 2019; Held, 2019a, 2020; Pawel and Held, 2022; Best et al., 2021).

The purpose of this paper is to highlight recent technical developments of Good’s basic idea which lead to inferential tools of practical value in the analysis of summary measures as reported in meta-analysis. As a running example we consider a recently published meta-analysis investigating the association between corticosteroids and mortality in hospitalized patients with COVID-19. Specifically, we show how Reverse-Bayes methods address the current concerns about the interpretation of new findings and their replication. We begin by illustrating the basics of the Reverse-Bayes approach for both hypothesis testing and parameter estimation. This is followed by a discussion of Reverse-Bayes methods for assessing effect estimates in Section 2. These allow the credibility of both new and existing research findings reported in

terms of NHST to be evaluated in the context of existing knowledge. This enables researchers to go beyond the standard dichotomy of statistical significance/non-significance, extracting further insight from their findings. We then discuss the use of the Reverse-Bayes approach in the most recalcitrant form of the Problem of Priors, involving the assessment of research findings which are unprecedented and thus lacking any clear source of prior support. We show how the concept of intrinsic credibility resolves this challenge, and puts recent calls to tighten  $p$ -value thresholds on a principled basis (Benjamin et al., 2017). In Section 3 we describe Reverse-Bayes methods with Bayes factors, the principled solution for Bayesian hypothesis testing. Finally, we describe in Section 4 Reverse-Bayes approaches to interpretational issues that arise in conventional statistical analysis based on  $p$ -values, and how they can be used to flag the risk of inferential fallacies. We close with some extensions and final conclusions.

## 1.1 Reverse-Bayes for hypothesis testing

The subjectivity involved in the specification of prior distributions is often seen as a weak point of Bayesian inference. The Reverse-Bayes approach can help to resolve this issue both in hypothesis testing and parameter estimation, we will start with the former.

Consider a null hypothesis  $H_0$  with prior probability  $\pi = \Pr(H_0)$ , so  $\Pr(H_1) = 1 - \pi$  is the prior probability of the alternative hypothesis  $H_1$ . Computation of the posterior probability of  $H_1$  is routine with Bayes' theorem:

$$\Pr(H_1 | \text{data}) = \frac{\Pr(\text{data} | H_1) \Pr(H_1)}{\Pr(\text{data} | H_0) \Pr(H_0) + \Pr(\text{data} | H_1) \Pr(H_1)}.$$

Bayes' theorem can be written in more compact form as

$$\frac{\Pr(H_1 | \text{data})}{\Pr(H_0 | \text{data})} = \frac{\Pr(\text{data} | H_1)}{\Pr(\text{data} | H_0)} \frac{\Pr(H_1)}{\Pr(H_0)}, \quad (1)$$

i. e., the posterior odds are the likelihood ratio times the prior odds. The standard 'forward-Bayes' approach thus fixes the prior odds (or one of the underlying probabilities), determines the likelihood ratio for the available data, and takes the product to compute the posterior odds. Of course, the latter can be easily back-transformed to the posterior probability  $\Pr(H_1 | \text{data})$ , if required. The Problem of Priors is now apparent: in order for us to update the odds in favour of  $H_1$ , we must first specify the prior odds. This can be problematic in situations where, for example, the evidence on which to base the prior odds is controversial or even non-existent.

However, as Good emphasised it is entirely justifiable to "flip" Bayes's theorem around, allowing us to ask the question: Which prior, when combined with the data, leads to our specified posterior?

$$\frac{\Pr(H_1)}{\Pr(H_0)} = \frac{\Pr(H_1 | \text{data})}{\Pr(H_0 | \text{data})} \bigg/ \frac{\Pr(\text{data} | H_1)}{\Pr(\text{data} | H_0)}. \quad (2)$$

For illustration we re-visit an example put forward by Good (1950, p. 35), perhaps the first published Reverse-Bayes calculation. It centres on a question for which the setting of an

initial prior is especially problematic: does an experiment provide convincing evidence for the existence of extra-sensory perception (ESP)? The substantive hypothesis  $H_1$  is that ESP exists, so that  $H_0$  asserts it does not exist. Imagine an experiment in which a person has to make  $n$  consecutive guesses of random digits (between 0 and 9) and all are correct, as the ESP hypothesis  $H_1$  would predict. The likelihood ratio is therefore

$$\frac{\Pr(\text{data} | H_1)}{\Pr(\text{data} | H_0)} = \frac{1}{(1/10)^n} = 10^n. \quad (3)$$

It is unlikely that sceptics and advocates of the existence of ESP would ever agree on what constitutes reasonable priors from which to start a standard Bayesian analysis of the evidence. However, Good argued that Reverse-Bayes offers a way forward by using it to set bounds on the prior probabilities for  $H_1$  and  $H_0$ . This is achieved via the outcome of a thought (Gedanken) experiment capable of demonstrating  $H_1$  is more likely than  $H_0$ , that is, of leading to posterior probabilities such that  $\Pr(H_1 | \text{data}) > \Pr(H_0 | \text{data})$ . Using this approach, which Good termed the *Device of Imaginary Results*, we see that if the ESP experiment produced  $n = 20$  correct consecutive guesses, (2) combined with (3) implies that ESP may be deemed more likely than not to exist by anyone whose priors satisfy  $\Pr(H_1)/\Pr(H_0) > 10^{-20}$ . In contrast, if only  $n = 3$  correct guesses emerged, then the existence of ESP could be rejected by anyone whose priors satisfy  $\Pr(H_1)/\Pr(H_0) < 10^{-3}$ . Using Bayes's Theorem in reverse has thus led to a quantitative statement of the prior beliefs that either advocates or sceptics of ESP must be able to justify in the face of results from a real experiment. The practical value of Good's approach was noted by Jaynes in his treatise: "[I]n the present state of development of probability theory, the device of imaginary results is usable and useful in a very wide variety of situations, where we might not at first think it applicable" (Jaynes, 2003, p. 125–126).

It is straightforward to extend (1) and (2) to hypotheses that involve unknown parameters  $\theta$ . The likelihood ratio  $\Pr(\text{data} | H_1)/\Pr(\text{data} | H_0)$  is then called a Bayes factor (Jeffreys, 1961; Kass and Raftery, 1995) where

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

is the marginal likelihood under hypothesis  $H_i$ ,  $i = 0, 1$ , obtained by integration of the ordinary likelihood with respect to the prior distribution  $f(\theta | H_i)$ . We will apply the Reverse-Bayes approach to Bayes factors in Section 3 and 4.

## 1.2 Reverse-Bayes for parameter estimation

We can also apply the Reverse-Bayes idea to continuous prior and posterior distributions of a parameter of interest  $\theta$ . Reversing Bayes' theorem

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

then leads to

$$f(\theta) = f(\text{data}) \frac{f(\theta | \text{data})}{f(\text{data} | \theta)}. \quad (4)$$

---

So the prior is proportional to the posterior divided by the likelihood with proportionality constant  $f(\text{data})$ .

Consider Bayesian inference for the mean  $\theta$  of a univariate normal distribution, assuming the variance  $\sigma^2$  is known. Let  $x$  denote the observed value from that  $N(\theta, \sigma^2)$  distribution and suppose the prior for  $\theta$  (and hence also the posterior) is conjugate normal. Each of them is determined by two parameters, usually the mean and the variance, but two distinct quantiles would also work. If we fix both parameters of the posterior, then the prior in (4) is – under a certain regularity condition – uniquely determined. For ease of presentation we work with the observational precision  $\kappa = 1/\sigma^2$  and denote the prior and posterior precision by  $\delta$  and  $\delta'$ , respectively. Finally let  $\mu$  and  $\mu'$  denote the prior and posterior mean, respectively.

Forward-Bayesian updating tells us how to compute the posterior precision and mean:

$$\delta' = \delta + \kappa, \quad (5)$$

$$\mu' = \frac{\mu\delta + x\kappa}{\delta'}. \quad (6)$$

For example, fixed-effect (FE) meta-analysis is based on iteratively applying (5) and (6) to the summary effect estimate  $x_i$  with standard error  $\sigma_i$  from the  $i$ -th study,  $i = 1, \dots, n$ , starting with an initial precision of zero. Reverse-Bayes simply inverts these equations, which leads to the following:

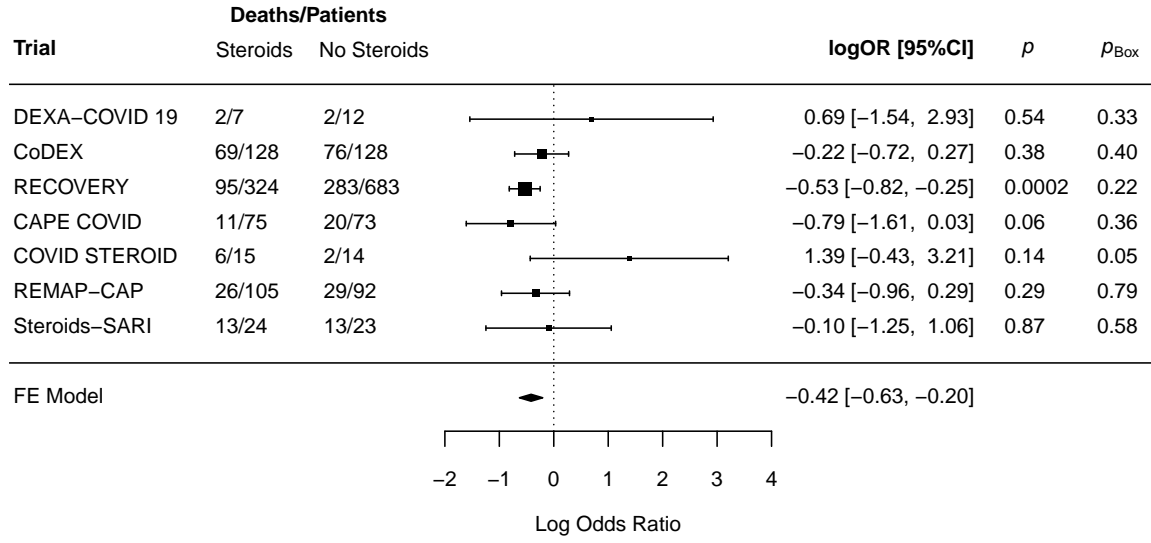
$$\delta = \delta' - \kappa, \quad (7)$$

$$\mu = \frac{\mu'\delta' - x\kappa}{\delta}, \quad (8)$$

provided  $\delta' > \kappa$ , i. e., the posterior precision must be larger than the observational precision.

We will illustrate the application of (7) and (8) as well as the methodology in the rest of this paper using a recent meta-analysis combining information from  $n = 7$  randomized controlled trials investigating the association between corticosteroids and mortality in hospitalized patients with COVID-19 ([WHO REACT Working Group, 2020](#)); its results are reproduced in Figure 1 (here and henceforth, odds ratios (ORs) are expressed as log odds ratios to transform the range from  $(0, \infty)$  to  $(-\infty, +\infty)$ , consistent with the assumption of normality). Let  $x_i = \hat{\theta}_i$  denote the maximum likelihood estimate (MLE) of the log odds ratio  $\theta$  in the  $i$ -th study with standard error  $\sigma_i$ . The meta-analytic odds ratio estimate under the fixed-effect model (the pre-specified primary analysis) is  $\widehat{\text{OR}} = 0.66$  [95% CI 0.53 to 0.82], respectively  $\hat{\theta} = -0.42$  [95% CI  $-0.63$  to  $-0.20$ ] for the log odds ratio  $\theta$ , indicating evidence for lower mortality of patients treated with corticosteroids compared to patients receiving usual care or placebo. The pooled effect estimate  $\hat{\theta}$  represents a posterior mean  $\mu'$  with posterior precision  $\delta' = 83.8$ .

With a meta-analysis such as this, it is of interest to quantify potential conflict among the effect estimates from the different studies. To do this, we follow [Presanis et al. \(2013\)](#) and compute a prior-predictive tail probability ([Box, 1980](#); [Evans and Moshonov, 2006](#)) for each study-specific estimate  $\hat{\theta}_i$ , with a meta-analytic estimate based on the remaining studies serving as the prior. As discussed above, fixed-effect meta-analysis is standard forward-Bayesian updating for normally distributed effect estimates with an initial flat prior. Hence, instead of fitting a reduced meta-analysis for each study, we can simply use the the Reverse-Bayes



**Figure 1:** Forest plot of fixed-effect meta-analysis of 7 randomized controlled trials investigating association between corticosteroids and mortality in hospitalized patients with COVID-19 (WHO REACT Working Group, 2020). Shown are number of deaths among total number of patients for treatment/control group, log odds ratio effect estimates with 95% confidence interval, two-sided  $p$ -values  $p$ , and prior-predictive tail probabilities  $p_{\text{Box}}$  with a meta-analytic estimate based on the remaining studies serving as the prior.

equations (7) and (8) together with the overall estimate to compute the parameters of the prior in the absence of the  $i$ -th study (denoted by the index  $-i$ ):

$$\begin{aligned}\delta_{-i} &= \delta' - 1/\sigma_i^2, \\ \mu_{-i} &= \frac{\mu' \delta' - \hat{\theta}_i / \sigma_i^2}{\delta_{-i}}.\end{aligned}$$

For example, through omitting the RECOVERY trial result  $\hat{\theta}_i = -0.53$  with standard error  $\sigma_i = 0.145$  we obtain  $\delta_{-i} = 36.1$  and  $\mu_{-i} = -0.26$ . A prior predictive tail probability using the approach from Box (1980) is then obtained by computing  $p_{\text{Box}} = \Pr(\chi_1^2 \geq t_{\text{Box}}^2)$  with

$$t_{\text{Box}} = \frac{\hat{\theta}_i - \mu_{-i}}{\sqrt{\sigma_i^2 + 1/\delta_{-i}}} = -1.24.$$

This leads to  $p_{\text{Box}} = 0.22$  for the RECOVERY trial, indicating very little prior-data conflict. The tail probabilities for the other studies are even larger, with the exception of the COVID STEROID trial ( $p_{\text{Box}} = 0.05$ ), see Figure 1. The lack of strong conflict can be seen as an informal justification of the assumptions of the underlying fixed-effect meta-analysis (Presanis et al., 2013; Ferkingstad et al., 2017). A related method in network meta-analysis is to assess consistency via “node-splitting” (Dias et al., 2010). Reverse-Bayes methods may also be useful



---

for conflict assessment in more general evidence synthesis methods where multiple distinct sources of data are combined (Goudie et al., 2019; Cunen and Hjort, 2021), but this may require more advanced numerical techniques.

Instead of determining the prior completely based on the posterior, one may also want to fix one parameter of the posterior and one parameter of the prior. This is of particular interest in order to challenge “significant” or “non-significant” findings through the Analysis of Credibility, as we will see in the following section.

## 2 Reverse-Bayes methods for the assessment of effect estimates

A more general question amenable to Reverse-Bayes methods is the assessment of effect estimates and their statistical significance or non-significance. This issue has recently attracted intense interest following the public statement of the American Statistical Association about the misuse and misinterpretation of the NHST concepts of statistical significance and non-significance (Wasserstein and Lazar, 2016). First investigated 20 years ago by Matthews (2001a,b), Reverse-Bayes methods for assessing both statistically significant and non-significant findings have been termed the Analysis of Credibility (Matthews, 2018) (or AnCred), whose principles and practice we now briefly review.

### 2.1 The Analysis of Credibility

Suppose the study gives rise to a conventional confidence interval for the unknown effect size  $\theta$  at level  $1 - \alpha$  with lower limit  $L$  and upper limit  $U$ . Assume that  $L$  and  $U$  are symmetric around the point estimate  $\hat{\theta}$  (assumed to be normally distributed with standard error  $\sigma$ ). AnCred then takes this likelihood and uses a Reverse-Bayes approach to deduce the prior required in order to generate credible evidence for the existence of an effect, in the form of a posterior that excludes no effect. As such, AnCred allows evidence deemed *statistically significant/non-significant* in the NHST framework to be assessed for its *credibility* in the Bayesian framework. As the latter conditions on the data rather than the null hypothesis, it is inferentially directly relevant to researchers. After a suitable transformation AnCred can be applied to a large number of commonly used effect measures such as differences in means, odds ratios, relative risks and correlations. We refer to the literature of meta-analysis for details about conversion among effect size scales, e. g., Cooper et al. (2019, chapter 11.6). The inversion of Bayes’s Theorem needed to assess credibility requires the form and location of the prior distribution to be specified. This in turn depends on whether the claim being assessed is statistically significant or non-significant; we consider each below.

#### Challenging statistically significant findings

A statistically significant finding at level  $\alpha$  is characterized by both  $L$  and  $U$  being either positive or negative. Equivalently  $z^2 > z_{\alpha/2}^2$  is required where  $z = \hat{\theta}/\sigma$  denotes the corresponding test statistic and  $z_{\alpha/2}$  the  $(1 - \alpha/2)$ -quantile of the standard normal distribution.

For significant findings, the idea is to ask how sceptical we would have to be not to find the apparent effect estimate convincing. To this end, a *sceptical prior* is derived such that the corresponding posterior credible interval just includes zero, the value of no effect. This critical prior interval can then be compared with internal or external evidence to assess if the finding is credible or not, despite being “statistically significant”.

More specifically, a Reverse-Bayes approach is applied to significant confidence intervals (at level  $\alpha$ ) based on a normally distributed effect estimate. The sceptical prior is a mean-zero normal distribution with variance  $\tau^2 = g \cdot \sigma^2$ , so the only free parameter is the relative prior variance  $g = \tau^2/\sigma^2$ . The posterior is hence also normal and either its lower  $\alpha/2$ -quantile (for positive  $\hat{\theta}$ ) or upper  $1 - \alpha/2$ -quantile (for negative  $\hat{\theta}$ ) is fixed to zero, so just represents “non-credible”. The sufficiently sceptical prior then has relative variance

$$g = \begin{cases} \frac{1}{z^2/z_{\alpha/2}^2 - 1} & \text{if } z^2 > z_{\alpha/2}^2 \\ \text{undefined} & \text{else} \end{cases} \quad (9)$$

see the Appendix in [Held \(2019a\)](#) for a derivation. The corresponding *scepticism limit* ([Matthews, 2018](#)), the upper bound of the equal-tailed sceptical prior credible interval at level  $1 - \alpha$ , is

$$SL = \frac{(U - L)^2}{4\sqrt{UL}}, \quad (10)$$

which holds for any value of  $\alpha$  provided the effect is significant at that level.

The left plot in Figure 2 illustrates the AnCred procedure for the finding from the RECOVERY trial, the only statistically significant result (at the conventional  $\alpha = 0.05$  level) shown in Figure 1. The trial found a decrease in COVID-19 mortality for patients treated with corticosteroids compared to usual care or placebo ( $\hat{\theta} = -0.53$  [95% CI  $-0.82$  to  $-0.25$ ]). The sufficiently sceptical prior has relative variance  $g = 0.39$ , so the sufficiently sceptical prior variance needs to be roughly 2.5 times smaller than the variance of the estimate to make the result non-credible. The scepticism limit on the log odds ratio scale turns out to be  $SL = 0.18$ , which corresponds to a critical prior interval with limits 0.84 and 1.19 on the odds ratio scale. Thus sceptics may still reject the RECOVERY trial finding as lacking credibility despite its statistical significance if external evidence suggests mortality reductions (in terms of odds) are unlikely to exceed  $1 - 0.84 \approx 16\%$ .

It is also possible to apply the approach to the meta-analytic log odds ratio estimate  $\hat{\theta} = -0.42$  [95% CI  $-0.63$  to  $-0.20$ ] from all 7 studies combined. Then  $SL = 0.13$ , so the meta-analytic estimate can be considered as non-credible if external evidence suggests that mortality reductions are unlikely to exceed  $1 - \exp(-SL) = 1 - 0.88 \approx 12\%$ . This illustrates that the meta-analytic estimate has gained credibility compared to the result from the RECOVERY study alone, despite the reduction in the effect estimate ( $\widehat{OR} = \exp(\hat{\theta}) = 0.66$  vs. 0.59 in the RECOVERY study).

---

## Challenging statistically non-significant findings

It is also possible to challenge “non-significant” findings (i.e., those for which the CI now includes zero, so  $z^2 < z_{\alpha/2}^2$ ) using a prior that pushes the posterior towards being credible in the Bayesian sense, with posterior credible interval no longer including zero, corresponding to no effect.

Matthews (2018) proposed the “advocacy prior” for this purpose, a normal prior with positive mean  $\mu$  and variance  $\tau^2$  chosen such that the  $\alpha/2$ -quantile is fixed to zero (for positive effect estimates  $\hat{\theta} > 0$ ). He showed that the “advocacy limit” AL, the  $(1 - \alpha/2)$ -quantile of the advocacy prior is

$$AL = -\frac{U+L}{2UL}(U-L)^2 \quad (11)$$

to reach credibility of the corresponding posterior at level  $\alpha$ . We show in Appendix A that the corresponding relative prior mean  $f = \mu/\hat{\theta}$  is

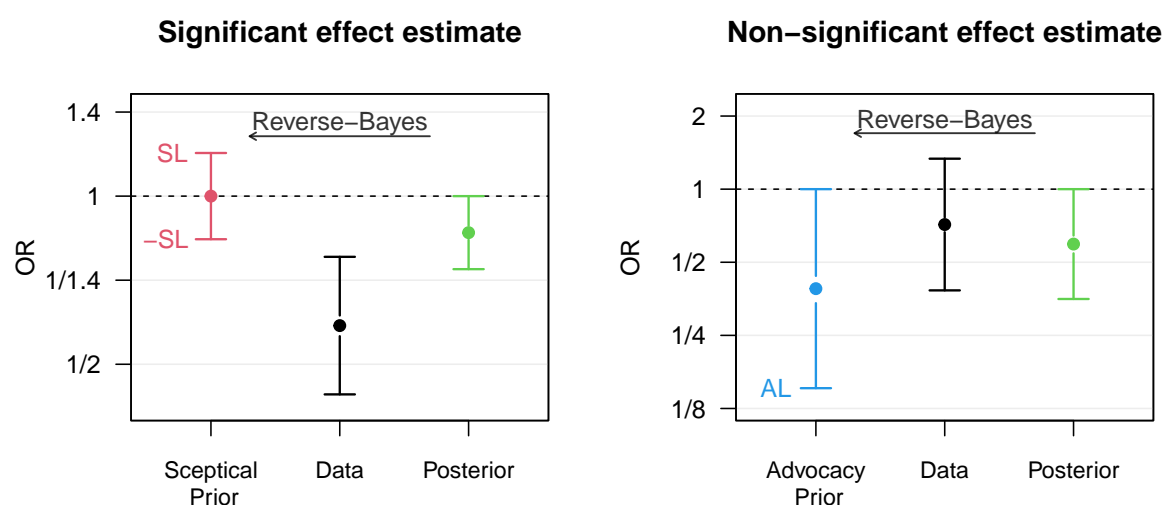
$$f = \begin{cases} \frac{2}{1 - z^2/z_{\alpha/2}^2} & \text{if } z^2 < z_{\alpha/2}^2 \\ \text{undefined} & \text{else.} \end{cases} \quad (12)$$

There are two important properties of the advocacy prior. First, the coefficient of variation CV is

$$CV = \tau/\mu = z_{\alpha/2}^{-1}.$$

The advocacy prior  $\theta \sim N(\mu, \tau^2 = \mu^2 CV^2)$  is hence characterized by a fixed coefficient of variation, so this prior has equal evidential weight (quantified in terms of  $\mu/\tau = z_{\alpha/2}$ ) as data which are “just significant” at level  $\alpha$ . Second, the advocacy limit AL defines the family of normal priors capable of rendering a “non-significant” finding credible at the same level. Such priors are summarized by the credible interval  $(L_o, U_o)$  where  $L_o \geq 0$ ,  $U_o \leq AL$ . Thus when confronted with a “non-significant” result – often, and wrongly, interpreted as indicating no effect – advocates of the existence of an effect may still claim the existence of the effect is credible to the same level if there exists prior evidence or insight compatible with the credible interval  $(L_o, U_o)$ . If the evidence for an effect is strong (weak), the resulting advocacy prior will be broad (narrow), giving advocates of an effect more (less) latitude to make their case under terms of AnCred. Note that (11) and (12) also hold for negative effect estimates, where we fix the  $(1 - \alpha/2)$ -quantile of the advocacy prior to zero and define the advocacy limit AL as the  $\alpha/2$ -quantile of the advocacy prior.

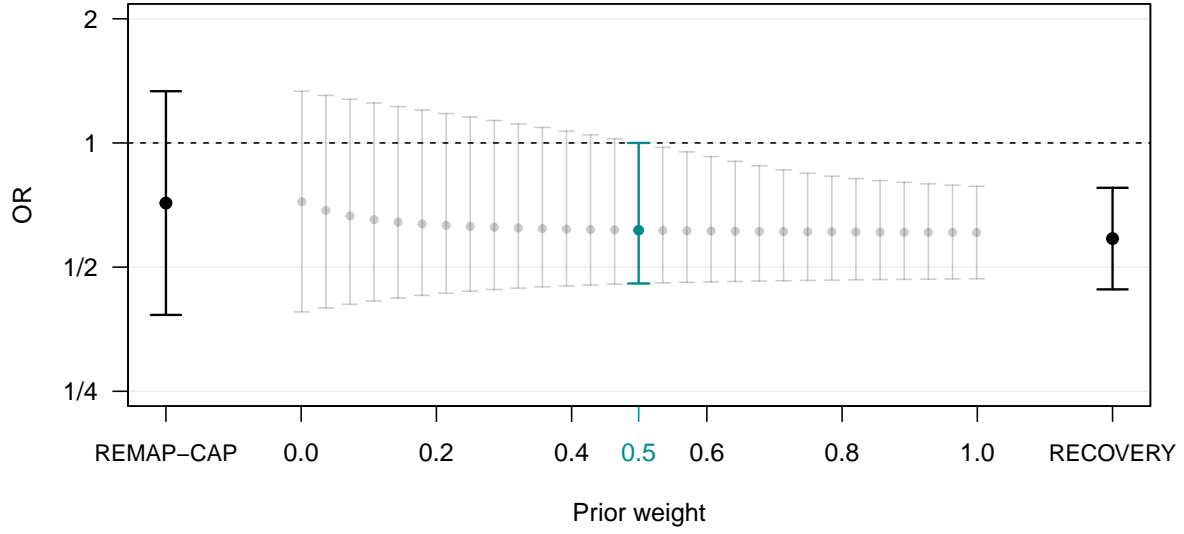
For illustration we consider the data from the REMAP-CAP trial that supported the RECOVERY trial finding of decreased COVID-19 mortality from corticosteroid use. However, this trial involved far fewer patients, and despite the point estimate showing efficacy, the relatively large uncertainty rendered the overall finding non-significant at the 5% level ( $\hat{\theta} = -0.34$  [95% CI  $-0.96$  to  $0.29$ ]). Such an outcome is frequently (and wrongly) taken to imply no effect. The use of AnCred leads to a more nuanced conclusion. The advocacy limit AL on the log odds ratio scale for REMAP-CAP is  $-1.89$ , i.e.,  $0.15$  on the odds ratio scale, see also the right plot in Figure 2. Thus advocates of the effectiveness of corticosteroids can regard the trial as providing credible evidence of effectiveness despite its non-significance if external evidence



**Figure 2:** Two examples of the Analysis of Credibility. Shown are point estimates within 95% confidence/credible intervals. The left plot illustrates how a sceptical prior is used to challenge the significant finding from the RECOVERY trial (RECOVERY Collaborative Group, 2020). The right plot illustrates how an advocacy prior is used to challenge a non-significant finding from the REMAP-CAP trial (REMAP-CAP Investigators, 2020). In both scenarios the posterior is fixed to be just non-credible/credible.

supports mortality reductions (in terms of odds) in the range 0% to 85%. So broad an advocacy range reflects the fact that this relatively small trial provides only modest evidential weight, and thus little constraint on prior beliefs about the effectiveness of corticosteroids.

Another way to push non-significant findings towards credibility is to use a prior based on data from another study or a different subgroup. For example, Best et al. (2021) consider results from the MENSA trial (Ortega et al., 2014) on the efficacy of Mepolizumab against placebo in 551 adult and 25 adolescent patients with severe asthma. The treatment effect was estimated to be positive in both subgroups but lacked significance among adolescents. Best et al. combine the data in the adolescent subgroup with a mixture prior based on a weak and an informative component. The weak component is a minimally informative normal prior with mean zero and large variance. The variance is chosen such that the information content of the prior is equivalent to that provided by a single subject or event (*unit-information prior*) (Kass and Wasserman, 1995). The other component is an informative prior based on the (significant) results from the adolescent subgroup. A Reverse-Bayes approach is used to determine how much prior weight one needs to assign to the informative component to obtain a credible posterior result with a 95% highest posterior credible interval no longer including zero. In the MENSA trial the required prior weight on the informative component was 0.7 (and thus 0.3 on the weak prior component) to achieve a credible result (Best et al., 2021). This illustrates that a considerable amount of “Bayesian borrowing” is required to extrapolate the results from adults to adolescents.



**Figure 3:** Illustration of the Reverse-Bayes borrowing method. The data from the (non-significant) REMAP-CAP trial are combined with a mixture prior consisting of the (significant) RECOVERY trial data and a unit-information prior (both estimates shown with 95% confidence interval). The resulting posterior medians with equal-tailed 95% credible intervals are shown for a range of mixing weights. The Reverse-Bayes mixing weight  $w = 0.5$  leads to the highlighted posterior with upper credible interval limit fixed at one.

In the meta-analytic setting we may ask a similar question: Suppose we want to combine the REMAP-CAP study results with a fraction of the RECOVERY trial data, how much weight do we need to assign to the RECOVERY trial to make the REMAP-CAP study credible? The unit-information prior for a logOR has variance  $\tau^2 = 4$  (Spiegelhalter et al., 2004, section 2.4.1), so the mixture prior is

$$\theta \sim w \cdot N(0, \tau^2 = 4) + (1 - w) \cdot N(\hat{\theta}_{\text{REC}}, \sigma_{\text{REC}}^2)$$

with  $w$  the mixing weight and point estimate  $\hat{\theta}_{\text{REC}}$  and squared standard error  $\sigma_{\text{REC}}^2$  of the RECOVERY trial, respectively. The resulting posterior is again a mixture of two normals with the posterior mean and variance of each component being the usual ones obtained from conjugate Bayesian updating, while the weights are proportional to the marginal likelihood of the data under each component (Best et al., 2021, section 3.5).

Figure 3 shows posterior medians with 95% equal-tailed credible intervals for a range of mixing weights. We see that a weight of at least  $w = 0.5$  is required to render the resulting posterior credible. Advocates of corticosteroids thus need to be able to justify such levels of prior beliefs, in order to conclude efficacy of corticosteroids also in the REMAP-CAP trial.

---

## Assessing credibility via equivalent prior study sizes

Reverse-Bayes credibility assessments can also be formulated in terms of the size and content of a prior study capable of challenging a claim of statistical significance/non-significance. This approach puts the weight of prior evidence in the context of the observed data, expressed as participant numbers. Greenland (2006) demonstrated the value of this approach in assessing the credibility of statistically significant findings from large observational studies in epidemiology. The same concept can, however, be extended to the assessment of both significant and non-significant outcomes more widely, such as small randomized controlled trials. For any study generating binary data in the form of event/non-event counts under two different conditions, the comparative effect measure can be expressed as a log-odds ratio with variance (squared standard error)

$$1/m_1 + 1/n_1 + 1/m_2 + 1/n_2. \quad (13)$$

where  $m_i$  and  $n_i$  are the numbers of events and non-events, respectively, in study arm  $i = 1, 2$ . This provides the link between the Reverse-Bayes prior distribution and the corresponding numbers of prior study participants. Using the simplifying assumptions of equal numbers of events  $m = m_1 = m_2$  and large numbers of non-events  $n_1$  and  $n_2$  in each arm ( $n_i > 100 m_i$ , say), the variance (13) reduces to  $2/m$ . The Reverse-Bayes sceptical prior defined in Section 2.1 has variance  $\tau^2 = \text{SL}^2/z_{\alpha/2}^2$ , where SL is the sceptical limit. Equating the two implies that such a prior is equivalent to a (large) hypothetical study with

$$m = 2/\tau^2 = 2z_{\alpha/2}^2/\text{SL}^2$$

events in both arms. The more compelling the data – that is, the smaller the value of SL – the larger the number of events  $m$  required in both arms of the hypothetical large prior study to render the result non-credible at level  $\alpha$ .

While the assumption of large studies can be appropriate with epidemiological studies involving rare events, it can be harder to justify for RCTs. Fortunately, the theory can be extended to encompass these and also the case of non-significant findings with little additional complexity. In the case of the sceptical prior, we simply require that the numbers of event  $m$  and non-events  $n$  are the same in both arms to constrain the mean to zero; the variance (13) then simplifies to  $2/m + 2/n$ . Adding the constraint that the event rate of the sceptical prior  $R = m/(m + n)$  matches that of the study under assessment, we then find

$$m = \frac{2}{\tau^2(1 - R)} \quad \text{and} \quad n = \frac{m(1 - R)}{R}.$$

For example, from Figure 1 the RECOVERY trial has an overall mortality rate  $R = (95 + 283)/(324 + 683) = 37.54\%$  and  $\text{SL} = 0.178$  at the  $\alpha = 5\%$  level ( $z_{\alpha/2} = 1.96$ ) corresponds to  $\tau^2 = 0.091^2$ , so  $m = 389$  and  $n = 648$  (these are integer approximations of exact computations), and thus a prior study capable of challenging the credibility of the RECOVERY trial requires 1037 patients and 389 deaths in each arm. At more than twice the size of the RECOVERY trial (2074 vs. 1007) patients and considerably more deaths in both arms, this level of sceptical prior evidence highlights the robustness of the trial finding.

A similar approach determines the characteristics of the hypothetical prior study needed to turn a “negative” non-significant finding into one that is credible to a specific  $\alpha$  level. The

---

Reverse-Bayes advocacy prior from [Matthews \(2018\)](#) described in Section 2.1 has a mean  $\mu = AL/2$  and variance  $\tau^2 = AL^2/(2z_{\alpha/2})^2$ . Under the large study assumption and equating the latter with (13) as before, the corresponding number of events needed to be observed in both arms of the hypothetical study is  $m = 2/\tau^2 = 8z_{\alpha/2}^2/AL^2$ . To incorporate the non-zero mean by which this prior represents advocacy, these  $m$  events are taken to have been observed among participants allocated to the two study arms in the ratio 1:  $K$  where  $K = \exp(\mu) = \exp(AL/2)$ , the allocation being such that it increases the relative evidential weight for the hypothesis “negated” by the non-significance.

As before, while the large study approximation may be justified in epidemiological examples, this is less likely to be true for RCTs. In such cases, we can adapt the approach used for sceptical priors, the size and composition of the advocacy prior being found by setting the numbers of events  $m$  in each arm the same, but this time allowing for different numbers of non-events in each arm via the allocation ratio  $K$ . The resulting variance is then

$$\tau^2 = \frac{2}{m} + \frac{K+1}{n}$$

where  $n$  is the number of non-events in the arm used to support the null hypothesis (e. g., the control arm in an RCT). With the control arm event rate  $R = m/(m+n)$  constrained to match that of the actual study, we find

$$m = \frac{2 - R(1 - K)}{\tau^2(1 - R)} \quad \text{and} \quad n = \frac{m(1 - R)}{R}.$$

As an example, we return to the REMAP-CAP trial, whose findings were consistent with a reduction of mortality but failed to reach statistical significance. As noted above, its advocacy limit ( $AL = -1.89$ ) implies this trial has relatively little evidential weight, and gives considerable scope for prior studies to make its outcome credible at the 95% level. With  $K = \exp(\mu) = \exp(AL/2) = 0.39$ ,  $R = 29/92$  and  $z_{\alpha/2} = 1.96$  we find  $m = 11$  and  $n = 25$ . Thus the hypothetical prior study comprises 11 deaths from  $11 + 25 = 36$  patients in the control arm and the same number of deaths from  $11 + (25/0.4) = 75$  patients in the treatment arm. At barely half the total size of REMAP-CAP but a considerably more impressive mortality reduction from  $R = 29/92 = 32\%$  in the control arm to  $11/75 = 15\%$  in the treatment arm (rather than  $26/105 = 25\%$  in REMAP-CAP), the nature of this hypothetical prior study confirms the paucity of evidence in the original trial.

### The fail-safe $N$ method

Another data representation of a sceptical prior forms the basis of the well-known “fail-safe  $N$ ” method, sometimes also called “file-drawer analysis”. This method, first introduced by [Rosenthal \(1979\)](#) and later refined by [Rosenberg \(2005\)](#), is commonly applied to the results from a meta-analysis and answers the question: “How many unpublished negative studies do we need to make the meta-analytic effect estimate non-significant?” A relatively large  $N$  of such unpublished studies suggests that the estimate is robust to potential null-findings, for



---

example due to publication bias. Calculations are made under the assumption that the unpublished studies have an average effect of zero and a precision equal to the average precision of the published ones.

While the method does not identify nor adjust for publication bias, it provides a quick way to assess how robust the meta-analytic effect estimate is. The method is available in common software packages such as *metafor* (Viechtbauer, 2010) and its simplicity and intuitive appeal have made it very popular among researchers.

AnCred and the fail-safe  $N$  are both based on the idea to challenge effect estimates such that they become “non-significant/not credible”, and it is easy to show that the methods are under some circumstances also technically equivalent. To illustrate this, we consider again the meta-analysis on the association between corticosteroids and COVID-19 mortality (WHO REACT Working Group, 2020) which gave the pooled log odds ratio estimate  $\hat{\theta} = -0.42$  with standard error  $\sigma = 0.11$ , posterior precision  $\delta' = 83.8$  and test statistic  $z = \hat{\theta}/\sigma = -3.81$ .

Using the Rosenberg (2005) approach (as implemented in the `fsn()` function from the *metafor* package) we find that at least  $N = 20$  additional but unpublished non-significant findings are needed to make the published meta-analysis effect non-significant. If instead, we challenge the overall estimate with AnCred, we obtain the relative prior variance  $g = 0.36$  using equation (9), so  $\tau^2 = 0.0043$ . Taking into account the average precision  $\delta'/n = 11.98$  of the different effect estimates estimates in the meta-analysis leads to  $N = n/(\delta' \cdot \tau^2) = 19.5$  which is equivalent to the fail-safe  $N$  result after rounding to the next larger integer.

## 2.2 Intrinsic credibility

The Problem of Priors is at its most challenging in the context of entirely novel “out of the blue” effects for which no obviously relevant external evidence exist. By their nature, such findings often attract considerable interest both within and beyond the research community, making their reliability of particular importance. Given the absence of external sources of evidence, Matthews (2018) proposed the concept of *intrinsic credibility*. This requires that the evidential weight of an unprecedented finding is sufficient to put it in conflict with the sceptical prior rendering it non-credible. In the AnCred framework, this implies a finding possesses intrinsic credibility at level  $\alpha$  if the estimate  $\hat{\theta}$  is outside the corresponding sceptical prior interval  $[-SL, SL]$  extracted using Reverse-Bayes from the finding itself, i. e.,  $\hat{\theta}^2 > SL^2$  with  $SL$  given in (10). Matthews showed this implies an unprecedented finding is intrinsically credible at level  $\alpha = 0.05$  if its  $p$ -value does not exceed 0.013.

Held (2019a) refined the concept by suggesting the use of a prior-predictive check (Box, 1980; Evans and Moshonov, 2006) to assess potential prior-data conflict. With this approach the uncertainty of the estimate  $\hat{\theta}$  is also taken into account since it is based on the prior-predictive distribution, in this case  $\hat{\theta} \sim N(0, \sigma^2 + \tau^2 = \sigma^2(1 + g))$  with  $g$  as given in (9). Intrinsic credibility is declared if the (two-sided) tail-probability

$$p_{\text{Box}} = \Pr(\chi_1^2 \geq \hat{\theta}^2/(\sigma^2 + \tau^2)) = \Pr(\chi_1^2 \geq z^2/(1 + g))$$



of  $\hat{\theta}$  under the prior-predictive distribution is smaller than  $\alpha$ . It turns out that the  $p$ -value associated with  $\theta$  needs to be at least as small as 0.0056 to obtain intrinsic credibility at level  $\alpha = 0.05$ , providing another principled argument for the recent proposition to lower the  $p$ -value threshold for the claims of new discoveries to 0.005 (Benjamin et al., 2017). A simple check for intrinsic credibility is based on the *credibility ratio*, the ratio of the upper to the lower limit (or vice versa) of a confidence interval for a significant effect size estimate. If the credibility ratio is smaller than 5.8 then the result is intrinsically credible (Held, 2019a). This holds for confidence intervals at all possible values of  $\alpha$ , not just for the 0.05 standard. For example, in the RECOVERY study the 95% confidence interval for the log-odds ratio ranges from  $-0.82$  to  $-0.25$ , so the credibility ratio is  $-0.82 / -0.25 = 3.27 < 5.8$  and the result is intrinsically credible at the standard 5% level.

### Replication of effect direction

Whether intrinsic credibility is assessed based on the prior or the prior-predictive distribution, it depends on the level  $\alpha$  in both cases. To remove this dependence, Held (2019a) proposed to consider the smallest level at which intrinsic credibility can be established, defining the  $p$ -value for intrinsic credibility

$$p_{\text{IC}} = 2 \left\{ 1 - \Phi \left( \frac{|z|}{\sqrt{2}} \right) \right\}, \quad (14)$$

see section 4 in Held (2019a) for the derivation. Now  $z = \hat{\theta}/\sigma$ , so compared to the standard  $p$ -value  $p = 2 \{1 - \Phi(|z|)\}$ , the  $p$ -value for intrinsic credibility is based on twice the variance  $\sigma^2$  of the estimate  $\hat{\theta}$ . Although motivated from a different perspective, inference based on intrinsic credibility thus mimics the *doubling the variance rule* advocated by Copas and Eguchi (2005) as a simple means of adjusting for model uncertainty.

Moreover, Held (2019a) showed that  $p_{\text{IC}}$  is connected to  $p_{\text{rep}}$  of Killeen (2006), the probability that a replication will result in an effect estimate  $\hat{\theta}_r$  in the same direction as the observed effect estimate  $\hat{\theta}$ , by  $p_{\text{rep}} = 1 - p_{\text{IC}}/2$ . Hence, an intrinsically credible estimate at a small level  $\alpha$  will have high chance of replicating since  $p_{\text{rep}} \geq 1 - \alpha/2$ . Note that  $p_{\text{rep}}$  lies between 0.5 and 1 with the extreme case  $p_{\text{rep}} = 0.5$  if  $\hat{\theta} = 0$ .

As an example, the  $p$ -value for intrinsic credibility for the RECOVERY trial finding (with  $p$ -value  $p = 0.0002$ ) cited earlier is  $p_{\text{IC}} = 0.01$  and thus the probability of the replication effect going in the same direction (i.e., reduced mortality in this case) is 0.995. In contrast, the finding from the smaller REMAP-CAP trial (with  $p = 0.29$ ) leads to  $p_{\text{IC}} = 0.46$ , and the probability of effect direction replication is hence only 0.77.

## 3 Reverse-Bayes methods with Bayes factors

The AnCred procedure as described above uses posterior credible intervals as a means of quantifying evidence. However, quantification of evidence with Bayes factors is a more principled solution for hypothesis testing in the Bayesian framework (Jeffreys, 1961; Kass and

Raftery, 1995). Bayes factors enable direct probability statements about null and alternative hypothesis and they can also quantify evidence *for* the null hypothesis, both are impossible with indirect measures of evidence such as  $p$ -values (Held and Ott, 2018). Reverse-Bayes approaches combined with Bayes factor methodology was pioneered by Carlin and Louis (1996) but then remained unexplored until Pawel and Held (2022) proposed an extension of AnCred where Bayes factors are used as a means of quantifying evidence. Rather than determining a prior such that a finding becomes “non-credible” in terms of a posterior credible interval, this approach determines a prior such that the finding becomes “non-compelling” in terms of a Bayes factor. In the second step of the procedure, the plausibility of this prior is quantified using external data from a replication study. Here, we will illustrate the methodology using only an original study; we mention extensions for replications in Section 5.1.

### 3.1 Sceptical priors

As before,  $\hat{\theta}$  denotes the estimate of the unknown mean  $\theta$  of a  $N(\theta, \sigma^2)$  distribution with known variance  $\sigma^2$ . A standard hypothesis test compares the null hypothesis  $H_0: \theta = 0$  to the alternative  $H_1: \theta \neq 0$ . Bayesian hypothesis testing requires specification of a prior distribution of  $\theta$  under  $H_1$ . A typical choice is a local alternative, a unimodal symmetric prior distribution centred around the null value (Johnson and Rossell, 2010). We consider again the conjugate sceptical prior  $\theta | H_1 \sim N(0, \tau^2 = g \cdot \sigma^2)$  with relative prior variance  $g$  for this purpose. This leads to the Bayes factor comparing  $H_0$  to  $H_1$  being

$$BF_{01} = \sqrt{1+g} \cdot \exp \left\{ -\frac{g}{1+g} \cdot \frac{z^2}{2} \right\}, \quad (15)$$

where  $z = \hat{\theta}/\sigma$ . Yet again, the amount of evidence which the data provide against the null hypothesis depends on the prior parameter  $g$ ; As  $g$  becomes smaller ( $g \downarrow 0$ ), the null hypothesis and the alternative will become indistinguishable, so the data are equally likely under both ( $BF_{01} \rightarrow 1$ ). On the other hand, for increasingly diffuse priors ( $g \rightarrow \infty$ ), the null hypothesis will always prevail ( $BF_{01} \rightarrow \infty$ ) due to the Jeffreys-Lindley paradox (Robert, 2014). In between, the  $BF_{01}$  reaches a minimum at  $g = \max \{z^2 - 1, 0\}$  leading to

$$\min BF_{01} = \begin{cases} |z| \cdot \exp \{ -z^2/2 \} \cdot \sqrt{e} & \text{if } |z| > 1 \\ 1 & \text{else} \end{cases} \quad (16)$$

which is an instance of a *minimum Bayes factor*, the smallest possible Bayes factor within a class of alternative hypotheses, in this case zero-mean normal alternatives (Edwards et al., 1963; Berger and Sellke, 1987; Sellke et al., 2001; Held and Ott, 2018).

Reporting of minimum Bayes factors is one attempt of solving the Problem of Priors in Bayesian inference. However, this bound may be rather small and the corresponding prior unrealistic. In contrast, the Reverse-Bayes approach makes the choice of the prior explicit by determining the relative prior variance parameter  $g$  such that the finding is no longer compelling, followed by assessing the plausibility of this prior. To do so, one first fixes  $BF_{01} = \gamma$ , where  $\gamma$  is a cut-off above which the result is no longer convincing, for example  $\gamma = 1/10$ , the

level for strong evidence according to the classification from [Jeffreys \(1961\)](#). The sufficiently sceptical relative prior variance is then given by

$$g = \begin{cases} -\frac{z^2}{q} - 1 & \text{if } -\frac{z^2}{q} \geq 1 \\ \text{undefined} & \text{else} \end{cases} \quad (17)$$

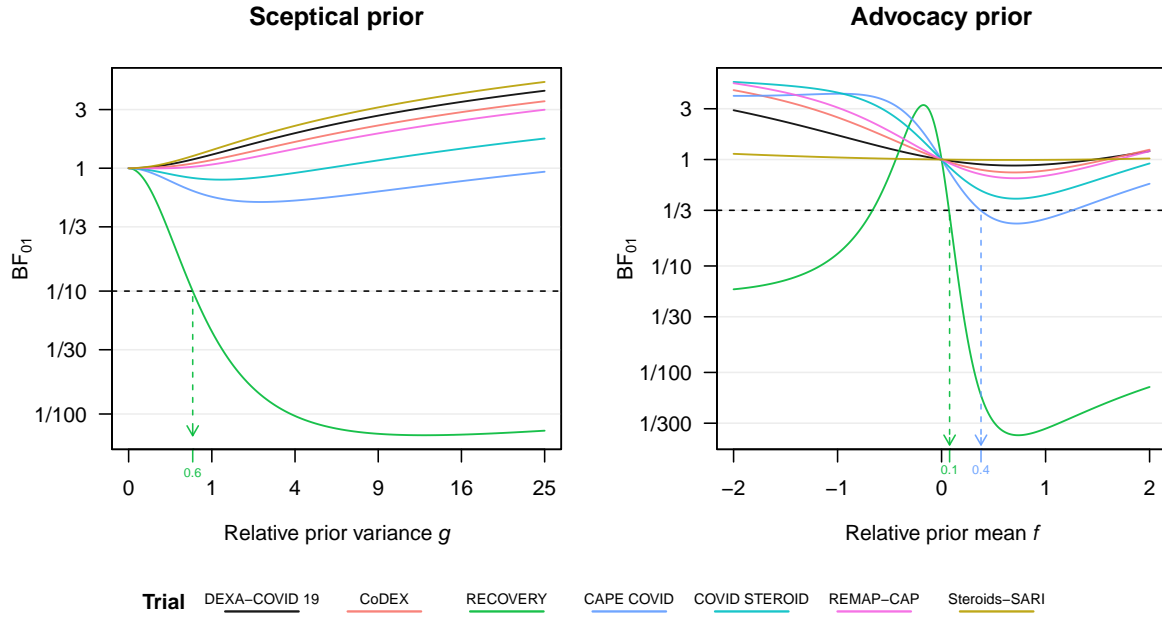
where  $q = W\left(-\frac{z^2}{\gamma^2} \cdot \exp\{-z^2\}\right)$

where  $W(\cdot)$  is the branch of the Lambert  $W$  function that satisfies  $\text{text}W(y) \leq -1$  for  $y \in [-e^{-1}, 0)$  ([Corless et al., 1996](#)), see the Appendix in ([Pawel and Held, 2022](#)) for a proof.

The sufficiently sceptical relative prior variance  $g$  exists only for a cut-off  $\gamma$  if  $\min \text{BF}_{01} \leq \gamma$ , similar to standard AnCred where it exists only at level  $\alpha$  if the original finding was significant at the same level. In contrast to standard AnCred, however, if the sufficiently sceptical relative prior variance  $g$  exists, there are always two solutions, a consequence of the Jeffreys-Lindley paradox: If  $\text{BF}_{01}$  decreases in  $g$  below the chosen cut-off  $\gamma$ , after attaining its minimum it will monotonically increase and intersect a second time with  $\gamma$ , admitting a second solution for the sufficiently sceptical prior.

We now revisit the meta-analysis example considered earlier: The left plot in Figure 4 shows the Bayes factor  $\text{BF}_{01}$  from (15) as a function of the relative prior variance  $g$  for each finding included in the meta-analysis. Most of them did not include a great number of participants and thus provide little evidence against the null hypothesis for any value of the relative prior variance  $g$ . In contrast, the finding from the RECOVERY trial provides more compelling evidence and can be challenged up to  $\min \text{BF}_{01} = 1/148.9$ . For example, we see in Figure 4 that the relative sceptical prior variance needs to be  $g \leq 0.59$  such that the finding is no longer compelling at level  $\gamma = 1/10$ . This translates to a 95% prior credible interval from 0.8 to 1.24 for the OR (or any narrower interval around 1). Hence, a sceptic might still consider the RECOVERY finding to be unconvincing, despite its minimum BF being very compelling, if external evidence supports ORs in that range. By applying the prior-to-data conversion method described in Section 2.1 we can further see that the evidential value of this prior is equivalent to a trial with 258 events and 429 non-events in both arms (so that the overall mortality rate is equivalent with the RECOVERY trial). For comparison, the sceptical prior from standard AnCred at  $\alpha = 0.05$  was equivalent to a trial with 389 events and 648 non-events, respectively.

The plausibility of the sufficiently sceptical prior can be evaluated in light of external evidence, but what should we do in the absence of such? We could again use the [Box \(1980\)](#) prior-predictive check as in Section 2.2, however, the resulting tail probability is difficult to compare to the Bayes-factor cut-off  $\gamma$ . When a specific alternative model to the null is in mind, [Box \(1980, p. 391\)](#) also suggested to use a Bayes factor for model criticism of the null model. Following this approach, [Pawel and Held \(2022\)](#) proposed to define a second Bayes factor contrasting the sufficiently sceptical prior to an optimistic prior, which they defined as  $\theta | H_2 \sim N(\hat{\theta}, \sigma^2)$  the posterior of  $\theta$  based on the data and the reference prior  $f(\theta) \propto 1$ . The optimistic prior therefore represents the position of a proponent who takes the original claim at face



**Figure 4:** Illustration of the AnCred with Bayes factors procedure using the findings from the meta-analysis on the association of COVID-19 mortality and corticosteroids. The left plot shows the Bayes factor  $BF_{01}$  as a function of the relative variance  $g$  of the sceptical prior. The result from the RECOVERY trial is challenged with a sceptical prior such that  $BF_{01} = 1/10$ , for the other trials such a prior does not exist. The right plot shows the Bayes factor  $BF_{01}$  as a function of the relative mean  $f = \mu/\hat{\theta}$  of the advocacy prior where the coefficient of variation from the prior is fixed to  $CV = \tau/\mu = 1/z(\gamma = 1/3) = 0.67$ , where  $z(\gamma)$  is given in (20). The RECOVERY and the CAPE COVID findings are challenged such that  $BF_{01} = 1/3$ , for the other trials such a prior does not exist.

value. This leads to the second Bayes factor being

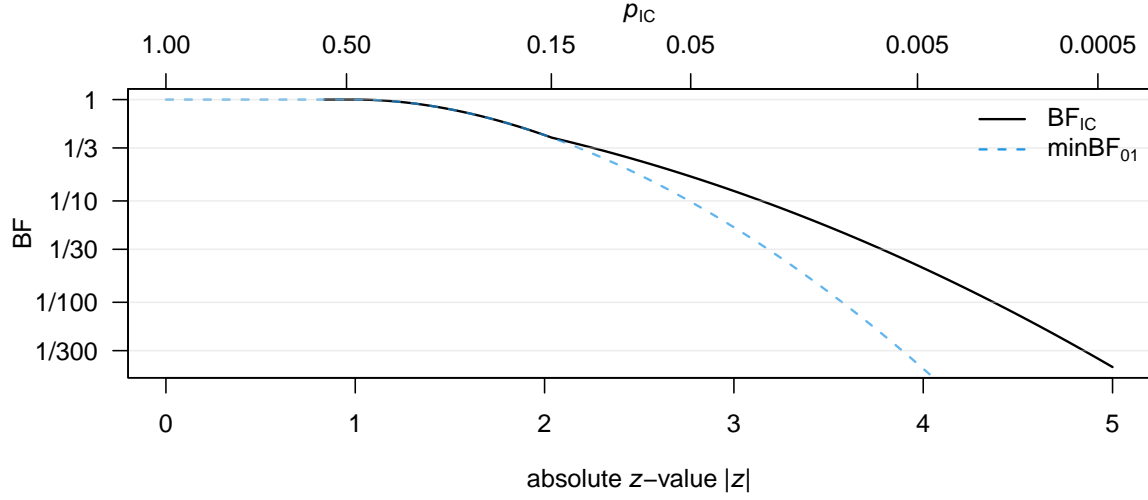
$$BF_{12} = \sqrt{\frac{2}{1+g}} \cdot \exp \left\{ -\frac{1}{2} \cdot \frac{z^2}{1+g} \right\}. \quad (18)$$

Analogously to the tail probability approach from Section 2.2, intrinsic credibility is established if the data support the optimistic over the sceptical prior at a higher level than they support the sceptical prior over the null hypothesis, i. e., if

$$BF_{12} \leq BF_{01}$$

with sufficiently sceptical relative prior variance  $g$  from (17) used in both Bayes factors. For example, if we challenge the RECOVERY trial finding such that the resulting Bayes factor is only  $BF_{01} = 1/10$ , we obtain with (9) the sufficiently sceptical relative prior variance  $g = 0.59$  and in turn with (18) the Bayes factor  $BF_{12} = 1/64$ , so the finding is intrinsically credible at  $\gamma = 1/10$ .

To remove the dependence on the choice of the level  $\gamma$ , one can determine the smallest level  $\gamma$  where intrinsic credibility can be established. This defines a Bayes factor for intrinsic credibility  $BF_{IC}$  similar to the definition of the  $p$ -value for intrinsic credibility  $p_{IC}$  from (14). Intrinsic



**Figure 5:** Comparison of the Bayes factor for intrinsic credibility  $BF_{IC}$ , the minimum Bayes factor  $\min BF_{01}$ , and the  $p$ -value for intrinsic credibility  $p_{IC}$  as a function of the absolute  $z$ -value  $|z|$ . The value  $p_{IC} = 0.15$  is at the breakpoint at  $|z| = 2.04$ .

credibility at level  $\gamma$  is then equivalent with  $BF_{IC} \leq \gamma$ . Details on the computation of  $BF_{IC}$  are given in Appendix B. For the RECOVERY finding, the Bayes factor for intrinsic credibility is  $BF_{IC} = 1/25$ . This means the data favour the optimistic prior over any sceptical prior that is capable of rendering the original result no longer convincing at  $\gamma = 1/25$ . For comparison the  $p$ -value for intrinsic credibility (14) is  $p_{IC} = 0.009$ .

Figure 5 shows the Bayes factor for intrinsic credibility  $BF_{IC}$  as a function of the  $z$ -value along with a comparison to the  $p$ -value for intrinsic credibility  $p_{IC}$  and the minimum Bayes factor  $\min BF_{01}$  from (16). We see that the  $BF_{IC}$  is undefined when  $|z| < \sqrt{\log 2} \approx 0.83$ . In this case the data are so unconvincing that any sceptical prior is better supported by the data than the optimistic prior. For  $z$ -values between  $\sqrt{\log 2} \leq |z| < 2.04$ , the  $BF_{IC}$  equals the minimum Bayes factor  $\min BF_{01}$ , whereas for larger  $z$ -values  $|z| \geq 2.04$ , the  $BF_{IC}$  is always larger (more conservative) than the  $\min BF_{01}$ . In the absence of any prior information, it may therefore be a useful evidential summary which formally takes into account both scepticism and optimism about the observed data.

A  $p$ -value less than 0.05 is usually regarded as sufficient evidence against the null hypothesis, but how much evidence does  $p = 0.05$  mean in terms of the Bayes factor for intrinsic credibility? From Figure 5, we see that the  $BF_{IC} = 1/2.1$  for  $|z| = 1.96$ , so at most “worth a bare mention” according to the classification from [Jeffreys \(1961\)](#). Thus, also from this perspective, the conventional  $p$ -value threshold of 0.05 for the claim of new discoveries seems too lax in terms of the evidential value that a finding at this threshold provides. We saw in Section 2.2 that an ordinary  $p$ -value needs to be at least as small as  $p \leq 0.0056$  for a finding to be intrinsically credible in terms of the  $p$ -value for intrinsic credibility  $p_{IC} \leq 0.05$ . A  $p$ -value of 0.0056 corresponds to  $|z| = 2.77$  where the Bayes factor for intrinsic credibility is  $BF_{IC} = 1/5.7$ , in-

dicating at least “substantial” evidence against the null hypothesis according to Jeffreys. To achieve intrinsic credibility at the level for strong evidence ( $\gamma = 1/10$ ) the requirements are even more stringent as the  $z$ -value needs to be at least  $|z| \geq 3.15$  (equivalent to  $\min\text{BF} \leq 1/27$ ,  $p \leq 0.002$ , or  $p_{\text{IC}} \leq 0.026$ ).

### 3.2 Advocacy priors

A natural question is whether we can also define an advocacy prior, a prior which renders an unconvincing finding convincing, in the AnCred framework with Bayes factors. In traditional AnCred, advocacy priors always exist since one can always find a prior that, when combined with the data, can overrule them. This is fundamentally different to inference based on Bayes factors, where the prior is not synthesized with the data, but rather used to predict them. A classical result due to [Edwards et al. \(1963\)](#) states that if we consider the class of all possible priors under  $H_1$ , the minimum Bayes factor is given by

$$\min\text{BF}_{01} = \exp \{-z^2/2\} \quad (19)$$

which is obtained for  $H_1: \theta = \hat{\theta}$ . This implies that a non-convincing finding can not be “rescued” further than to this bound. For example, for the finding from the REMAP-CAP trial the bound is unsatisfactorily  $\min\text{BF}_{01} = 1/1.7$ , so at most “worth a bare mention” according to the classification from [Jeffreys \(1961\)](#).

Putting these considerations aside, we may still consider the class of  $N(\mu, \tau^2)$  priors under the alternative  $H_1$ . The Bayes factor contrasting  $H_0$  to  $H_1$  is then given by

$$\text{BF}_{01} = \sqrt{1 + \tau^2/\sigma^2} \cdot \exp \left\{ -\frac{1}{2} \left[ \frac{\hat{\theta}^2}{\sigma^2} - \frac{(\hat{\theta} - \mu)^2}{\sigma^2 + \tau^2} \right] \right\}.$$

The Reverse-Bayes approach now determines the prior mean  $\mu$  and variance  $\tau^2$  which lead to the Bayes factor  $\text{BF}_{01}$  being just at some cut-off  $\gamma$ . However, if both parameters are free, there are infinitely many solutions to  $\text{BF}_{01} = \gamma$ , if any exist at all. The traditional AnCred framework resolves this by restricting the class of possible priors to advocacy priors with fixed coefficient of variation of  $\text{CV} = \tau/\mu = 1/z_{\alpha/2}$ . We can translate this idea to the Bayes factor AnCred framework and fix the prior’s coefficient of variation to  $\text{CV} = 1/z(\gamma)$ , where

$$z(\gamma) = \sqrt{-2 \log \gamma}, \quad (20)$$

obtained by solving (19) for  $z$  with  $\min\text{BF}_{01} = \gamma$ . The advocacy prior thus carries the same evidential weight as data with  $\min\text{BF}_{01} = \gamma$ . Moreover, the determination of the prior parameters becomes more feasible since there is only one free parameter left (either  $\mu$  or  $\tau^2$ ).

The right plot in Figure 4 illustrates application of the procedure on data from the meta-analysis on association between COVID-19 mortality and corticosteroids. The coefficient of variation of the advocacy prior is fixed to  $\text{CV} = 1/z(\gamma = 1/3) = 0.67$  (for comparison, the CV of the advocacy prior in traditional AnCred at  $\alpha = 0.05$  is  $\text{CV} = 1/z_{\alpha/2} = 0.51$ ) and thus the Bayes factor  $\text{BF}_{01}$  only depends on the relative mean  $f = \mu/\hat{\theta}$ . Under the sceptical prior only the RECOVERY finding could be challenged at  $\gamma = 1/3$  (where  $z(\gamma) = 1.5$  corresponds



to  $\alpha = 13\%$ ). With the advocacy prior this is now also possible for the CAPE COVID finding (Dequin et al., 2020), where a prior with mean  $\mu = f \cdot \hat{\theta} = 0.37 \cdot (-0.79) = -0.29$  and standard deviation  $\tau = \text{CV} \cdot \mu = 0.2$  is able to make the finding compelling at  $\gamma = 1/3$ . The corresponding prior credible interval for the OR at level  $1 - \alpha$  ranges from 0.55 to 1, so advocates may still consider the “non-compelling” finding as providing moderate evidence in favour of a benefit, if external evidence supports mortality reductions in that range. Using the prior-to-data conversion described in Section 2.1, the prior can be translated to a trial with 69 events in both arms, but 206 non-events in the treatment and 182 non-events in the control arm (such that the mortality rate in the control arm is the same as in the CAPE-COVID trial). Note that the advocacy prior may not be unique, e. g., for the CAPE COVID finding the prior with relative mean  $f' = 1.26$  and standard deviation  $\tau' = 0.67$  also renders the data as just compelling at  $\gamma = 1/3$ . We recommend to choose the prior with  $f$  closer to zero, as it is the more conservative choice.

## 4 Reverse-Bayes analysis of the False Positive Risk

Application of the Analysis of Credibility with Bayes factors as described in Section 3 assumes some familiarity with Bayes factors as measures of evidence. Colquhoun (2019) argued that very few nonprofessional users of statistics are familiar with the notion of Bayes factors or likelihood ratios. He proposes to quantify evidence with the *false positive risk*, “if only because that is what most users still think, mistakenly, that that is what the  $p$ -value tells them”. More specifically, Colquhoun (2019) defines the FPR as the posterior probability that the point null hypothesis  $H_0$  of no effect is true given the observed  $p$ -value  $p$ , i. e.,  $\text{FPR} = \Pr(H_0 | p)$ . As before,  $H_0$  corresponds to the point null hypothesis  $H_0: \theta = 0$ . Note also that we take the exact (two-sided)  $p$ -value  $p$  as the observed “data”, regardless of whether or not it is significant at some pre-specified level, the so-called “ $p$ -equals” interpretation of NHST (Colquhoun, 2017).

FPR can be calculated based on the Bayes factor associated with  $p$ . For ease of presentation we invert Bayes’ theorem (1) and obtain

$$\frac{\text{FPR}}{1 - \text{FPR}} = \frac{\Pr(H_0 | p)}{\Pr(H_1 | p)} = \text{BF}_{01} \frac{\Pr(H_0)}{\Pr(H_1)}, \quad (21)$$

where  $\text{BF}_{01} = 1/\text{BF}_{10}$  is the Bayes factor for  $H_0$  against  $H_1$ , computed directly from the observed  $p$ -value  $p$ .

The common ‘forward-Bayes’ approach is to compute the FPR from the prior probability  $\Pr(H_0)$  and the Bayes factor with (21). However, the prior probability  $\Pr(H_0)$  is usually unknown in practice and often hard to assess. This can be resolved via the Reverse-Bayes approach (Colquhoun, 2017, 2019): Given a  $p$ -value and a false positive risk value, calculate the corresponding prior probability  $\Pr(H_0)$  that is needed to achieve that false positive risk. Of specific interest is the value  $\text{FPR} = 5\%$ , because many scientists believe that a Type-I error of 5% is equivalent to a FPR of 5% (Greenland et al., 2016). This is of course not true and we

follow Example 1 from [Berger and Sellke \(1987\)](#) and use the Reverse-Bayes approach to derive the necessary prior assumptions on  $\Pr(H_0)$  to achieve  $\text{FPR} = 5\%$  with Equation (21):

$$\Pr(H_0) = \left[ 1 + \frac{1 - \text{FPR}}{\text{FPR}} \cdot \text{BF}_{01} \right]^{-1}. \quad (22)$$

[Colquhoun \(2017\)](#) uses a Bayes factor based on the  $t$ -test, but for compatibility with the previous sections we assume normality of the underlying test statistic. We consider Bayes factors under all simple alternatives, but also Bayes factors under local normal priors, see [Held and Ott \(2018\)](#) for a detailed comparison.

Instead of working with a Bayes factor for a specific prior distribution, we prefer to work with the minimum Bayes factor  $\min\text{BF}_{01}$  as introduced in Section 3.1. In what follows we will use the minimum Bayes factor based on the  $z$ -test, see Section 2.1 and 2.2 in [Held and Ott \(2018\)](#).

Let  $\min\text{BF}_{01}$  denote the minimum Bayes factor over a specific class of alternatives. From equation (22) we obtain the inequality

$$\Pr(H_0) \leq \left[ 1 + \frac{1 - \text{FPR}}{\text{FPR}} \cdot \min\text{BF}_{01} \right]^{-1}. \quad (23)$$

The right-hand side is thus an upper bound on the prior probability  $\Pr(H_0)$  for a given  $p$ -value to achieve a pre-specified FPR value.

There are also  $\min\text{BF}$ s not based on the  $z$ -test statistic as (16), but directly on the (two-sided)  $p$ -value  $p$ , the so-called “ $-e p \log p$ ” ([Sellke et al., 2001](#)) calibration

$$\min\text{BF} = \begin{cases} -e p \log p & \text{for } p < 1/e \\ 1 & \text{otherwise,} \end{cases} \quad (24)$$

and the “ $-e q \log q$ ” calibration, where  $q = 1 - p$ , see Section 2.3 in [Held and Ott \(2018\)](#):

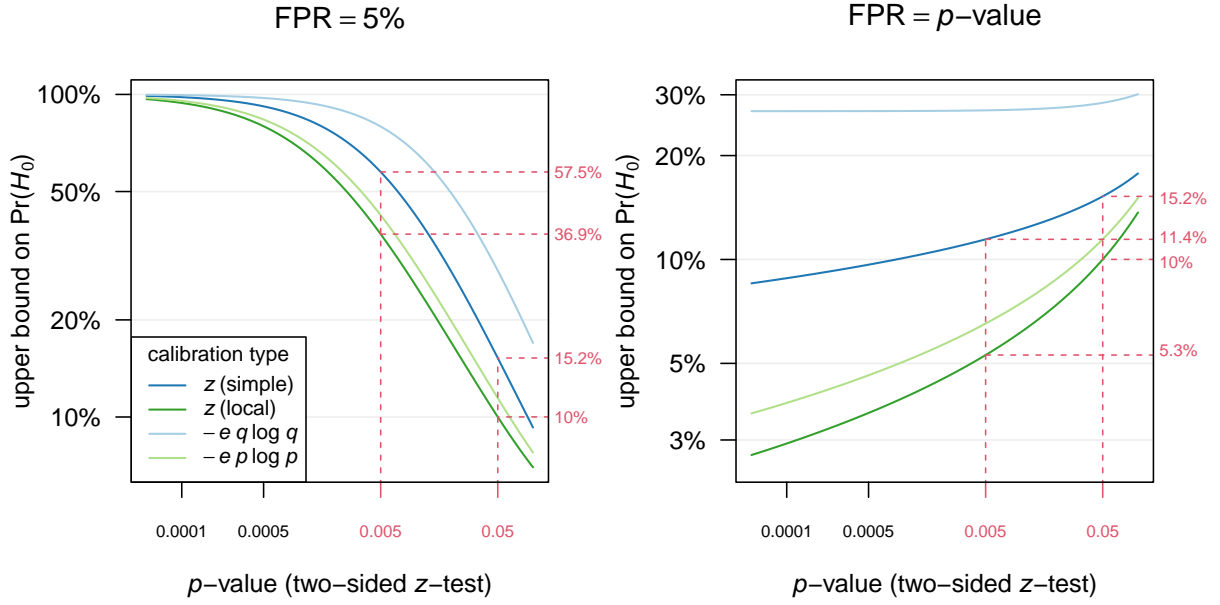
$$\min\text{BF} = \begin{cases} -e (1 - p) \log(1 - p) & \text{for } p < 1 - 1/e \\ 1 & \text{otherwise.} \end{cases} \quad (25)$$

For small  $p$ , equation (25) can be simplified to  $\min\text{BF} \approx e p$ , which mimics the [Good \(1958\)](#) transformation of  $p$ -values to Bayes factors ([Held, 2019b](#)).

The two  $p$ -based calibrations carry less assumptions than the minimum Bayes factors based on the  $z$ -test under normality and can be used as alternative expressions in (23). The “ $-e p \log p$ ” provides a general bound under all unimodal and symmetrical local priors for  $p$ -values from  $z$ -tests, see Section 3.2 in [Sellke et al. \(2001\)](#). The “ $-e q \log q$ ” calibration is more conservative and gives a smaller bound on the Bayes factor than the “ $-e p \log p$ ” calibration. It can be viewed as a general lower bound under simple alternatives where the direction of the effect is taken into account, see Section 2.1 and 2.3 in [Held and Ott \(2018\)](#).

The left plot in Figure 6 shows the resulting upper bound on the prior probability  $\Pr(H_0)$  as a function of the two-sided  $p$ -value if the FPR is fixed at 5%. For  $p = 0.05$ , the “ $-e p \log p$ ” bound is around 11% and 28% for the “ $-e q \log q$ ” calibration. The corresponding values based on the  $z$ -test are slightly smaller (10% and 15%, respectively). All the probabilities are





**Figure 6:** The left plot shows the upper bound on the prior probability  $\Pr(H_0)$  to achieve a false positive risk of 5% as a function of the  $p$ -value calibrated with either a  $z$ -test calibration (simple or local alternatives) or with the “ $-e p \log p$ ” or “ $-e q \log q$ ” calibrations, respectively. The right plot shows the upper bound on  $\Pr(H_0)$  as a function of the  $p$ -value using the same calibrations but assuming the  $p$ -value equals the FPR.

below the 50% value of equipoise, illustrating that borderline significant result with  $p \approx 0.05$  do not provide sufficient evidence to justify an FPR value of 5%. For  $p = 0.005$ , the upper bounds are closer to 50% (37% for local and 57% for simple alternatives).

Turning again to the example from the RECOVERY trial, the  $p$ -value associated with the estimated treatment effect is  $p = 0.0002$ . The left plot in Figure 6 shows that the false positive risk can safely be assumed to be around 5% (or lower), since the upper bound on  $\Pr(H_0)$  are all very large for such a small  $p$ -value.

Fixing FPR at the 5% level may be considered as arbitrary. Another widespread misconception is the belief that the FPR is equal to the  $p$ -value. Held (2013) used a Reverse-Bayes approach to investigate which prior assumptions are required such that  $\text{FPR} = p$  holds. Combining (22) with the “ $-e p \log p$ ” calibration (24) gives the explicit condition

$$\Pr(H_0) \leq 1 / \{1 - e(1 - p) \log(p)\}$$

whereas the “ $-e q \log q$ ” calibration (25) leads to

$$\Pr(H_0) \leq 1 / \left\{ 1 - e \frac{(1 - p)^2}{p} \log(1 - p) \right\} \approx 1 / \{1 + e(1 - p)\},$$

which is approximately  $1/(1 + e) = 26.9\%$  for small  $p$ .

---

The right plot in Figure 6 compares the bounds based on these two calibrations with the ones obtained from simple respectively local alternatives. We can see that strong assumptions on  $\Pr(H_0)$  are needed to justify the claim  $\text{FPR} = p$ :  $\Pr(H_0)$  cannot be larger than 15.2% if the  $p$ -value is conventionally significant ( $p < 0.05$ ). For  $p < 0.005$ , the bound drops further to 11.4%. Even under the conservative “ $-e q \log q$ ” calibration, the upper bound on  $\Pr(H_0)$  is 26.9% for small  $p$  and increases only slightly for larger values of  $p$ . This illustrates that the misinterpretation  $\text{FPR} = p$  only holds if the prior probability of  $H_0$  is substantially smaller than 50%, an assumption which is questionable in the absence of strong external knowledge.

## 5 Discussion

### 5.1 Extensions, work in progress and outlook

The Reverse-Bayes methods described above have focused on the comparison of the prior needed for credibility with findings from other studies and/or more general insights. However, replication studies make an obvious additional source of external evidence, as these are typically conducted to confirm original findings by repeating their experiments as closely as possible. The question is then whether the original findings have been successfully “replicated”, currently of considerable concern to the research community. To date, there remains no consensus on the precise meaning of replication in a statistical sense. The proposal of [Held \(2020\)](#) (see also [Held et al., 2022](#)) was to challenge the original finding using AnCred, as described in Section 2.1, and then evaluate the plausibility of the resulting prior using a prior-predictive check on the data from a replication study. A similar procedure but using AnCred based on Bayes factors as in Section 3 was proposed in [Pawel and Held \(2022\)](#). Reverse-Bayes inference seems to fit naturally into this setting as it provides a formal framework to challenge and substantiate scientific findings.

Apart from using data from a replication study, there are also other possible extensions of AnCred: We proposed to derive Reverse-Bayes priors using posterior tail probabilities (or credible intervals) or Bayes factors as measures of evidence, but also other measures such as relative belief ratios ([Evans, 2015](#)) could be used. When testing point null hypotheses, relative belief ratios are equivalent to Bayes factors due to the Savage-Dickey density ratio ([Evans, 2015](#), p. 98). Therefore, determining the sceptical prior variance through fixing the resulting Bayes factor is equivalent to fixing the resulting relative belief ratio. However, there is no connection to relative belief in prior-data conflict assessment based on the Bayes factor contrasting the sceptical to the optimistic prior since both are composite. Further research is needed on Reverse-Bayes procedures in the relative belief framework, candidate methods for prior-data conflict assessment are prior to posterior divergence ([Nott et al., 2020](#)) and prior expansions ([Nott et al., 2021](#)) as these methods have an interpretation in terms of relative beliefs. Moreover, we either used prior-predictive checks ([Box, 1980](#); [Evans and Moshonov, 2006](#)) or Bayes-factors ([Jeffreys, 1961](#); [Kass and Raftery, 1995](#)) for the formal evaluation of the plausibility of the priors derived through Reverse-Bayes. Other methods could be used for this purpose, for example, Bayesian measures of surprise ([Bayarri and Morales, 2003](#)). Furthermore, AnCred in its current state is derived assuming a normal likelihood for the effect

---

estimate  $\hat{\theta}$ . This is the same framework as in standard meta-analysis and provides a good approximation for studies with reasonable sample size (Carlin, 1992). For the comparison of binomial outcomes with small counts, the normal approximation of the log odds ratio could be improved with a Yates continuity correction (Spiegelhalter et al., 2004, section 2.4.1) or replaced with the exact profile likelihood of the log odds ratio (Held and Sabanés Bové, 2014, section 5.3), see also Section 4 in Pawel and Held (2022) which shows AnCred with Bayes factors using either a non-central  $t$  or a Binomial likelihood. Likewise, the conjugate normal prior could be replaced by a more robust prior distribution such as a mixture of normals (as considered in Section 2.1), a double-exponential, or a Student  $t$ -distribution (Pericchi and Smith, 1992). For example, Fúquene et al. (2009) investigate the use of robust priors in an application to binomial data from a randomized controlled trial. In general, any distribution from the location-scale family can be used, whereby the scale parameter takes over the role of the sceptical prior standard deviation, while the location parameter is fixed to the null value.

## 5.2 Conclusions

The inferential advantages of Bayesian methods are increasingly recognised within the statistical community. However, among the majority of working researchers they have failed to make any serious headway, and retain a reputation for complex and “controversial”. We have outlined how an idea that began with Jack Good’s proposal for resolving the “Problem of Priors” over 70 years ago (Good, 1950) has experienced a renaissance over recent years. The basic idea is to invert Bayes’ theorem: a specified posterior is combined with the data to obtain the Reverse-Bayes prior, which is then used for further inference. This approach is useful in situations where it is difficult to decide what constitutes a reasonable prior, but easy to specify the posterior which would lead to a particular decision. A subsequent prior-to-data conversion (Greenland, 2006) helps to assess the weight of the Reverse-Bayes prior in relation to the actual data.

We have shown that the Reverse-Bayes methodology is useful to extract more insights from the results typically reported in a meta-analysis. It facilitates the computation of prior-predictive checks for conflict diagnostics (Presanis et al., 2013) and has been shown capable of addressing many common inferential challenges, including assessing the credibility of scientific findings (Spiegelhalter et al., 2004; Greenland, 2011), making sense of “out of the blue” discoveries with no prior support (Matthews, 2018; Held, 2019a), estimating the probability of successful replications (Held, 2019a, 2020), and extracting more insight from standard  $p$ -values while reducing the risk of misinterpretation (Held, 2013; Colquhoun, 2017, 2019). The appeal of Reverse-Bayes techniques has recently been widened by the development of inferential methods using both posterior probabilities and Bayes Factors (Carlin and Louis, 1996; Pawel and Held, 2022).

These developments come at a crucial time for the role of statistical methods in research. Despite the many serious – and now well-publicised – inadequacies of NHST (Wasserstein and Lazar, 2016), the research community has shown itself to be remarkably reluctant to abandon NHST. Techniques based on the Reverse-Bayes methodology of the kind described

---

in this review could encourage the wider use of Bayesian inference by researchers. As such, we believe they can play a key role in the scientific enterprise of the 21<sup>th</sup> century.

## Software and Data Availability

All analyses were performed in the R programming language version 4.2.2 (R Core Team, 2022). Minimum Bayes factors were computed using the package pCalibrate (Held and Ott, 2018). The package metafor (Viechtbauer, 2010) was used for meta-analysis and forest plots. Data and code to reproduce all analyses are available at <https://gitlab.uzh.ch/samuel.pawel/Reverse-Bayes-Code>.

## Acknowledgments

We are grateful to Sander Greenland for helpful comments on a previous version of this article. We also acknowledge detailed comments by an Associate Editor and two referees, that helped to improve the manuscript.

## A Mean of the advocacy prior

Suppose that the estimate  $\hat{\theta}$  is not significant at level  $\alpha$ , so  $z^2/z_{\alpha/2}^2 < 1$ . With  $U, L = \hat{\theta} \pm z_{\alpha/2} \sigma$  we have  $U + L = 2\hat{\theta}$ ,  $UL = \hat{\theta}^2 - z_{\alpha/2}^2 \sigma^2$  and  $U - L = 2z_{\alpha/2} \sigma$ .

We therefore obtain with (11):

$$\mu = \frac{AL}{2} = -\frac{2\hat{\theta}}{2(\hat{\theta}^2 - z_{\alpha/2}^2 \sigma^2)} \frac{(2z_{\alpha/2} \sigma)^2}{2} = \frac{2\hat{\theta} z_{\alpha/2}^2 \sigma^2}{z_{\alpha/2}^2 \sigma^2 - \hat{\theta}^2} = \frac{2\hat{\theta}}{1 - z^2/z_{\alpha/2}^2}.$$

Dividing by the effect estimate  $\hat{\theta}$  leads to the relative mean  $f = \mu/\hat{\theta}$  as in (12). The advocacy standard deviation is  $\tau = AL/(2z_{\alpha/2}) = \mu/z_{\alpha/2}$  and the coefficient of variation is therefore  $CV = \tau/\mu = z_{\alpha/2}^{-1}$ .

## B Bayes factor for intrinsic credibility

Intrinsic credibility at level  $\gamma$  is established when

$$BF_{12} \leq BF_{01} = \gamma \tag{26}$$

and we are interested in the Bayes factor for intrinsic credibility  $BF_{IC}$  which is the smallest level  $\gamma \in (0, 1]$  where (26) holds. The  $BF_{IC}$  is therefore a special case of the sceptical Bayes

factor from Pawel and Held (2022) where the same data is used in both Bayes factors (instead of the data from a replication study for  $\text{BF}_{12}$ ). It is hence given by

$$\text{BF}_{\text{IC}} = \begin{cases} \sqrt{-z^2/k} \cdot \exp\{-(z^2 + k)/2\} & \text{if } |z| \geq \sqrt{\{-2W(-e^{-1}/\sqrt{2})\}} \\ \min\text{BF}_{01} & \text{if } \sqrt{\log 2} \leq |z| < \sqrt{\{-2W(-e^{-1}/\sqrt{2})\}} \\ \text{undefined} & \text{if } |z| < \sqrt{\log 2} \end{cases} \quad (27)$$

with  $k = W(-z^2 \exp\{-z^2/2\}/\sqrt{2})$  and  $W(\cdot)$  the branch of the Lambert  $W$  function that satisfies  $W(y) \leq -1$  for  $y \in [-e^{-1}, 0)$ . If  $|z| \geq \sqrt{-2W(-e^{-1}/\sqrt{2})} \approx 2.04$ ,  $\text{BF}_{\text{IC}}$  is located at the intersection between  $\text{BF}_{12}$  and  $\text{BF}_{01}$  in the relative prior variance  $g$ , so equation (7) from Pawel and Held (2022) can be used. For  $|z| \geq \sqrt{\log 2} \approx 0.83$ ,  $\text{BF}_{12}$  remains below  $\text{BF}_{01}$  for all  $g$  and hence  $\text{BF}_{\text{IC}}$  is given by  $\min\text{BF}_{01}$ , the minimum of  $\text{BF}_{01}$  from (16). Finally, when  $|z| < \sqrt{\log 2}$ , equation (26) cannot be satisfied for any valid sufficiently sceptical relative prior variance  $g$ , hence the  $\text{BF}_{\text{IC}}$  is undefined.

## Bibliography

- Bayarri, M. and Morales, J. (2003). Bayesian measures of surprise for outlier detection. *Journal of Statistical Planning and Inference*, 111(1-2):3–22. doi:[10.1016/s0378-3758\(02\)00282-3](https://doi.org/10.1016/s0378-3758(02)00282-3).
- 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](https://doi.org/10.1038/s41562-017-0189-z).
- 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](https://doi.org/10.2307/2289131).
- Best, N., Price, R. G., Pouliquen, I. J., and Keene, O. N. (2021). Assessing efficacy in important subgroups in confirmatory trials: An example using Bayesian dynamic borrowing. *Pharmaceutical Statistics*, 20(3):551–562. doi:[10.1002/pst.2093](https://doi.org/10.1002/pst.2093).
- Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *Journal of the Royal Statistical Society: Series A (General)*, 143:383–430. doi:[10.2307/2982063](https://doi.org/10.2307/2982063).
- Carlin, B. P. and Louis, T. A. (1996). Identifying prior distributions that produce specific decisions, with application to monitoring clinical trials. In Berry, D., Chaloner, K., and Geweke, J., editors, *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*, pages 493–503. Wiley, New York.
- Carlin, J. B. (1992). Meta-analysis for  $2 \times 2$  tables: A Bayesian approach. *Statistics in Medicine*, 11(2):141–158. doi:[10.1002/sim.4780110202](https://doi.org/10.1002/sim.4780110202).
- Colquhoun, D. (2017). The reproducibility of research and the misinterpretation of  $p$ -values. *Royal Society Open Science*, 4(12). doi:[10.1098/rsos.171085](https://doi.org/10.1098/rsos.171085).

- 
- Colquhoun, D. (2019). The false positive risk: A proposal concerning what to do about  $p$ -values. *The American Statistician*, 73(sup1):192–201. doi:[10.1080/00031305.2018.1529622](https://doi.org/10.1080/00031305.2018.1529622).
- 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](https://doi.org/10.7758/9781610448864).
- Copas, J. and Eguchi, S. (2005). Local model uncertainty and incomplete-data bias (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):459–513. doi:[10.1111/j.1467-9868.2005.00512.x](https://doi.org/10.1111/j.1467-9868.2005.00512.x).
- 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](https://doi.org/10.1007/bf02124750).
- Cox, D. R. (2006). *Principles of Statistical Inference*. Cambridge University Press, Cambridge.
- Cunen, C. and Hjort, N. L. (2021). Combining information across diverse sources: The II-CC-FF paradigm. *Scandinavian Journal of Statistics*. doi:[10.1111/sjos.12530](https://doi.org/10.1111/sjos.12530).
- Dequin, P.-F., Heming, N., Meziani, F., Plantefève, G., Voiriot, G., Badié, J., François, B., Aubron, C., Ricard, J.-D., Ehrmann, S., Jouan, Y., Guillon, A., Leclerc, M., Coffre, C., Bourgoin, H., Lengellé, C., Caille-Fénérol, C., Tavernier, E., Zohar, S., Giraudeau, B., Annane, D., and and, A. L. G. (2020). Effect of hydrocortisone on 21-day mortality or respiratory support among critically ill patients with COVID-19. *JAMA*, 324(13):1298. doi:[10.1001/jama.2020.16761](https://doi.org/10.1001/jama.2020.16761).
- Dias, S., Welton, N. J., Caldwell, D. M., and Ades, A. E. (2010). Checking consistency in mixed treatment comparison meta-analysis. *Statistics in Medicine*, 29(7-8):932–944. doi:[10.1002/sim.3767](https://doi.org/10.1002/sim.3767).
- 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](https://doi.org/10.1037/h0044139).
- Evans, M. (2015). *Measuring statistical evidence using relative belief*. Number 144 in Monographs on statistics and applied probability. CRC Press, Taylor & Francis Group, Boca Raton.
- Evans, M. and Moshonov, H. (2006). Checking for prior-data conflict. *Bayesian Analysis*, 1(4):893–914. doi:[10.1214/06-ba129](https://doi.org/10.1214/06-ba129).
- Ferkingstad, E., Held, L., and Rue, H. (2017). Fast and accurate Bayesian model criticism and conflict diagnostics using R-INLA. *Stat*, 6(1):331–344. doi:[10.1002/sta4.163](https://doi.org/10.1002/sta4.163).
- Fúquene, J. A., Cook, J. D., and Pericchi, L. R. (2009). A case for robust Bayesian priors with applications to clinical trials. *Bayesian Analysis*, 4(4):817–846. doi:[10.1214/09-BA431](https://doi.org/10.1214/09-BA431).
- Gelman, A. and Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102(6):460–465. doi:[10.1511/2014.111.460](https://doi.org/10.1511/2014.111.460).
- 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. doi:[10.1080/01621459.1958.10501480](https://doi.org/10.1080/01621459.1958.10501480).
-



- 
- Good, I. J. (1983). *Good Thinking: The Foundations of Probability and Its Applications*. University of Minnesota Press, Minneapolis.
- Goudie, R. J. B., Presanis, A. M., Lunn, D., Angelis, D. D., and Wernisch, L. (2019). Joining and Splitting Models with Markov Melding. *Bayesian Analysis*, 14(1):81–109. doi:[10.1214/18-BA1104](https://doi.org/10.1214/18-BA1104).
- Green, P., Łatuszyński, K., Pereyra, M., and Robert, C. (2015). Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing*, 25(6):835–862. doi:[10.1007/s11222-015-9574-5](https://doi.org/10.1007/s11222-015-9574-5).
- Greenland, S. (2006). Bayesian perspectives for epidemiological research: I. foundations and basic methods. *International Journal of Epidemiology*, 35:765–775. doi:[10.1093/ije/dyi312](https://doi.org/10.1093/ije/dyi312).
- Greenland, S. (2011). Null misinterpretation in statistical testing and its impact on health risk assessment. *Preventive Medicine*, 53:225–228. doi:[10.1016/j.ypmed.2011.08.010](https://doi.org/10.1016/j.ypmed.2011.08.010).
- Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31(4):337–350. doi:[10.1007/s10654-016-0149-3](https://doi.org/10.1007/s10654-016-0149-3).
- Held, L. (2013). Reverse-Bayes analysis of two common misinterpretations of significance test. *Clinical Trials*, 10:236–242. doi:[10.1177/1740774512467707](https://doi.org/10.1177/1740774512467707).
- 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](https://doi.org/10.1098/rsos.181534).
- Held, L. (2019b). On the Bayesian interpretation of the harmonic mean  $p$ -value. *Proceedings of the National Academy of Sciences*, 116(13):5855–5856. doi:[10.1073/pnas.1900671116](https://doi.org/10.1073/pnas.1900671116).
- 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](https://doi.org/10.1111/rssa.12493).
- Held, L., Micheloud, C., and Pawel, S. (2022). The assessment of replication success based on relative effect size. *The Annals of Applied Statistics*, 16(2):706–720. doi:[10.1214/21-aos1502](https://doi.org/10.1214/21-aos1502).
- Held, L. and Ott, M. (2018). On  $p$ -values and Bayes factors. *Annual Review of Statistics and Its Application*, 5(1):393–419. doi:[10.1146/annurev-statistics-031017-100307](https://doi.org/10.1146/annurev-statistics-031017-100307).
- Held, L. and Sabanés Bové, D. (2014). *Applied Statistical Inference*. Springer-Verlag Berlin Heidelberg. doi:[10.1007/978-3-642-37887-4](https://doi.org/10.1007/978-3-642-37887-4).
- Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK New York, NY. doi:[10.1017/cbo9780511790423](https://doi.org/10.1017/cbo9780511790423).
- Jeffreys, H. (1961). *Theory of Probability*. Oxford: Clarendon Press, third edition.
- Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):143–170. doi:[10.1111/j.1467-9868.2009.00730.x](https://doi.org/10.1111/j.1467-9868.2009.00730.x).

- 
- 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](https://doi.org/10.1080/01621459.1995.10476572).
- Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90(431):928–934. doi:[10.1080/01621459.1995.10476592](https://doi.org/10.1080/01621459.1995.10476592).
- Killeen, P. (2006). An alternative to null-hypothesis significance tests. *Psychological Science*, 16:345–353. doi:[10.1111/j.0956-7976.2005.01538.x](https://doi.org/10.1111/j.0956-7976.2005.01538.x).
- Matthews, R. A. J. (2001a). Methods for assessing the credibility of clinical trial outcomes. *Drug Information Journal*, 35:1469–1478. doi:[10.1177/009286150103500442](https://doi.org/10.1177/009286150103500442).
- Matthews, R. A. J. (2001b). Why should clinicians care about Bayesian methods? *Journal of Statistical Planning and Inference*, 94:43–71. doi:[10.1016/S0378-3758\(00\)00232-9](https://doi.org/10.1016/S0378-3758(00)00232-9).
- Matthews, R. A. J. (2017). The ASA’s  $p$ -value statement, one year on. *Significance*, 14(2):38–40. doi:[10.1080/00031305.2016.1154108](https://doi.org/10.1080/00031305.2016.1154108).
- 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](https://doi.org/10.1098/rsos.171047).
- McElreath, R. (2018). *Statistical Rethinking*. Chapman and Hall/CRC. doi:[10.1201/9781315372495](https://doi.org/10.1201/9781315372495).
- McGrayne, S. B. (2011). *The Theory That Would Not Die*. Yale University Press, New Haven, CT.
- Nott, D. J., Seah, M., Al-Labadi, L., Evans, M., Ng, H. K., and Englert, B.-G. (2021). Using prior expansions for prior-data conflict checking. *Bayesian Analysis*, 16(1). doi:[10.1214/20-ba1204](https://doi.org/10.1214/20-ba1204).
- Nott, D. J., Wang, X., Evans, M., and Englert, B.-G. (2020). Checking for prior-data conflict using prior-to-posterior divergences. *Statistical Science*, 35(2). doi:[10.1214/19-sts731](https://doi.org/10.1214/19-sts731).
- O’Hagan, A. and Forster, J. (2004). *Kendall’s Advanced Theory of Statistic 2B*. Wiley, second edition.
- Ortega, H. G., Liu, M. C., Pavord, I. D., Brusselle, G. G., FitzGerald, J. M., Chetta, A., Humbert, M., Katz, L. E., Keene, O. N., Yancey, S. W., and Chanez, P. (2014). Mepolizumab treatment in patients with severe eosinophilic asthma. *New England Journal of Medicine*, 371(13):1198–1207. doi:[10.1056/NEJMoa1403290](https://doi.org/10.1056/NEJMoa1403290).
- Pawel, S. and Held, L. (2022). The sceptical Bayes factor for the assessment of replication success. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. doi:[10.1111/rssb.12491](https://doi.org/10.1111/rssb.12491).
- Pericchi, L. R. and Smith, A. F. M. (1992). Exact and approximate posterior moments for a normal location parameter. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(3):793–804. doi:[10.1111/j.2517-6161.1992.tb01452.x](https://doi.org/10.1111/j.2517-6161.1992.tb01452.x).
- Presanis, A. M., Ohlssen, D., Spiegelhalter, D. J., and Angelis, D. D. (2013). Conflict diagnostics in directed acyclic graphs, with applications in Bayesian evidence synthesis. *Statistical Science*, 28(3):376–397. doi:[10.1214/13-sts426](https://doi.org/10.1214/13-sts426).



- 
- 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/>.
- RECOVERY Collaborative Group (2020). Dexamethasone in hospitalized patients with covid-19 – preliminary report. *New England Journal of Medicine*. doi:[10.1056/nejmoa2021436](https://doi.org/10.1056/nejmoa2021436).
- REMAP-CAP Investigators (2020). Effect of hydrocortisone on mortality and organ support in patients with severe COVID-19. *JAMA*, 324(13):1317. doi:[10.1001/jama.2020.17022](https://doi.org/10.1001/jama.2020.17022).
- Robert, C. P. (2014). On the Jeffreys-Lindley paradox. *Philosophy of Science*, 81(2):216–232. doi:[10.1086/675729](https://doi.org/10.1086/675729).
- Rosenberg, M. S. (2005). The file-drawer problem revisited: A general weighted method for calculating fails-safe numbers in meta-analysis. *Evolution*, 59(2):464–468. doi:[10.1111/j.0014-3820.2005.tb01004.x](https://doi.org/10.1111/j.0014-3820.2005.tb01004.x).
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3):638–641. doi:[10.1037/0033-2909.86.3.638](https://doi.org/10.1037/0033-2909.86.3.638).
- Sellke, T., Bayarri, M. J., and Berger, J. O. (2001). Calibration of  $p$  values for testing precise null hypotheses. *The American Statistician*, 55:62–71. doi:[10.1198/000313001300339950](https://doi.org/10.1198/000313001300339950).
- Spiegelhalter, D. J., Abrams, R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. New York: Wiley.
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3). doi:[10.18637/jss.v036.i03](https://doi.org/10.18637/jss.v036.i03).
- Wagenmakers, E.-J., Lee, M., Lodewyckx, T., and Iverson, G. J. (2008). *Bayesian Versus Frequentist Inference*, pages 181–207. Springer New York, New York, NY.
- Wasserstein, R. L. and Lazar, N. A. (2016). The ASA’s statement on  $p$ -values: Context, process, and purpose. *The American Statistician*, 70(2):129–133. doi:[10.1080/00031305.2016.1154108](https://doi.org/10.1080/00031305.2016.1154108).
- Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. (2019). Moving to a world beyond “ $p < 0.05$ ”. *The American Statistician*, 73(sup1):1–19. doi:[10.1080/00031305.2019.1583913](https://doi.org/10.1080/00031305.2019.1583913).
- WHO REACT Working Group (2020). Association between administration of systemic corticosteroids and mortality among critically ill patients with COVID-19: A meta-analysis. *JAMA*, 324(13):1330–1341. doi:[10.1001/jama.2020.17023](https://doi.org/10.1001/jama.2020.17023).



---

**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>.

---



---

## Abstract

We examine the concept of Bayesian Additional Evidence (BAE) recently proposed by Sondhi et al. We derive simple closed-form expressions for BAE and compare its properties with other methods for assessing findings in the light of new evidence. We find that while BAE is easy to apply, it lacks both a compelling rationale and clarity of use needed for reliable decision-making.

**Key words:** Advocacy prior, Analysis of Credibility, Bayesian Additional Evidence, Reverse-Bayes

## 1 Introduction

We read with great interest the article by [Sondhi et al. \(2021\)](#), which introduces the concept of *Bayesian Additional Evidence* (BAE). The authors use a reverse-Bayes argument to define BAE, and apply it to the important issue of how new evidence affects the overall credibility of an existing finding. As they state, BAE is thus closely related to another reverse-Bayes approach known as *Analysis of Credibility* (AnCred) proposed by [Matthews \(2018\)](#); see also the recent review of Reverse-Bayes methods ([Held et al., 2022](#)). In what follows, we comment on the similarities and differences of the two approaches and their inferential consequences. We find that decision making based on the BAE approach is limited by the restrictive assumption that the additional evidence must have equal or smaller variance than the variance of the observed data.

## 2 Bayesian additional evidence

We begin by showing that fortunately – and contrary to the statement by Sondhi et al. on page 4 of their article – there is a closed-form solution for what they term the BAE “tipping point”, which is key to their approach.

Assume, as per Sondhi et al., that both the likelihood of an effect estimate  $\hat{\theta}$  (the “data”) and the prior of the underlying effect size  $\theta$  are represented by normal distributions  $\hat{\theta} | \theta \sim N(\theta, \sigma^2)$  and  $\theta \sim N(\mu, \tau^2)$ , with the latter evidence coming either from pre-existing insight/studies or from a subsequent replication. Bayes’s Theorem then implies a posterior distribution  $\theta | \hat{\theta} \sim N(\mu_p, \tau_p^2)$  whose mean and variance satisfy

$$\frac{\mu_p}{\tau_p^2} = \frac{\hat{\theta}}{\sigma^2} + \frac{\mu}{\tau^2} \quad \text{and} \quad \frac{1}{\tau_p^2} = \frac{1}{\sigma^2} + \frac{1}{\tau^2}$$

Sondhi et al. further assume that  $\tau^2 = \sigma^2$ , that is, the prior variance  $\tau^2$  is equal to the data variance  $\sigma^2$  which itself is equal to the squared (known) standard error  $\sigma$  of the effect estimate

---

$\hat{\theta}$ . It then follows that the posterior mean is the mean of the data and the prior mean, and that the posterior variance is half the data variance

$$\mu_p = \frac{\hat{\theta} + \mu}{2} \quad \text{and} \quad \tau_p^2 = \frac{\sigma^2}{2} \quad (1)$$

The BAE “tipping point” is then defined as the least extreme prior mean that results in a posterior credible interval which excludes the null value. If the substantive hypothesis is for positive effect estimates (e.g.,  $\log(\text{HR}) > 0$ ) the BAE is the prior mean which leads to the lower limit  $L_p$  of the  $100(1 - \alpha)\%$  posterior credible interval being zero

$$L_p = \mu_p - z_{\alpha/2} \tau_p = 0 \quad (2)$$

while for negative effect estimates the upper limit  $U_p$  is fixed to zero

$$U_p = \mu_p + z_{\alpha/2} \tau_p = 0 \quad (3)$$

with  $z_{\alpha/2}$  the  $1 - \alpha/2$  quantile of the standard normal distribution. Combining Eq. (1) with Eq. (2), respectively Eq. (3), leads to

$$\text{BAE} = \text{sign}(\hat{\theta}) \sqrt{2} z_{\alpha/2} \sigma - \hat{\theta} \quad (4)$$

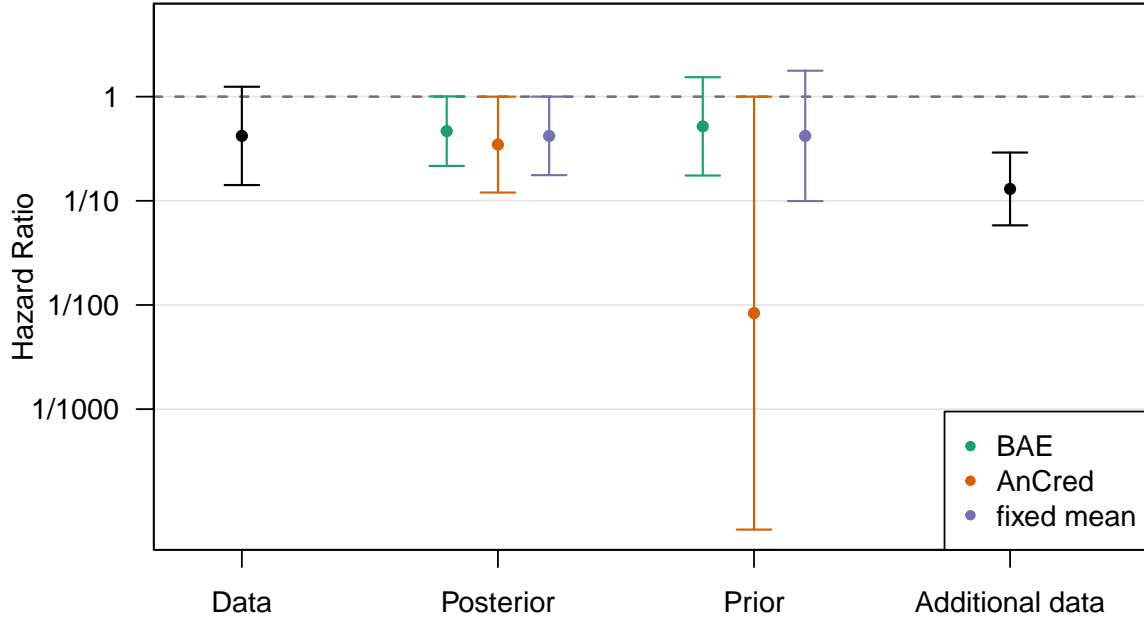
where  $\text{sign}(\hat{\theta}) = 1$  when  $\hat{\theta} > 0$  and  $\text{sign}(\hat{\theta}) = -1$  otherwise. Re-written in terms of the upper and lower  $100(1 - \alpha)\%$  confidence interval (CI) limits  $U$  and  $L$  of the effect estimate  $\hat{\theta}$  we obtain

$$\text{BAE} = \frac{\text{sign}(\hat{\theta}) \sqrt{2}(U - L) - (U + L)}{2} \quad (5)$$

We see from Eq. (4) that Sondhi et al.’s proposal has the intuitive property that as the study becomes more convincing (through larger effect sizes  $|\hat{\theta}|$  and/or smaller standard errors  $\sigma$ ), the BAE will decrease (increase) for positive (negative)  $\hat{\theta}$ , indicating that less additional evidence is needed to push a non-significant study towards credibility. Eq. (4) and Eq. (5) also hold for significant studies but the BAE then represents the mean of a “sceptical” prior which renders the study non-significant.

These closed-form solutions greatly simplify the use of the BAE methodology. For example, Sondhi et al. use a comparison of monoclonals to show how it identifies additional evidence which, when combined with a non-significant finding, leads to overall credibility. The trial estimated the hazard ratio of the bevacizumab+chemo patients compared to the cetuximab+chemo patients as  $\text{HR} = 0.42$  (95% CI: 0.14 to 1.23), a non-significant finding with  $p = 0.11$ . Expressed as  $\log(\text{HR})$ , we have  $L = -1.97$  and  $U = 0.21$ . We use Eq. (5) and find that on log hazard ratio scale  $\text{BAE} = -0.66$  equivalent to an HR of 0.52. Figure 1 shows the corresponding prior mean with 95% prior credible interval.

Thus additional evidence in the form of prior insight or a subsequent replication supporting an HR at least as impressive as this (i.e., an  $\text{HR} < 0.52$  in this case), and a CI at least as tight as that of the original study will render this non-significant result credible at the 95% level. Sondhi et al. cite prior evidence from [Innocenti et al. \(2019\)](#) who found an  $\text{HR} = 0.13$  (95% CI: 0.06 to 0.30) which meets both criteria set by the BAE, and renders the original study credible.



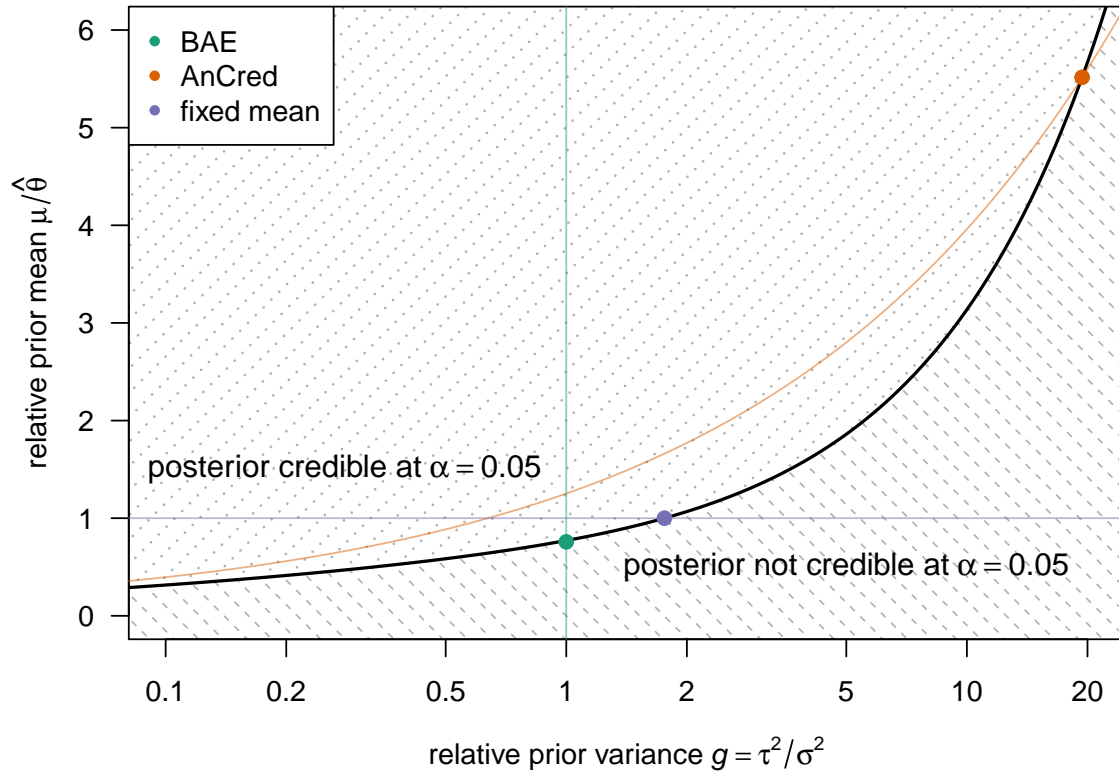
**Figure 1:** Comparison of BAE, AnCred, and fixed mean 95% prior and posterior credible intervals for the data from [Sondhi et al. \(2021\)](#). Additional data from [Innocenti et al. \(2019\)](#) are also shown.

### 3 Alternatives approaches

In order to get a unique solution for the BAE, Sondhi et al. make the assumption that the prior variance equals the data variance, but also other possibilities exist. An alternative rationale would be to set the mean of the additional evidence, rather than variance, to that of the original finding (i.e.,  $\mu = \hat{\theta}$ ), and determining the prior variance  $\tau^2$  such that the posterior credible interval includes the null value. Under this approach, the prior variance is given by

$$\tau^2 = \frac{\sigma^2}{z_{\alpha/2}^2 / z^2 - 1}$$

with  $z = \hat{\theta} / \sigma$ . The resulting prior represents a study with identical effect estimate but different precision compared to the observed one. As the observed study becomes more convincing (with larger effect estimates  $|\hat{\theta}|$  and/or smaller standard errors  $\sigma$ ), the prior will become more diffuse, so less additional evidence is needed to render the finding credible. We see in Figure 1 that prior and posterior are similar to BAE for the clinical trial data from Sondhi et al.



**Figure 2:** Relative prior mean vs. relative prior variance for the data from Sondhi et al. The dashed region represents parameter values, which do not lead to posterior credibility, whereas values in the dotted region lead to posterior credibility (at  $\alpha = 5\%$ ). The colored lines indicate the parameters which fulfil the side-constraints of the respective method.

Figure 1 also illustrates that the BAE and the fixed mean approach can lead to priors which support effect sizes opposing that of the original finding. This is not possible with the AnCred advocacy prior whose prior credible interval is fixed to the null value so that the prior adheres to the Principle of Fairminded Advocacy (Matthews, 2018). Held et al. (2022) showed that this constraint is equivalent to fixing the coefficient of variation from the prior to  $\tau/\mu = z_{\alpha/2}^{-1}$ . Hence, its mean and variance are given by

$$\mu = \frac{2\hat{\theta}}{1 - z^2/z_{\alpha/2}^2} \quad \text{and} \quad \tau^2 = \frac{\mu^2}{z_{\alpha/2}^2}.$$

We see that – as with the fixed mean approach – the AnCred prior becomes more diffuse for increasingly convincing studies. However, at the same time the prior mean also increases (decreases) for positive (negative) effect estimates, so that only effect sizes in the correct direction are supported.



Figure 1 shows that the AnCred advocacy prior credible interval is far wider compared to the other approaches. Perhaps this observation led Sondhi et al. to state that AnCred is harder to interpret than BAE, and that it can lead to prior intervals “wide enough to effectively contain any effect size, which is unhelpful for decision making”. We would argue that broad priors are a valuable diagnostic of when little additional evidence is needed to achieve posterior credibility, as it is the case with the example Sondhi et al. consider. Moreover, we would argue that AnCred priors are very helpful in decision making since any additional evidence whose confidence interval is contained in the AnCred prior credible interval will *necessarily* lead to posterior credibility when combined with the observed data (Held et al., 2022). In contrast, the BAE approach requires decision makers to keep in mind the variance of the additional evidence, since only additional evidence with a point estimate that is more extreme as the BAE and with confidence interval at least as tight as the observed confidence interval from the study is guaranteed to lead to posterior credibility. Assume, for example, the additional data from Innocenti et al. had been more impressive, say,  $HR = 0.05$ , with a 95% CI from 0.015 to 0.16. Intuition suggests, and direct calculation confirms, that this would be even more capable of making the original finding credible. However, this would not be clear to a decision maker using the BAE approach as currently formulated, as the confidence interval is wider than the one of the observed study (on the log scale).

While Sondhi et al. acknowledge the dependence of the BAE on the choice of the prior variance, they do not give clear guidance on when it should be set to a value different from the observed data variance. Fortunately, when the prior and data variances differ, there is again a closed form solution for the BAE “tipping point”

$$BAE(g) = \text{sign}(\hat{\theta}) \sqrt{g(1+g)} z_{\alpha/2} \sigma - g \hat{\theta} \quad (6)$$

with relative prior variance  $g = \tau^2/\sigma^2$ . We see from Figure 2 that Eq. (6) substantially depends on the chosen prior variance and that the BAE based on  $g = 1$  only captures a limited range of priors which lead to posterior credibility. Unfortunately, Sondhi et al. do not give a clear rationale for the default choice of  $g = 1$ . It may therefore be more helpful for decision makers to base their decision on the more principled AnCred advocacy prior or on a visualisation of the prior parameter space as in Figure 2.

## 4 Conclusion

In summary, we welcome BAE as an interesting application of reverse-Bayes methods, and we hope our derivation of closed-form solutions will encourage further research. However, as currently formulated BAE lacks both a clear rationale for the constraints on which it is based, and a sufficiently detailed explanation allowing reliable decision-making.

## Software and data

Summary statistics on the case study were taken from Sondhi et al. (2021). The code to reproduce our analyses is available at <https://github.com/SamCH93/BAEcomment>. A snapshot

---

of the Git repository at the time of writing is archived at <https://doi.org/10.5281/zenodo.XXXXXX>.

## Acknowledgments

We thank two anonymous referees and the third referee Riko Kelter for their helpful suggestions. Financial support from the Swiss National Science Foundation (Project #189295) is gratefully acknowledged. The funder had no role in study design, data collection, data analysis, data interpretation, decision to publish, or preparation of the manuscript.

## Bibliography

- Held, L., Matthews, R., Ott, M., and Pawel, S. (2022). Reverse-Bayes methods for evidence assessment and research synthesis. *Research Synthesis Methods*, 13(3):295–314. doi:[10.1002/jrsm.1538](https://doi.org/10.1002/jrsm.1538).
- Innocenti, F., Ou, F.-S., Qu, X., Zemla, T. J., Niedzwiecki, D., Tam, R., Mahajan, S., Goldberg, R. M., Bertagnolli, M. M., Blanke, C. D., Sanoff, H., Atkins, J., Polite, B., Venook, A. P., Lenz, H.-J., and Kabbarah, O. (2019). Mutational analysis of patients with colorectal cancer in CALGB/SWOG 80405 identifies new roles of microsatellite instability and tumor mutational burden for patient outcome. *Journal of Clinical Oncology*, 37(14):1217–1227. doi:[10.1200/jco.18.01798](https://doi.org/10.1200/jco.18.01798).
- 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](https://doi.org/10.1098/rsos.171047).
- 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](https://doi.org/10.1186/s12874-021-01432-5).

---

## **Pitfalls and Potentials in Simulation Studies**

*Samuel Pawel, Lucas Kook, Kelly Reeve*

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

---

