Multisite generalizations of replicability measures

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

Working paper version September 26, 2023

Abstract

Multisite replication studies aim to repeat an original study in order to assess whether similar results can be obtained with new data across different study sites. While a variety of statistical methods have been proposed for the analysis of singlesite replication studies, fewer methods are available for the multisite setting. Here we discuss several extensions of singlesite methods that have not yet been generalized to the multisite setting, both frequentist (the two-trials rule) and Bayesian (the sceptical p-value, the replication Bayes factor, and the sceptical Bayes factor). A key challenge is to account for between-replication heterogeneity, and we present different approaches for doing so. These generalizations provide analysts with a suite of methods for assessing different aspects of replicability. We illustrate their properties using data from several multisite replication projects.

Keywords: Bayes factor, heterogeneity, multivariate, sceptical p-value, two-trials rule

1 Introduction

A fundamental aspect of the credibility of a research finding is whether it is *replicable*, that is, whether a similar finding can be obtained when a study is repeated with new subjects (National Academies of Sciences, Engineering, and Medicine, 2019). The "replication crisis" in the social and life sciences led to an increase in the conduct of replication studies, and several journals and funders now actively promote such studies (NWO, 2016; NSF, 2018; Nature Communications, 2022). While many replication projects have focused on *one-to-one* or *singlesite* replication studies – conducting a single replication study for a single original study (e.g., Open Science Collaboration, 2015; Camerer et al., 2016, 2018; Errington et al., 2021) – there has also been an increasing interest in *many-to-one* or *multisite* replication studies – conducting multiple replication studies for a single original study (for example, Klein et al., 2014; Ebersole et al., 2016; Klein et al., 2018; Wagenmakers et al., 2016; Protzko et al., 2020; Arroyo-Araujo et al., 2022). This study design allows not only to assess replicability of a finding, but also potential between-replication heterogeneity.

There is no universally agreed definition of "replicability" or "replication success". However, while in the case of singlesite replications, methodologists have proposed diverse approaches to quantify replicability (Verhagen and Wagenmakers, 2014; Simonsohn, 2015; Anderson and Maxwell, 2016; Patil et al., 2016; Bonett, 2020; Held, 2020; Pawel and Held, 2022, among others), fewer approaches have been proposed in the case of multisite replications. In practice, researchers have often used meta-analytic approaches to pool results from different replication studies and quantify their heterogeneity, but also "vote-counting" approaches have been used, e.g., counting how many of the individual replication p-values are smaller than some threshold. Most of the methodological contributions to date have focused on the extension of meta-analytic methods to the replication setting, mostly from a frequentist perspective. For example, Hedges and Schauer (2019) have proposed various tests for (non-)replicability based on testing for the absence/presence of between-replication heterogeneity, or Mathur and VanderWeele (2020) have proposed measures of consistency between original and replication effect sizes. The aim of our article is therefore to extend other measures of replicability that fall outside the meta-analytic realm considered previously and that have not yet been adapted to the multisite setting. In particular, we consider different generalizations of the two-trials rule (Section 4), the sceptical p-values (Section 5), and various Bayes factor methods (Section 6).

2 Running examples

Throughout this article, we will apply the developed methodology to data from the three multisite replication studies, some of them shown in Figure 1. Table 1 shows different measures of replicability applied to them, we will develop and discuss these measures throughout the article.

Facial feedback replications This multisite replication study by Wagenmakers et al. (2016) attempted to replicate the original study from Strack et al. (1988) which tested the facial feedback hypothesis. The original study found that participant gave higher funniness ratings to cartoons if they were smiling as opposed to showing discontent (estimated mean difference of 0.82 units on a 10-point Likert scale, with 95% confidence interval from -0.05 to 1.69). On the other hand, the pooled replication mean difference was very close to zero (estimated mean difference of -0.03 with 95% confidence interval from -0.11 to 0.16). In addition, the individual replication mean differences hardly showed any heterogeneity ($p_Q = 0.91$).

Moral credentials replications This multisite replication study by Ebersole et al. (2016) attempted to replicate the original study by Monin and Miller (2001) on "moral credentialling". The original study found that participants were more likely to indicate preference for hiring male candidates in an imagined hiring scenario if they were assigned to a credentialling condition where they had to indicate (dis)agreement with a sexist statement as opposed to a control condition (Fisher z-transformed correlation of 0.21 with 95% confidence interval from 0.1 to 0.32). The replication pooled effect estimate was slightly smaller (0.07 with 95% confidence interval from 0.03 to 0.11) and there was some heterogeneity among the individual replication effect estimates ($p_Q = 0.10$).

Prospective replication project Protzko et al. (2020) conducted a prospective replication project. Each of the four participating laboratories conducted original studies as well as replication studies of their own and the other original studies. Figure 1 shows original and replication studies from one of the 16 experiments, Table 1 shows summary statistics for all of them. The fast social desirability (FSD)

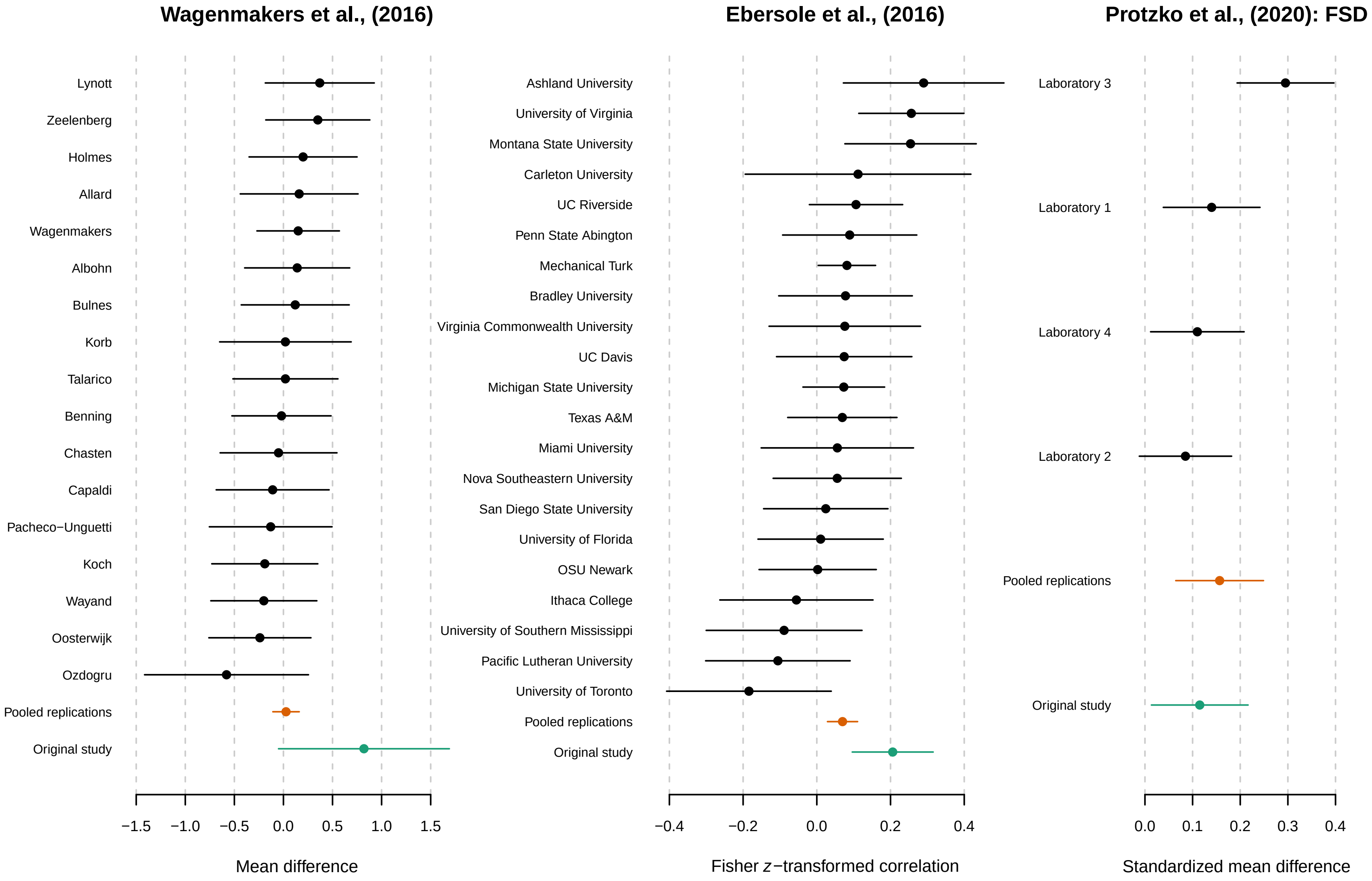

Figure 1: Forest plots of effect estimates from facial feedback studies (Wagenmakers et al., 2016), moral credentials studies (Ebersole et al., 2016) and fast social desirability (FSD) studies (Protzko et al., 2020).

Table 1: Replicability assessment of facial feedback replications (Wagenmakers et al., 2016), moral credentials replications (Ebersole et al., 2016), and replications from prospective replication project (Protzko et al., 2020). Shown are one-sided original p-value p_o, original and pooled replication effect estimates $\hat{\theta}_o$ and $\hat{\theta}_r$, REML estimated between-replication standard deviation $\hat{\tau}$, p-value from Q-test for between-replication heterogeneity p_Q, meta-analytic replication p-value p_r, sceptical p-value p_S (multivariate and pooling approach versions), default Bayes factors BF_{01} (using one-sided standard normal priors under the alternative), replication Bayes factor BF_R, and sceptical Bayes factor BF_S.

Study	p_o	$\hat{\theta}_o$ [95% CI]	$\hat{\theta}_r$ [95% CI]	$\hat{\tau}$ [95% CI]	p_Q	p_r	p_S (multi)	p_S (pool)	$\text{BF}_{01}(\hat{\theta}_o)$	$\text{BF}_{01}(\hat{\theta}_r)$	BF_R	BF_S
Facial feedback	0.032	0.82 [−0.05, 1.69]	0.03 [−0.11, 0.16]	0.00 [0.00, 0.16]	0.91	0.35	0.47	0.39	1/3.2	10	38	nonexistent
Moral credentials	0.0001	0.21 [0.10, 0.32]	0.07 [0.03, 0.11]	0.04 [0.00, 0.13]	0.10	0.0004	0.072	0.034	1/94	1/11	1/7	1/1.5
Redemption	0.074	−0.08 [−0.18, 0.03]	0.09 [−0.01, 0.20]	0.10 [0.04, 0.40]	0.003	0.96	0.84	0.87	3.7	50	8.3	nonexistent
Misreporting	0.35	0.02 [−0.08, 0.12]	0.03 [−0.03, 0.09]	0.04 [0.00, 0.23]	0.21	0.20	0.35	0.38	14	14	1.1	nonexistent
Prediction	0.064	0.08 [−0.02, 0.18]	0.15 [0.10, 0.20]	0.00 [0.00, 0.07]	0.89	< 0.0001	0.12	0.086	3.2	< 1/1000	< 1/1000	1/1.3
FSD	0.013	0.12 [0.01, 0.22]	0.16 [0.06, 0.25]	0.08 [0.01, 0.35]	0.017	0.0004	0.052	0.035	1/1.2	1/24	1/150	1/3.2
Minimal Groups	0.0008	0.15 [0.06, 0.25]	0.10 [0.06, 0.15]	0.00 [0.00, 0.07]	0.88	< 0.0001	0.049	0.018	1/15	1/310	< 1/1000	1/6.1
Labels	< 0.0001	0.20 [0.11, 0.30]	0.23 [0.09, 0.38]	0.14 [0.07, 0.54]	< 0.0001	0.0008	0.013	0.004	1/353	1/20	1/111	1/32
Referrals	< 0.0001	0.20 [0.11, 0.30]	0.22 [0.18, 0.26]	0.02 [0.00, 0.11]	0.46	< 0.0001	0.001	0.0005	1/435	< 1/1000	< 1/1000	1/154
Ads	< 0.0001	0.22 [0.11, 0.32]	0.22 [0.14, 0.30]	0.06 [0.00, 0.30]	0.061	< 0.0001	0.002	0.0009	1/576	< 1/1000	< 1/1000	1/102
Cookies	< 0.0001	0.24 [0.14, 0.33]	0.26 [0.21, 0.31]	0.00 [0.00, 0.12]	0.68	< 0.0001	0.0002	< 0.0001	< 1/1000	< 1/1000	< 1/1000	1/948
Tumor	< 0.0001	0.32 [0.21, 0.42]	0.34 [0.29, 0.40]	0.03 [0.00, 0.20]	0.20	< 0.0001	< 0.0001	< 0.0001	< 1/1000	< 1/1000	< 1/1000	< 1/1000
Ostracism	< 0.0001	0.35 [0.25, 0.46]	0.35 [0.28, 0.42]	0.05 [0.00, 0.26]	0.11	< 0.0001	< 0.0001	< 0.0001	< 1/1000	< 1/1000	< 1/1000	< 1/1000
Self-Control	< 0.0001	0.36 [0.26, 0.47]	0.28 [0.23, 0.32]	0.00 [0.00, 0.02]	0.97	< 0.0001	< 0.0001	< 0.0001	< 1/1000	< 1/1000	< 1/1000	< 1/1000
Misattribution	< 0.0001	0.46 [0.36, 0.56]	0.34 [0.29, 0.39]	0.00 [0.00, 0.17]	0.44	< 0.0001	< 0.0001	< 0.0001	< 1/1000	< 1/1000	< 1/1000	< 1/1000
Fairness	< 0.0001	0.47 [0.37, 0.56]	0.43 [0.36, 0.50]	0.06 [0.00, 0.29]	0.051	< 0.0001	< 0.0001	< 0.0001	< 1/1000	< 1/1000	< 1/1000	< 1/1000
Orientation	< 0.0001	0.51 [0.42, 0.60]	0.55 [0.49, 0.60]	0.01 [0.00, 0.18]	0.39	< 0.0001	< 0.0001	< 0.0001	< 1/1000	< 1/1000	< 1/1000	< 1/1000
Worse	< 0.0001	0.61 [0.51, 0.71]	0.39 [0.24, 0.53]	0.14 [0.07, 0.54]	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 1/1000	< 1/1000	< 1/1000	< 1/1000

experiment found that forcing people to answer questions quickly made them give more socially desirable answers (estimated standardized mean difference of 0.12 with 95% confidence interval from 0.01 to 0.22). The replication found an even larger pooled effect estimate (estimated standardized mean difference of 0.16 with 95% confidence interval from 0.06 to 0.25). However, there was considerable heterogeneity across the four replications ($p_Q = 0.017$).

3 Notation and assumptions

Denote by $\hat{\theta}_o$ and σ_o the effect estimate of the unknown effect size θ_o and its standard error obtained from the original study. Similarly, denote by $\hat{\boldsymbol{\theta}}_r = (\hat{\theta}_{r1}, \dots, \hat{\theta}_{rn})^\top$ and $\boldsymbol{\sigma}_r = (\sigma_{r1}, \dots, \sigma_{rn})^\top$ the effect estimates of the unknown effect sizes $\boldsymbol{\theta}_r = (\theta_{r1}, \dots, \theta_{rn})^\top$ and their standard errors obtained from n replication studies. Throughout, we will assume that effect estimates are normally distributed around their unknown effect size with variance equal to their squared standard error, i.e.,

$$\hat{\theta}_o \mid \theta_o \sim \mathrm{N}_1(\theta_o, \sigma_o^2) \qquad \text{and} \qquad \hat{\boldsymbol{\theta}}_r \mid \boldsymbol{\theta}_r \sim \mathrm{N}_n\{\boldsymbol{\theta}_r, \operatorname{diag}(\boldsymbol{\sigma}_r^2)\} \tag{1}$$

with $\boldsymbol{\sigma}_r^2 = (\sigma_{r1}^2, \dots, \sigma_{rn}^2)^\top$ and $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denoting the p-variate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. For some effect size types a transformation may be required to make the normality assumption more accurate (for example, a log transformation for an odds ratio effect size). In addition, denote by

$$\hat{\theta}_r = \frac{\sum_{i=1}^n w_{ri}\hat{\theta}_{ri}}{\sum_{i=1}^n w_{ri}} \qquad \text{and} \qquad \sigma_r = \frac{1}{\sqrt{\sum_{i=1}^n w_{ri}}} \tag{2}$$

the pooled replication effect estimate and standard error with weights $\boldsymbol{w}_r = (w_{r1}, \dots, w_{rn})^\top$. There are three typical cases for the weights: first, $w_{ri} = \sigma_{ri}^{-2}$, which arises from a common-effect model

$$\theta_r = \theta_{r1} = \cdots = \theta_{rn}, \tag{3}$$

second, $w_{ri} = (\tau^2 + \sigma_{ri}^2)^{-1}$, which arises from an additive heterogeneity model

$$\boldsymbol{\theta}_r \mid \theta_r \sim \mathrm{N}_n\{\theta_r \mathbf{1}_n, \tau^2 \operatorname{diag}(\mathbf{1}_n)\} \tag{4}$$

with $\mathbf{1}_n$ a vector of n ones and heterogeneity variance $\tau^2 \geq 0$, and third, $w_{ri} = (\phi\sigma_{ri}^2)^{-1}$, which arises from a multiplicative heterogeneity model

$$\boldsymbol{\theta}_r \mid \theta_r \sim \mathrm{N}_n\{\theta_r \mathbf{1}_n, (\phi - 1) \operatorname{diag}(\boldsymbol{\sigma}_r^2)\} \tag{5}$$

with heterogeneity multiplier $\phi > 0$. For all three models, the likelihood of the replication effect estimates (potentially marginalized over the study-specific effect sizes) can be written as

$$\hat{\boldsymbol{\theta}}_r \mid \theta_r \sim \mathrm{N}_n\{\theta_r \mathbf{1}_n, \operatorname{diag}(\boldsymbol{w}_r^{-1})\} \tag{6}$$

with $\boldsymbol{w}_r^{-1} = (w_{r1}^{-1}, \dots, w_{rn}^{-1})^\top$, which is often the most useful form for derivations and computations.

An important question is what value to choose for the heterogeneity parameters τ^2 and ϕ. Frequentist approaches typically plug in an estimate based on the data (see Veroniki et al., 2015; Baker and Jackson, 2012, for an overview of available estimators), while full Bayes approaches specify a prior for the respective parameter (Röver et al., 2021). Here, we will only consider the plug-in approach and outline generalizations to full Bayes approaches in the discussion.

4 Multisite generalization of the two-trials rule

The most commonly used criterion for replication success in the singlesite setting is to require that both the original and the replication p-values (two-sided) are smaller than some threshold α, typically $\alpha = 0.05$, and that their effect estimates go in the same direction. Consistency of effect direction can alternatively be accounted for by using one-sided p-values and halving the threshold, conventionally $\alpha = 0.025$. This criterion is also known as the *two-trials rule* in drug regulation and is typically required for a new drug to be approved for sale on the market (Senn, 2007, Section 12.2.8).

Replication researchers have adapted the criterion to the multisite setting in two ways: (i) a *meta-analytic significance* criterion, i.e., pooling the replications with meta-analysis and then computing a meta-analytic p-value from the pooled estimate and standard error, typically assuming an additive heterogeneity model, (ii) a *vote counting* criterion, i.e., computing the proportion of individual replication p-values that are significant in the same direction as the original one. While both approaches seem intuitive, they are not the only possible way for generalizing the two-trials rule and come with their own shortcomings. In particular, the vote counting approach has been criticized for being too stringent since its power may decrease as the number of studies increases (Mathur and VanderWeele, 2020).

Rosenkranz (2022) and Held (2023) discussed various generalizations of the two-trials rule to more than two studies in drug regulation, and these are equally applicable to the replication settings. The idea is that a generalized method should not become more stringent as more studies are added, as would be the case, for example, if the same level α were used to threshold each p-value. To achieve this, Rosenkranz (2022) suggested that the type I error rate of the procedure should be controlled at the same level as the two-trials rule for two studies, typically $\alpha^2 = 0.025^2 = 0.0625\%$ for one-sided p-values, even if there are more than two studies. Assuming that the original study was significant at level α and that n replication studies are conducted, this can be achieved by using a level $\alpha^{1/n}$ for the thresholding of the replication p-values. An equivalent but perhaps easier to interpret approach is to take the maximum of the n replication p-values raised to the power of n ($p_r = \max\{p_{r1}, \dots, p_{rn}\}^n$) and compare it to the ordinary level α, since p_r is a valid p-value with a uniform distribution under the null hypothesis of no effect in all studies. Held (2023) called this approach the *n-trials rule*, and he noted that it is only one of several possible methods for combining the p-values from replication studies. There is a large literature on p-value combination methods (see, for example, Hedges and Olkin, 1985, Chapter 3), and Table 2 gives several other methods. The p-value combination perspective also highlights that the commonly used meta-analytic criterion significance is only one possible way of doing this, as it corresponds to Stouffer's method using common-effect weights $w_{ri} = 1/\sigma^2_{ri}$.

The obvious question to ask is which method is the most useful in the replication setting? This depends on what properties the combined p-value should have. Mathematical texts have focused

Table 2: Different methods for combining p-values $\boldsymbol{p}_r = (p_{r1}, \dots, p_{rn})^\top$ from n replication studies into an overall replication p-value p_r. The first example combines the one-sided p-values $\boldsymbol{p}_r = (10^{-4}, 2 \times 10^{-4}, 0.003, 0.005, 0.02, 0.05, 0.098, 0.17, 0.18, 0.2, 0.21, 0.23, 0.24, 0.27, 0.3, 0.39, 0.45, 0.49, 0.7, 0.79, 0.85, 0.95)^\top$ from the facial feedback replications (Wagenmakers et al., 2016), the second example combines the one-sided p-values $\boldsymbol{p}_r = (0.032, 0.096, 0.098, 0.24, 0.24, 0.3, 0.3, 0.33, 0.47, 0.48, 0.53, 0.57, 0.65, 0.66, 0.75, 0.77, 0.82, 0.91)^\top$ from the moral credentialling replications (Ebersole et al., 2016), and the third example combines the one-sided p-values $\boldsymbol{p}_r = (6.2 \times 10^{-9}, 0.0034, 0.014, 0.043)^\top$ from the FSD replications (Protzko et al., 2020).

Method	Combined replication p-value	Ex. 1	Ex. 2	Ex. 3
n-trials rule	$p_r = \max\{p_{r1}, \dots, p_{rn}\}^n$	0.21	0.32	7.7×10^{-5}
Held	$p_r = \Pr(\chi^2_1 > h)$ with $h = \sum_{i=1}^n n^2/\Phi^{-1}(1 - p_{ri})^2$	0.16	0.22	5.9×10^{-6}
Pearson	$p_r = \Pr(\chi^2_{2n} \leq k)$ with $k = -2\sum_{i=1}^n \log(1 - p_{ri})$	0.23	0.009	3.1×10^{-5}
Edgington	$p_r = 1 - \frac{1}{n!}\sum_{k=0}^{\lfloor s \rfloor}(-1)^k \binom{n}{k}(s - k)^n$ with $s = \sum_{i=1}^n p_{ri}$	0.41	0.001	3×10^{-5}
Tippet	$p_r = 1 - (1 - \min\{p_{r1}, \dots, p_{rn}\})^n$	0.82	0.004	1.9×10^{-8}
Stouffer	$p_r = 1 - \Phi(z_r)$ with $z_r = \dfrac{\sum_{i=1}^n \Phi^{-1}(1 - p_{ri})\sqrt{w_{ri}}}{\sqrt{\sum_{i=1}^n w_{ri}}}$	0.7	3.7×10^{-5}	4.1×10^{-8}
Fisher	$p_r = \Pr(\chi^2_{2n} > f)$ with $f = -2\sum_{i=1}^n \log(p_{ri})$	0.62	3.4×10^{-5}	1.4×10^{-9}

on the optimality properties of admissibility and efficiency. A combination method is admissible if it provides a most powerful test of a null hypothesis against some alternative (there may be multiple most powerful tests), while efficiency quantifies how fast the evidence against the null hypothesis grows with sample size. Based on these criteria, Stouffer's, Tippet's, and Fisher's methods have often been recommended, while Pearson's and Edgington's methods have been dismissed (Hedges and Olkin, 1985; Hartung et al., 2008).

Held (2023) noted that inadmissible methods, such as the n-trials rule, Pearson's, and Held's methods, may have desirable properties in settings such as drug development or replication studies because they require *all* studies to be convincing to some degree (making them inadmissible), whereas Stouffer's, Tippet's, and Fisher's methods do not. This can be formalized by the *partial type I error rate* of a method which is related to the partial or *no-replicability null hypothesis* where only some studies have a true null effect (Heller et al., 2014). It turns out that the n-trials rule, Held's method, Pearson's method, and Edgington's method control the partial type I error rate to some extent whereas for Fisher's method, Stouffer's method, and Tippets's method there is no non-trivial bound on the partial type I error rate (Micheloud et al., 2023; Held, 2023). While the n-trials rule controls the partial type I error rate at level α for any number of studies, the number of studies influences the partial type I error rate of the other methods.

The behavior of the different methods is illustrated by the examples in Table 2: In the first example, the combined p-values are of the same order of magnitude across the different methods, as these examples do not show too much heterogeneity in their p-values. In contrast, in the second example there are several very convincing replications with very small p-values, but also quite a few replications with large p-values. In this case, the n-trials rule and Held's method lead to the largest combined p-values, followed by Pearson, Edgington, and Tippet whose p-values are one order of magnitude smaller, followed by Stouffer and Fisher whose p-values are two orders of magnitude smaller. A similar difference can be seen in third example, where the p-values from the Tippet, Stouffer, and Fisher are several orders of magnitude smaller than the ones from the n-trials rule, Held, Pearson and Edgington, although all of them are substantially below the conventional threshold of $\alpha = 0.025$.

The desire to quantify and account for p-value heterogeneity may be the intention of the vote-counting approach used in some replication projects. Methods for combining p-values that control the partial type I error rate may be principled ways to achieve this goal, although it is still an open question as to which of these should be used as the default. An additional advantage of the p-value combination perspective is that the only assumption is the validity of each individual p-value. A p-value can be computed from an exact distribution, a bootstrap, or a permutation test, which are typically used in situations with small sample sizes or rare events where traditional meta-analysis methods often have poor performance due to their normality assumptions. Finally, one may want not only a combined p-value, but also a combined estimate. In this case, replication p-values can be calculated for different null hypotheses and the combined p-value can then be visualized as a function of the null value (a *p-value function*, see, for example, Fraser, 2019; Infanger and Schmidt-Trucksäss, 2019). The function can be cut at some level α and the null values with larger p-values form a $(1 - \alpha)$ confidence set. Similarly, the null value(s) at which the curve peaks can be taken as the point estimate(s). We will report on the details of this approach in future work.

5 Multisite generalization of the sceptical *p*-value

The sceptical p-value was introduced by Held (2020) as a replicability measure based on a reverse-Bayes approach. The idea is to assume a common effect size θ underlying the original and replication studies, and then to determine the variance of a zero-mean normal prior distribution for θ such that the resulting posterior credible interval includes zero, thereby indicating no longer evidence for an effect. This "sceptical" prior represents the position of a sceptic who does not believe in the presence of a genuine effect. The aim of the replication study is then to prove the sceptic wrong by showing that there is conflict between the replication data and the sceptical prior. The sceptical p-value is a summary measure that quantifies the degree of conflict. In the following, we describe two possible generalizations of the procedure when there is more than one replication study – a pooling approach that first synthesizes the replication studies and then employs the singlesite procedure, and a multivariate approach that takes into account the multivariate structure of the data. Both approaches are illustrated in Figure 2.

5.1 Pooling approach

A first approach to generalizing the sceptical p-value is to pool the replication estimates and their standard errors, and then apply the singlesite procedure. This "pooling" approach was suggested by Held (2020) in the discussion of his article, and we will now describe it in more detail: Assume that there is an effect size underlying both studies ($\theta = \theta_o = \theta_r$) and determine the variance λ^2 of a zero-mean normal prior $\theta \sim \mathrm{N}_1(0, \lambda^2)$ such that the posterior of θ based on the original data no

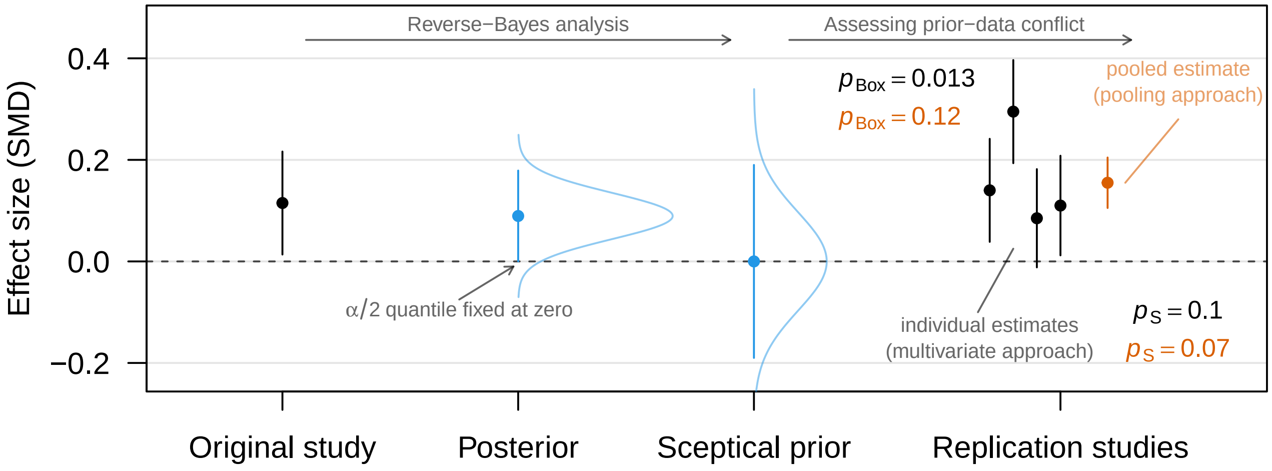

Figure 2: Shown are point estimates/means with 95% confidence/credible intervals. The original effect estimate from the FSD experiment (Protzko et al., 2020) is challenged with a sceptical prior such that the resulting posterior is no longer credible at level $\alpha = 5\%$ (two-sided). The prior-predictive tail probability p_{Box} quantifies the conflict between the sceptical prior and the replication data. Replication success at level α is achieved if $p_{\text{Box}} \leq \alpha$. The smallest level at which replication success can be achieved defines the sceptical p-value p_{S}. The pooling approach (orange) assesses conflict between the pooled replication estimate and the sceptical prior, whereas the multivariate approach (black) uses the joint distribution of the replication estimates.

longer provides evidence against a non-zero effect. That is, the variance is determined such that the $(1 - \alpha)$ posterior credible interval just includes the value of zero, which can be derived to be

$$\lambda_\alpha^2 = \begin{cases} \dfrac{\sigma_o^2}{z_o^2/z_{\alpha/2}^2 - 1} & \text{if } z_o^2 > z_{\alpha/2}^2 \\ \text{undefined} & \text{else} \end{cases} \tag{7}$$

where $z_o = \hat{\theta}_o/\sigma_o$ and $z_{\alpha/2}$ is the $(1 - \alpha/2)$ quantile of the standard normal distribution. The predictive distribution of the pooled replication estimate under this sceptical prior is then $\hat{\theta}_r \mid \lambda_\alpha^2 \sim \text{N}_1(0, \sigma_r^2 + \lambda_\alpha^2)$. A prior-predictive tail probability p_{Box} is then defined as the probability of the set of effect estimates with lower density than the observed replication effect estimate $\hat{\theta}_r$ under the prior-predictive distribution (Box, 1980), which is given by

$$p_{\text{Box}} = 2\left\{1 - \Phi\left(\frac{|\hat{\theta}_r|}{\sqrt{\lambda_\alpha^2 + \sigma_r^2}}\right)\right\} \tag{8}$$

with $\Phi(\cdot)$ the cumulative distribution function of the standard normal distribution. Held (2020) then defined replication success at level α whenever $p_{\text{Box}} \leq \alpha$, that is, 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. Mathematically, this means we have replication success at level α when

$$\hat{\theta}_r^2 \left(\frac{\sigma_o^2}{z_o^2/z_{\alpha/2}^2 - 1} + \sigma_r^2 \right)^{-1} \geq z_{\alpha/2}^2. \tag{9}$$

Clearly, p_{Box} depends on the level α. To remove this dependence, we can determine the smallest level α at which the success condition (9) holds. The solution is the sceptical p-value

$$p_{\text{S}}' = 2\left\{1 - \Phi(|z_{\text{S}}|)\right\} \tag{10}$$

where

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

with arithmetic mean $z_{\text{A}}^2 = (z_o^2 + z_r^2)/2$ and harmonic mean $z_{\text{H}}^2 = 2/(1/z_o^2 + 1/z_r^2)$ of the squared z-statistics $z_i = \hat{\theta}_i/\sigma_i$ for $i \in \{o, r\}$, and variance ratio $c = \sigma_o^2/\sigma_r^2$. A sceptical p-value $p_{\text{S}}' \leq \alpha$ is then equivalent to replication success at level α.

Defined in this way, the sceptical p-value does not take the direction of the effect estimates into account, and replication success may hence be achieved even if original and replication estimates go in opposite direction. Held (2020) therefore defined a one-sided version through

$$p_{\text{S}} = \begin{cases} 1 - \Phi(|z_{\text{S}}|) & \text{if } \operatorname{sign}(\hat{\theta}_o) = \operatorname{sign}(\hat{\theta}_r) \\ \Phi(|z_{\text{S}}|) & \text{if } \operatorname{sign}(\hat{\theta}_o) \neq \operatorname{sign}(\hat{\theta}_r), \end{cases} \tag{11}$$

which ensures that such a "replication paradox" (Ly et al., 2018) cannot occur.

The sceptical p-value in the form (11), called the *nominal sceptical p-value*, is not a proper p-value with uniform distribution under a null hypothesis, and its interpretation can be challenging. To address this, two calibrations have been developed, and both can also be applied to the pooling version of the sceptical p-value: First, the *golden sceptical p-value*

$$p_{\text{S}g} = 1 - \Phi\{\Phi^{-1}(1 - p_{\text{S}})\sqrt{\phi}\}$$

with $\phi = (1 + \sqrt{5})/2 \approx 1.62$ the golden ratio. This calibration ensures that for an original study just significant at level α, replication success at level α, i.e., $p_{\text{S}g} \leq \alpha$, is only possible if the replication effect estimate does not shrink compared to the original one, i.e., $\hat{\theta}_r/\hat{\theta}_o \geq 1$ (Held et al., 2022). This calibration hence takes into account effect shrinkage, which is often a serious concern in the replication setting. Second, the *calibrated sceptical p-value*

$$p_{\text{S}c} = F_c(p_{\text{S}})$$

with F_c a transformation that depends on the variance ratio $c = \sigma_o^2/\sigma_r^2$, and which ensures that $p_{\text{S}c}$ has a uniform distribution under the null hypothesis $H_0\colon \theta = 0$ (Micheloud et al., 2023). Calibrated in this form, the sceptical p-value can be interpreted as an ordinary frequentist p-value.

5.2 Multivariate approach

Instead of pooling the replication effect estimates via (2) and then applying the singlesite sceptical p-value procedure we will now generalize the procedure to use the multivariate likelihood (1) for the replication effect estimates. As before, assume that the same effect size underlies both studies ($\theta = \theta_o = \theta_r$), and in addition one of the three replication effect size models (common-effect, additive heterogeneity, multiplicative heterogeneity) with appropriate weights vector $\boldsymbol{w}_r$ so that marginally the likelihood is

$$\hat{\boldsymbol{\theta}}_r \mid \theta \sim \mathrm{N}_n\{\theta \mathbf{1}_n, \mathrm{diag}(\boldsymbol{w}_r^{-1})\}.$$

The prior predictive distribution of $\hat{\boldsymbol{\theta}}_r$ based on the sceptical prior for θ with sufficiently sceptical prior variance λ^2_α from (7) is then given by

$$\hat{\boldsymbol{\theta}}_r \mid \lambda^2_\alpha \sim \mathrm{N}_n\{0\mathbf{1}_n, \boldsymbol{\Sigma}_r = \mathrm{diag}(\boldsymbol{w}_r^{-1}) + \lambda^2_\alpha \mathbf{J}_n\}$$

where $\mathbf{J}_n = \mathbf{1}_n\mathbf{1}_n^\top$ denotes an $n \times n$ matrix of ones. The replication effect estimates are now correlated (i.e., the off-diagonal of Σ_r is non-zero), because all replication effect estimates share the same prior for θ. Box's tail probability p_{Box} is defined as the probability of the set of effect estimates with lower density than the observed replication effect estimates $\hat{\boldsymbol{\theta}}_r$, i.e.,

$$p_{\text{Box}} = \Pr\left\{\mathrm{N}_n\left(0\mathbf{1}_n, \boldsymbol{\Sigma}_r\right) \leq \mathrm{N}_n\left(\hat{\boldsymbol{\theta}}_r \mid 0\mathbf{1}_n, \boldsymbol{\Sigma}_r\right)\right\} = \Pr\left\{\chi^2_n > T(\hat{\boldsymbol{\theta}}_r)\right\}$$

with χ^2_n a chi-squared random variable with n degrees of freedom and test-statistic

$$T(\hat{\boldsymbol{\theta}}_r) = \hat{\boldsymbol{\theta}}_r^\top \Sigma_r^{-1} \hat{\boldsymbol{\theta}}_r \tag{12}$$

$$= \sum_{i=1}^n w_{ri}\,(\hat{\theta}_{ri} - \hat{\theta}_r)^2 + \frac{\hat{\theta}_r^2}{\lambda^2_\alpha + \sigma^2_r} \tag{13}$$

and where $\hat{\theta}_r$ and σ_r are the pooled replication estimate and its standard error from (2), see Appendix A for a proof. We see that the test-statistic (13) consists of two terms. The first term is the generalized Q-statistic

$$Q = \sum_{i=1}^n w_{ri}\,(\hat{\theta}_{ri} - \hat{\theta}_r)^2 \tag{14}$$

and it quantifies the heterogeneity among the effect estimates, reducing to the ordinary Cochrane's Q statistic with common-effect weights ($w_{ri} = 1/\sigma^2_{ri}$). More heterogeneous estimates than what would be expected under the sceptical prior lead to larger Q, and thereby increase the degree of prior-data conflict. The second term quantifies how much the pooled replication effect estimate is different from zero, larger values indicating more prior-data conflict.

As in the singlesite case, we define replication success at level α when

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

and want to find the smallest value at which replication success can be established, the *sceptical p-value*

$$p_{\text{S}}' = \inf\{\alpha : p_{\text{Box}} \leq \alpha\}\,. \tag{15}$$

Plugging the expression for the sceptical prior variance (7) into (13), we see that replication success at level α is achieved when

$$T(\hat{\boldsymbol{\theta}}_r) = Q + \hat{\theta}_r^2 \left(\frac{\sigma_o^2}{z_o^2/z_{\alpha/2}^2 - 1} + \sigma_r^2\right)^{-1} \geq \chi_n^2(1-\alpha) \tag{16}$$

with $\chi_n^2(1-\alpha)$ the $(1-\alpha)$ quantile of the chi-squared distribution with n degrees of freedom. Comparing the success condition from the pooling approach (9) with the success condition from the multivariate approach (16), we see that the latter differs through the addition of the Q-statistic and a different reference quantile (with n rather than one degrees of freedom). When only one replication is conducted ($n = 1$), we have that $Q = 0$ and $\chi_1^2(1-\alpha) = z_{\alpha/2}^2$, and can therefore obtain the closed-form solution for the sceptical p-value from (10). When more than one replications are conducted ($n > 1$), this is not possible anymore with the multivariate approach, but a solution can nevertheless be obtained numerically.

The test-statistic $T(\hat{\boldsymbol{\theta}}_r)$ does not directly take the direction of the replication effect estimates $\hat{\theta}_r$ into account. This means that the sceptical p-value may indicate a large degree of replication success despite that the replication effect estimates go in opposite direction of the original effect estimate. Taking into account effect direction for computing p-values in multivariate settings is not as straightforward as in the univariate setting (Follmann, 1996). A pragmatic choice is to only look at the effect direction of the pooled replication effect estimate $\hat{\theta}_r$ and define a one-sided sceptical p-value by

$$p_{\text{S}} = \begin{cases} p_{\text{S}}'/2 & \text{if } \operatorname{sign}(\hat{\theta}_r) = \hat{\theta}_o \\ 1 - p_{\text{S}}'/2 & \text{if } \operatorname{sign}(\hat{\theta}_r) \neq \hat{\theta}_o \end{cases}$$

with p_{S}' the non-directional sceptical p-value (15). This choice reduces to the singlesite one-sided sceptical p-value when $n = 1$.

Table 1 shows the one-sided sceptical p-values for the example replications calculated using both the pooling and multivariate approaches. For two studies, the multivariate version is slightly smaller than the pooling version. For example, in the "Redemption" study, the multivariate sceptical p-value ($p_{\text{S}} = 0.84$) is slightly smaller than the pooling version ($p_{\text{S}} = 0.87$), although both indicate hardly any replication success. The same is also true for the "Misreporting" study. For the remaining sixteen studies, the multivariate version is larger than the pooling version, leading to a more conservative assessment of replicability.

Again, it may be desirable to calibrate the sceptical p-value, either to an ordinary frequentist p-value or via the relative effect size as discussed at the end of Section 5.1. However, the fact that the sceptical p-value under the multivariate approach is not available in closed-form complicates the derivation of both calibrations. One brute-force way to obtain a frequentist calibration is to simulate original and replication effect estimates under the null hypothesis of no effect ($\theta_o = \theta_r = 0$) with the same standard errors as the observed ones, calculate sceptical p-values from them, and then calibrate the actually observed sceptical p-value against its simulated null distribution. In contrast, it is conceptually less clear how to generalize the relative effect size calibration to the multivariate setting: There is more than one replication effect estimate, so in principle there are n relative effect sizes to consider. One approach would be to consider only the relative effect size based on the pooled replication effect estimate $\hat{\theta}_r$. Furthermore, the condition for replication success (16) depends not only on the relative effect estimate(s) and standard error(s), but also on the number of studies n and the Q-statistic. For a given n and Q, a monotone transformation of the sceptical p-value may be determined numerically, which ensures that replication success at level γ is only possible if the pooled replication estimate does not shrink compared to an original estimate that is just significant at the same level. However, this transformation will be different for each n and Q, and may perhaps confuse rather than help.

6 Multisite generalization of Bayes factor methods

Several Bayes factor based methods have been proposed to quantify replicability, for example, default Bayes factors (Verhagen and Wagenmakers, 2014), replication Bayes factors (Verhagen and Wagenmakers, 2014; Ly et al., 2018; Harms, 2019) or sceptical Bayes factors (Pawel and Held, 2022). Pawel et al. (2023) illustrated how the replication Bayes factor may be generalized to the multisite setting under an additive heterogeneity model, but neither of these methods have been extended to the multisite setting considered here. Researchers have rather applied singlesite default and replication Bayes factors to each replication study and then counted how many of them indicated non-anecdotal evidence for either the null hypothesis or the alternative (Wagenmakers et al., 2016). In the following, we illustrate how possible generalizations may be made within the framework of normally distributed effect estimates.

6.1 The default Bayes factor

The default Bayes factor approach is similar to the two-trials rule in the sense that the evidence for the null hypothesis $H_0 \colon \theta = 0$ against the alternative $H_1 \colon \theta \neq 0$ is assessed independently in the original study and its replication(s), but using Bayes factors instead of p-values. Replication success is established if all Bayes factors indicate convincing evidence for the alternative H_1. Typically, the classification from Jeffreys (1961) is used, e.g., a Bayes factor $\text{BF}_{01} < 1/3$ indicates substantial evidence and a Bayes factor $\text{BF}_{01} < 1/10$ indicates strong evidence against the null hypothesis. The Bayes factor is called "default" because it uses a "default prior distribution" that has certain objective Bayes properties (Bayarri et al., 2012) and is not informed by external knowledge, for example, a standard Cauchy or a unit-information normal distribution (Kass and Wasserman, 1995). Here, we examine default Bayes factors based on normal priors because they are intuitive to interpret and specify, are available in closed-form, and are not too different from Bayes factors based on Cauchy priors in typical situations.

For original ($i = o$) or replication data ($i = r$), consider the default Bayes factor for testing $H_0 \colon \theta_i = 0$ against $H_1 \colon \theta_i \neq 0$ with normal prior $\theta_i \mid H_1 \sim \text{N}_1(\mu, \lambda^2)$ truncated to the interval $a \leq \theta_i \leq b$

assigned to the underlying effect size θ_i under the alternative H_1. Two-sided alternatives are obtained with $(a \downarrow -\infty, b \uparrow +\infty)$ whereas one-sided alternatives are obtained with $(a = 0, b \uparrow +\infty)$ or $(a \downarrow -\infty, b = 0)$. Using the result from Appendix B, the Bayes factor can be derived in closed-form

$$\mathrm{BF}_{01}(\hat{\theta}_i) = \frac{f(\hat{\theta}_i \mid H_0)}{f(\hat{\theta}_i \mid H_1)} = \sqrt{1 + \lambda^2/\sigma_i^2} \times \exp\left[-\frac{1}{2}\left\{\frac{\hat{\theta}_i^2}{\sigma_i^2} - \frac{(\hat{\theta}_i - \mu)^2}{\sigma_i^2 + \lambda^2}\right\}\right] \times \frac{\{\Phi(\frac{b-\mu}{\lambda}) - \Phi(\frac{a-\mu}{\lambda})\}}{\{\Phi(\frac{b-\mu'}{\lambda'}) - \Phi(\frac{a-\mu'}{\lambda'})\}} \tag{17}$$

with updated variance $(\lambda')^2 = (\lambda^{-2} + \sigma_i^{-2})^{-1}$, updated mean $\mu' = (\hat{\theta}\sigma_i^{-2} + \mu\lambda^{-2})(\lambda')^2$, and, in case of the replication data, $\hat{\theta}_r$ and σ_r the pooled replication effect estimate and standard error from (2). Remarkably, the same Bayes factor (17) is obtained regardless of whether the replication effect estimates $\hat{\boldsymbol{\theta}}_r$ and standard errors $\boldsymbol{\sigma}_r$ are first pooled into $\hat{\theta}_r$ and σ_r, or whether the full vector-valued data are used directly to compute the Bayes factor. Thus, there is no difference between a "pooling" and a "multivariate" approach as there is for the sceptical p-value.

6.2 The replication Bayes factor

In contrast to the default Bayes factor, the replication Bayes factor uses the posterior distribution of θ based on the original data (assuming a common effect size underlying original and replication studies) as prior under the alternative $H_{\mathrm{A}} \colon \theta \sim f(\theta \mid \hat{\theta}_o)$, with the subscript A denoting *advocacy*, as the prior should represent the position of an advocate of the original finding. Based on an initial flat prior for θ, the posterior is a normal distribution $\theta \sim \mathrm{N}_1(\hat{\theta}_o, \sigma_o^2)$ which may similarly be truncated to an interval $a \leq \theta \leq b$ as the default Bayes factor to account for effect direction. The replication Bayes factor is hence a special case of the default Bayes factor (17) with $\mu = \hat{\theta}_o$ and $\lambda = \sigma_o$, i.e.,

$$\mathrm{BF}_{\mathrm{R}} = \mathrm{BF}_{0\mathrm{A}}(\hat{\boldsymbol{\theta}}_r) = \sqrt{1 + \sigma_o^2/\sigma_r^2} \times \exp\left[-\frac{1}{2}\left\{\frac{\hat{\theta}_r^2}{\sigma_r^2} - \frac{(\hat{\theta}_r - \hat{\theta}_o)^2}{\sigma_r^2 + \sigma_o^2}\right\}\right] \times \frac{\{\Phi(\frac{b-\hat{\theta}_o}{\sigma_o}) - \Phi(\frac{a-\hat{\theta}_o}{\sigma_o})\}}{\{\Phi(\frac{b-\mu'}{\lambda'}) - \Phi(\frac{a-\mu'}{\lambda'})\}} \tag{18}$$

with updated variance $(\lambda')^2 = (\sigma_r^{-2} + \sigma_o^{-2})^{-1}$ and updated mean $\mu' = (\hat{\theta}_r\sigma_r^{-2} + \hat{\theta}_o\sigma_o^{-2})(\lambda')^2$. The alternative is either two-sided $(a \downarrow -\infty, b \uparrow +\infty)$ or one-sided in the direction of the original effect estimate, e.g., $(a = 0, b \uparrow +\infty)$ for a positive original effect estimate. As for the default Bayes factor, the same replication Bayes factor is obtained regardless of whether a pooling or multivariate approach is used.

6.3 The sceptical Bayes factor

The sceptical Bayes factor was proposed by Pawel and Held (2022) as a Bayes factor analog to the sceptical p-value. The idea is similar: first, one determines a sceptical prior such that the original data no longer provide evidence against a null hypothesis, second, one assesses the conflict between the sceptical prior and the replication data, the more conflict, the larger the degree of replication success. In contrast to the sceptical p-value, however, Bayes factors instead of tail probabilities are used for quantifying evidence and prior-data conflict. We will now outline how the procedure can be generalized to the multisite setting.

In a first step, the original estimate $\hat{\theta}_o$ with standard error σ_o is used to compute the Bayes factor (17) with sceptical prior $\theta \mid H_1 \sim \mathrm{N}_1(0, \lambda^2)$ under the alternative. The variance λ^2_γ of the prior is then determined such that the Bayes factor no longer indicates evidence against the null hypothesis at level γ (i.e, such that $\mathrm{BF}_{01}(\hat{\theta}_o) = \gamma$), which can be derived in closed-form

$$\lambda^2_\gamma = \begin{cases} -\sigma^2_o \left(\dfrac{z^2_o}{q} + 1\right) & \text{if } -\dfrac{z^2_o}{q} \geq 1 \\ \text{undefined} & \text{else} \end{cases} \tag{19}$$
$$\text{where } q = \mathrm{W}_{-1}\left\{-\frac{z^2_o}{\gamma^2} \exp\left(-z^2_o\right)\right\}$$

with $\mathrm{W}_{-1}(\cdot)$ the branch of the Lambert W function (Corless et al., 1996) that satisfies $\mathrm{W}(y) \leq -1$ for $y \in [-e^{-1}, 0)$, see Appendix A in Pawel and Held (2022) for details.

In a second step, the replication effect estimates $\hat{\boldsymbol{\theta}}_r$ and standard errors $\boldsymbol{\sigma}_r$ are used to contrast the evidence for the sceptic's hypothesis $H_\mathrm{S}\colon \theta \sim \mathrm{N}_1(0, \lambda^2_\gamma)$ to the advocate's hypothesis $H_\mathrm{A}\colon \theta \sim f(\theta \mid \hat{\theta}_o)$ with potential truncation of the prior $a \leq \theta \leq b$. Using the result from Appendix B, the Bayes factor can be derived to be

$$\mathrm{BF}_{\mathrm{SA}}(\hat{\boldsymbol{\theta}}_r) = \sqrt{\frac{\sigma^2_r + \sigma^2_o}{\sigma^2_r + \lambda^2_\gamma}} \times \exp\left[-\frac{1}{2}\left\{\frac{\hat{\theta}^2_r}{\sigma^2_r + \lambda^2_\gamma} - \frac{(\hat{\theta}_r - \hat{\theta}_o)^2}{\sigma^2_r + \sigma^2_o}\right\}\right] \times \frac{\left\{\Phi\left(\frac{b - \hat{\theta}_o}{\sigma_o}\right) - \Phi\left(\frac{a - \hat{\theta}_o}{\sigma_o}\right)\right\}}{\left\{\Phi\left(\frac{b - \mu'}{\lambda'}\right) - \Phi\left(\frac{a - \mu'}{\lambda'}\right)\right\}} \tag{20}$$

with pooled replication effect estimate $\hat{\theta}_r$ and standard error σ_r from (2), and updated variance $(\lambda')^2 = (\sigma_r^{-2} + \sigma_o^{-2})^{-1}$ and mean $\mu' = (\hat{\theta}_r\sigma_r^{-2} + \hat{\theta}_o\sigma_o^{-2})(\lambda')^2$. Replication success at level γ is then defined as when the Bayes factor (20) is less than or equal to γ, i.e., $\mathrm{BF}_{\mathrm{SA}}(\hat{\boldsymbol{\theta}}_r) \leq \gamma$, since then there is at least as much evidence for the advocate over the sceptic as there was evidence against the null hypothesis. Similar to the sceptical p-value, the sceptical Bayes factor BF_S is the smallest level at which replication success can be established. Apart from the special case of equal original and replication standard errors ($\sigma_o = \sigma_r$), there is no closed-form solution, but the sceptical Bayes factor must be determined numerically. However, since both pooling and multivariate approaches lead to the same Bayes factor (20), the singlesite implementation of the sceptical Bayes factor can be conveniently applied to the pooled replication estimates for this purpose.

6.4 Comparison of Bayes factors on running examples

Table 1 shows default, replication, and sceptical Bayes factors for the example replication studies. We see that the default Bayes factor $\mathrm{BF}_{01}(\hat{\theta}_r)$ and the replication Bayes factor BF_R most of them time agree, at least qualitatively. On the other hand, the sceptical Bayes factor BF_S indicates in all cases a smaller degree of replication success, and is even nonexistent in three replication studies as these are so unconvincing such that replication success cannot be established at any level. In this case, the default and the replication Bayes factor indicate evidence for the null hypothesis, e.g., in the "Facial feedback" study. Interestingly, Wagenmakers et al. (2016) also conducted a Bayes factor analysis for the "Facial feedback" replications with the conclusion:

> "16 replication Bayes factors provide evidence in favor of the null hypothesis, and 12 do this in a nonanecdotal manner (i.e., [$\mathrm{BF}_\mathrm{R} > 3$]). As before, these Bayes factors are not

> independent and hence may not be multiplied." (Wagenmakers et al., 2016, p. 923)

Our replication Bayes factor of $\mathrm{BF}_{\mathrm{R}} = 38$ agrees with this conclusion, indicating strong evidence for the null hypothesis, but does so using all replication studies simultaneously.

7 Discussion

We have shown how several singlesite replicability measures – the two-trials rule, the sceptical p-value, and the default/replication/sceptical Bayes factor – can be generalized to the multisite setting. In the case of the Bayes factor methods, it turns out that it is as simple as applying the singlesite procedure to a pooled estimate, even when the multivariate structure of the data is taken into account, whereas for the two-trials rule and the sceptical p-value this is not so straightforward.

We have not considered a full Bayesian approach with a prior on the heterogeneity parameters. Such an approach could potentially increase the efficiency of the methods, especially in scenarios with only few replication studies, but at the cost of losing closed-form expressions for marginal likelihoods, and additional complexity in specifying the prior hyperparameters.

An important aspect is the design of new replication studies, in particular their sample size determination. Hedges and Schauer (2021) and Pawel et al. (2023) have developed frequentist and Bayesian approaches, respectively, for doing so. Since the approach from Pawel et al. (2023) requires only the "success region" of the replication effect estimate(s), it can be readily applied to the multisite generalizations discussed in this paper.

Software and data

The three data sets used in this study were extracted from the supplementary material of Protzko et al. (2020, `https://osf.io/42ef9/`), Figure 4 in Wagenmakers et al. (2016), and in the case of the Ebersole et al. (2016) data from Mathur and VanderWeele (2020, `https://osf.io/36ed5/`) who reanalyzed the same data set. All analyses were conducted in the R programming language version 4.3.1 (R Core Team, 2020). The code and data to reproduce our analyses is openly available at `https://doi.org/10.5281/zenodo.8379956`.

Acknowledgments

This work was partially supported by the Swiss National Science Foundation (#189295).

Conflict of interest

The authors declare no conflict of interest.

Appendix A Decomposition of the test-statistic

By using the Sherman-Morrison formula, we can write the inverse of the prior predictive covariance matrix $\boldsymbol{\Sigma}_r$ as

$$\begin{aligned}\boldsymbol{\Sigma}_r^{-1} &= \left\{\operatorname{diag}(\boldsymbol{w}_r^{-1}) + \lambda_\alpha^2 \mathbf{1}_n \mathbf{1}_n^\top\right\}^{-1} \\ &= \operatorname{diag}(\boldsymbol{w}_r) - \frac{\operatorname{diag}(\boldsymbol{w}_r)\lambda_\alpha^2 \mathbf{1}_n \mathbf{1}_n^\top \operatorname{diag}(\boldsymbol{w}_r)}{1 + \lambda_\alpha^2 \mathbf{1}_n^\top \operatorname{diag}(\boldsymbol{w}_r)\mathbf{1}_n} \\ &= \operatorname{diag}(\boldsymbol{w}_r) - \frac{\boldsymbol{w}_r \boldsymbol{w}_r^\top}{\lambda_\alpha^{-2} + \sigma_r^{-2}}\end{aligned}$$

with σ_r the pooled standard error from (2). Hence, the test-statistic $T(\hat{\boldsymbol{\theta}}_r)$ is

$$\begin{aligned}T(\hat{\boldsymbol{\theta}}_r) &= \hat{\boldsymbol{\theta}}_r^\top \boldsymbol{\Sigma}_r^{-1} \hat{\boldsymbol{\theta}}_r \\ &= \hat{\boldsymbol{\theta}}_r^\top \operatorname{diag}(\boldsymbol{w}_r)\hat{\boldsymbol{\theta}}_r - \hat{\boldsymbol{\theta}}_r^\top \frac{\boldsymbol{w}_r \boldsymbol{w}_r^\top}{\lambda_\alpha^{-2} + \sigma^{-2}} \hat{\boldsymbol{\theta}}_r \\ &= \sum_{i=1}^n w_{ri}\hat{\theta}_{ri}^2 - \left(\sum_{i=1}^n w_{ri}\hat{\theta}_{ri}^2\right)^2 \frac{1}{\lambda_\alpha^{-2} + \sigma_r^{-2}}. \end{aligned} \tag{21}$$

Using the fact that we can decompose

$$\sum_{i=1}^n w_{ri}\hat{\theta}_{ri}^2 = \sum_{i=1}^n w_{ri}(\hat{\theta}_{ri} - \hat{\theta}_r)^2 + \frac{\hat{\theta}_r^2}{\sigma_r^2}$$

and that

$$\sum_{i=1}^n w_{ri}\hat{\theta}_{ri} = \hat{\theta}_r \sum_{i=1}^n w_{ri} = \frac{\hat{\theta}_r}{\sigma_r^2}$$

with $\hat{\theta}_r$ the pooled estimate of the replication effect estimates from (2), the test-statistic (21) can be written as

$$\begin{aligned}T(\hat{\boldsymbol{\theta}}_r) &= \sum_{i=1}^n w_{ri}(\hat{\theta}_{ri} - \hat{\theta}_r)^2 + \frac{\hat{\theta}_r^2}{\sigma_r^2}\left(1 - \frac{1}{1 + \sigma_r^2 \lambda_\alpha^{-2}}\right) \\ &= \sum_{i=1}^n w_{ri}(\hat{\theta}_{ri} - \hat{\theta}_r)^2 + \frac{\hat{\theta}_r^2}{\lambda_\alpha^2 + \sigma_r^2}.\end{aligned}$$

Appendix B Marginal likelihood under a truncated normal prior

Assume the data are given by effect estimates $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_n)$ with variances $\boldsymbol{w}^{-1} = (w_1^{-1}, \dots, w_n^{-1})$ and a likelihood $\hat{\boldsymbol{\theta}} \mid \theta \sim \mathrm{N}_n\{\theta \mathbf{1}_n, \operatorname{diag}(\boldsymbol{w}^{-1})\}$. In addition, assume a hypothesis H_k that assigns a normal prior $\theta \mid H_k \sim \mathrm{N}_1(\mu, \lambda^2)$ truncated to the interval $a \leq \theta \leq b$ to the underlying effect size θ. Using straightforward but tedious algebraic manipulations, the marginal likelihood of $\hat{\boldsymbol{\theta}}$ under hypothesis

H_k can be derived to be

$$
\begin{aligned}
f(\hat{\boldsymbol{\theta}} \mid H_k) &= \int f(\hat{\boldsymbol{\theta}} \mid \theta) f(\theta \mid H_k)\,\mathrm{d}\theta \\
&= \int_a^b \mathrm{N}_n\{\hat{\boldsymbol{\theta}} \mid \theta\mathbf{1}_n, \operatorname{diag}(\boldsymbol{w}^{-1})\}\,\mathrm{N}_1(\theta \mid \mu, \lambda^2)\left\{\Phi\left(\frac{b-\mu}{\lambda}\right) - \Phi\left(\frac{a-\mu}{\lambda}\right)\right\}^{-1} \mathrm{d}\theta \\
&= \sqrt{\frac{\prod_{i=1}^n w_i}{(2\pi)^n}}\left\{\Phi\left(\frac{b-\mu}{\lambda}\right) - \Phi\left(\frac{a-\mu}{\lambda}\right)\right\}^{-1} \int_a^b \exp\left[-\frac{1}{2}\underbrace{\left\{\sum_{i=1}^n w_i(\hat{\theta}_i - \theta)^2\right\}}_{=Q+\sigma^{-2}(\hat{\theta}-\theta)^2}\right] \mathrm{N}_1(\theta \mid \mu, \lambda^2)\,\mathrm{d}\theta \\
&= \underbrace{\sqrt{\frac{\prod_{i=1}^n w_i}{(2\pi)^n}}\exp\left(-\frac{Q}{2}\right)\sqrt{2\pi\sigma^2}}_{=K(Q)}\left\{\Phi\left(\frac{b-\mu}{\lambda}\right) - \Phi\left(\frac{a-\mu}{\lambda}\right)\right\}^{-1} \underbrace{\int_a^b \mathrm{N}_1(\hat{\theta} \mid \theta, \sigma^2)\,\mathrm{N}_1(\theta \mid \mu, \lambda^2)\,\mathrm{d}\theta}_{=\mathrm{N}_1(\hat{\theta} \mid \mu, \sigma^2+\lambda^2)\times\left\{\Phi\left(\frac{b-\mu'}{\lambda'}\right)-\Phi\left(\frac{a-\mu'}{\lambda'}\right)\right\}} \\
&= K(Q) \times \mathrm{N}_1(\hat{\theta} \mid \mu, \sigma^2 + \lambda^2) \times \frac{\left\{\Phi\left(\frac{b-\mu'}{\lambda'}\right) - \Phi\left(\frac{a-\mu'}{\lambda'}\right)\right\}}{\left\{\Phi\left(\frac{b-\mu}{\lambda}\right) - \Phi\left(\frac{a-\mu}{\lambda}\right)\right\}}
\end{aligned}
$$

with pooled variance $\sigma^2 = (\sum_{i=1}^n w_i)^{-1}$, pooled estimate $\hat{\theta} = (\sum_{i=1}^n w_i \hat{\theta}_i)\sigma^2$, generalized Q-statistic $Q = \sum_{i=1}^n w_i(\hat{\theta}_i - \hat{\theta})^2$, updated variance $(\lambda')^2 = (\lambda^{-2} + \sigma^{-2})^{-1}$, and updated mean $\mu' = (\hat{\theta}\sigma^{-2} + \mu\lambda^{-2})(\lambda')^2$.

References

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.

Arroyo-Araujo, M., Voelkl, B., Laloux, C., Novak, J., Koopmans, B., Waldron, A.-M., Seiffert, I., Stirling, H., Aulehner, K., Janhunen, S. K., Ramboz, S., Potschka, H., Holappa, J., Fine, T., Loos, M., Boulanger, B., Würbel, H., and Kas, M. J. (2022). Systematic assessment of the replicability and generalizability of preclinical findings: Impact of protocol harmonization across laboratory sites. *PLOS Biology*, 20(11):e3001886. doi:10.1371/journal.pbio.3001886.

Baker, R. D. and Jackson, D. (2012). Meta-analysis inside and outside particle physics: two traditions that should converge? *Research Synthesis Methods*, 4(2):109–124. doi:10.1002/jrsm.1065.

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.

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

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

Camerer, C. F., Dreber, A., Forsell, E., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351:1433–1436. doi:10.1126/science.aaf0918.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B., et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behavior*, 2:637–644. doi:10.1038/s41562-018-0399-z.

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.

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.

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.

Follmann, D. (1996). A simple multivariate test for one-sided alternatives. *Journal of the American Statistical Association*, 91(434):854–861. doi:10.1080/01621459.1996.10476953.

Fraser, D. A. S. (2019). The *p*-value function and statistical inference. 73(sup1):135–147. doi:10.1080/00031305.2018.1556735.

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

Hartung, J., Knapp, G., and Sinha, B. K. (2008). *Statistical Meta-Analysis with Applications*. Wiley. doi:10.1002/9780470386347.

Hedges, L. V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Elsevier. doi:10.1016/c2009-0-03396-0.

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: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. (2023). Beyond the two-trials rule. doi:10.48550/ARXIV.2307.04548. arXiv preprint.

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-aoas1502.

Heller, R., Bogomolov, M., and Benjamini, Y. (2014). Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. *Proceedings of the National Academy of Sciences*, 111(46):16262–16267. doi:10.1073/pnas.1314814111.

Infanger, D. and Schmidt-Trucksäss, A. (2019). *P*-value functions: An underused method to present research results and to promote quantitative reasoning. 38(21):4189–4197. doi:10.1002/sim.8293.

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

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.

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

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

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

Mathur, M. B. and VanderWeele, T. 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.

Micheloud, C., Balabdaoui, F., and Held, L. (2023). Assessing replicability with the sceptical *p*-value: Type-i error control and sample size planning. *Statistica Neerlandica*. doi:10.1111/stan.12312.

Monin, B. and Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology*, 81(1):33–43. doi:10.1037/0022-3514.81.1.33.

National Academies of Sciences, Engineering, and Medicine (2019). *Reproducibility and Replicability in Science*. National Academies Press. doi:10.17226/25303.

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

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. (2023). Bayesian approaches to designing replication studies. *Psychological Methods*. doi:10.1037/met0000604.

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.

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

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

Rosenkranz, G. K. (2022). A generalization of the two trials paradigm. *Therapeutic Innovation & Regulatory Science*, 57(2):316–320. doi:10.1007/s43441-022-00471-4.

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.

Senn, S. (2007). *Statistical issues in drug development*. John Wiley & Sons, Chichester, England Hoboken, NJ.

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

Strack, F., Martin, L. L., and Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, 54(5):768–777. doi:10.1037/0022-3514.54.5.768.

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.

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D., and Salanti, G. (2015). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, 7(1):55–79. doi:10.1002/jrsm.1164.

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., Dijkstra, K., Fischer, A. H., Foroni, F., Hess, U., Holmes, K. J., Jones, J. L. H., Klein, O., Koch, C., Korb, S., Lewinski, P., Liao, J. D., Lund, S., Lupianez, J., Lynott, D., Nance, C. N., Oosterwijk, S., Ozdoğru, A. A., Pacheco-Unguetti, A. P., Pearson, B., Powis, C., Riding, S., Roberts, T.-A., Rumiati, R. I., Senden, M., Shea-Shumsky, N. B., Sobocko, K., Soto, J. A., Steiner, T. G., Talarico, J. M., van Allen, Z. M., Vandekerckhove, M., Wainwright, B., Wayand, J. F., Zeelenberg, R., Zetzer, E. E., and Zwaan, R. A. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, 11(6):917–928. doi:10.1177/1745691616674458.