Bayesian approaches to designing replication studies
Supplementary materials

Samuel Pawel⋆, Guido Consonni†, Leonhard Held⋆

⋆ Department of Biostatistics, University of Zurich

† Dipartimento di Scienze Statistiche, Università Cattolica del Sacro Cuore

E-mail: samuel.pawel@uzh.ch

November 3, 2022

In this document we provide additional information on computing the predictive distribution of the replication effect estimate when a prior is assigned to the heterogeneity variance τ^2 (Section 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 2 to 7). 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 8). 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 9).

1   Prior on the heterogeneity variance

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

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

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

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

To compute the marginal posterior density of τ^2 one numerical integration is hence required. The updating of the prior depends on the distance between prior mean μ_θ and the original effect estimate $\hat{\theta}_o$ relative to the prior variance σ_θ^2 and the squared standard error σ_o^2. If an improper uniform prior is assigned to θ

$(\sigma_\theta^2 \to \infty)$, the posterior reduces to the prior

$$\lim_{\sigma_\theta^2 \to \infty} f(\tau^2 \,|\, \hat{\theta}_o, \sigma_o) = \lim_{\sigma_\theta^2 \to \infty} \frac{f(\tau^2)\,\mathrm{N}(\hat{\theta}_o \,|\, \mu_\theta, \tau^2 + \sigma_o^2 + \sigma_\theta^2)}{\int_0^{+\infty} f(\tau_*^2)\,\mathrm{N}(\hat{\theta}_o \,|\, \mu_\theta, \tau_*^2 + \sigma_o^2 + \sigma_\theta^2)\,\mathrm{d}\tau_*^2}$$

$$= \lim_{\sigma_\theta^2 \to \infty} \int_0^{+\infty} \frac{f(\tau^2)}{f(\tau_*^2)} \underbrace{\sqrt{\frac{\tau_*^2 + \sigma_o^2 + \sigma_\theta^2}{\tau^2 + \sigma_o^2 + \sigma_\theta^2}}}_{\to 1} \exp\left[-\frac{1}{2}\left\{ \underbrace{\frac{(\hat{\theta}_o - \mu_\theta)^2}{\tau^2 + \sigma_o^2 + \sigma_\theta^2}}_{\downarrow 0} - \underbrace{\frac{(\hat{\theta}_o - \mu_\theta)^2}{\tau_*^2 + \sigma_o^2 + \sigma_\theta^2}}_{\downarrow 0} \right\} \right] \mathrm{d}\tau_*^2$$

$$= f(\tau^2),$$

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 τ^2 which intuitively makes sense as estimation of a variance requires at least two observations. The phenomenon is illustrated in Figure 1 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.
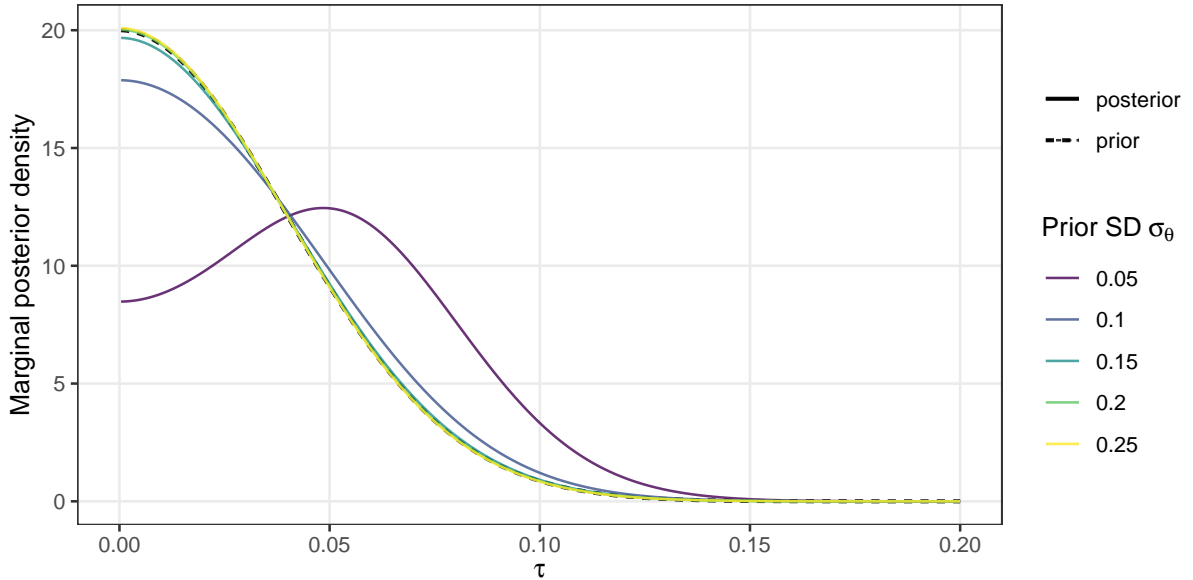

Figure 1: Marginal posterior distribution of heterogeneity variance τ^2 based on data from experiment "Labels" from Protzko et al. (2020) with original effect estimate $\hat{\theta}_o = 0.205$ and standard error $\sigma_o = 0.051$. A $\theta \sim \mathrm{N}(0, \sigma_\theta^2)$ prior is assigned to the effect size θ and a half normal prior with standard deviation 0.04 is assigned to τ.

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

$$\Pr(\hat{\theta}_r \in S \,|\, \hat{\theta}_o, \sigma_o, \sigma_r) = \int_S \int_0^{+\infty} f(\hat{\theta}_r \,|\, \hat{\theta}_o, \sigma_o, \sigma_r, \tau^2)\, f(\tau^2 \,|\, \hat{\theta}_o, \sigma_o)\, \mathrm{d}\hat{\theta}_r\, \mathrm{d}\tau^2$$

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

This mean computing the probability of replication success with a prior on τ^2 requires two-dimensional numerical integration. However, in the common case when a uniform prior is assigned to θ, the marginal posterior distribution of τ^2 reduces to the prior, and only one numerical integration is required.
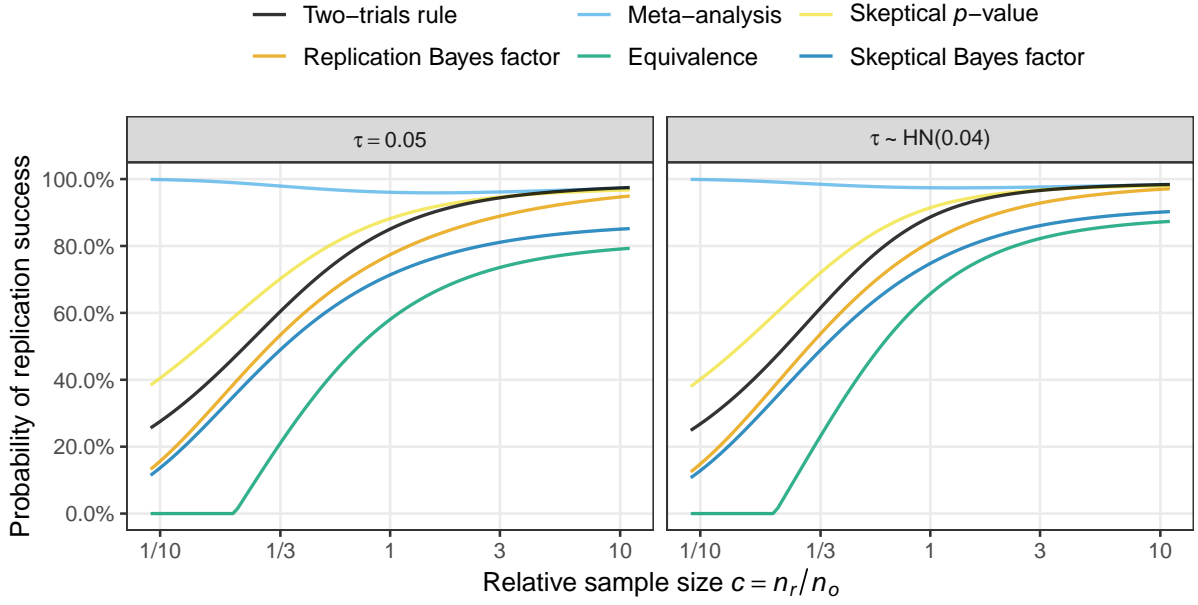
Figure 2: Probability of replication success as a function of relative sample size $c = n_r/n_o$ for experiment "Labels" with original effect estimate $\hat{\theta}_o = 0.205$ and standard error $\sigma_o = 0.051$ for uniform initial prior for effect size θ and either fixed $\tau = 0.05$ (as in main manuscript) or half normal prior with standard deviation 0.04 assigned to τ. 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 $\Delta = 0.2$, skeptical p-value at level $\alpha = 0.062$, and skeptical Bayes factor at level $\gamma = 1/10$.

Figure 2 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 τ which is a typical prior distribution used for heterogeneity modeling in meta-analysis (Röver 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.

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 α (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 $\hat{\theta}_i \,|\, \theta \sim \mathrm{N}(\theta, \sigma_i^2)$ with $\hat{\theta}_i$ the estimate of the unknown effect size θ from study i and σ_i is the corresponding standard error (assumed to be know). The p-values for testing $H_0\colon \theta = 0$ versus $H_1\colon \theta > 0$ are then $p_i = 1 - \Phi(\hat{\theta}_i/\sigma_i)$ whereas for the alternative $H_1\colon \theta < 0$ they are $p_i = \Phi(\hat{\theta}_i/\sigma_i)$. Suppose the original effect estimate was statistically significant at level α, i. e., $p_o \leq \alpha$. Replication success at level α is then established if the replication effect estimate $\hat{\theta}_r$ is also statistically significant at level α, 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

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

2.1   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 $\hat{\theta}_{r1}, \ldots, \hat{\theta}_m$ with standard errors $\sigma_{r1}, \ldots, \sigma_{rm}$. 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 $\sigma_{r*} = 1/\sqrt{\{\sum_i^m 1/(\sigma_{ri}^2 + \tau_r^2)\}}$ can be computed. If the between-replication heterogeneity variance τ_r^2 is set to zero this corresponds to the fixed effects estimate of θ, while estimating τ_r^2 from the data corresponds to the random effects estimate. Replication success at level α is then established if the replication p-value is smaller than α, 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

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

3   Fixed effects meta-analysis

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

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

which are also equivalent to the mean and standard deviation of a posterior distribution for the effect size θ based on the data from original and replication study and an initial flat prior for θ. Fixed effects meta-analysis is typically used because estimating a heterogeneity variance from two studies is highly unstable. Replication success at level α is established if the one-sided meta-analytic p-value (in the direction of the original effect estimate $\hat{\theta}$) is significant at level α, 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

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

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

$\hat{\theta}_i \mid \theta_i \sim \mathrm{N}(\theta_i, \sigma_i^2)$ 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 δ is then given by

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

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

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

5   The replication Bayes factor

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

$$\mathrm{BF_R} = \frac{f(\hat{\theta}_r \mid H_0)}{f(\hat{\theta}_r \mid H_1)} = \sqrt{1 + \sigma_o^2/\sigma_r^2} \exp\left[-\frac{1}{2}\left\{\frac{\hat{\theta}_r^2}{\sigma_r^2} - \frac{(\hat{\theta}_r - \hat{\theta}_o)^2}{\sigma_o^2 + \sigma_r^2}\right\}\right]. \tag{1}$$

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

$$S_{\mathrm{BF_R}} = \left(-\infty, -\sqrt{A} - (\hat{\theta}_o \sigma_r^2)/\sigma_o^2\right] \bigcup \left[\sqrt{A} - (\hat{\theta}_o \sigma_r^2)/\sigma_o^2, \infty\right)$$

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

5.1   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 $\hat{\boldsymbol{\theta}}_r = (\hat{\theta}_{r1}, \ldots, \hat{\theta}_{rm})^\top$ with corresponding standard error vector $\boldsymbol{\sigma}_r = (\sigma_{r1}, \ldots, \sigma_{rm})^\top$, and we assume the data model $\hat{\boldsymbol{\theta}}_r \mid \theta \sim \mathrm{N}_m\{\theta\, \mathbf{1}_m, \mathrm{diag}(\boldsymbol{\sigma}^2 + \tau_r^2\, \mathbf{1}_m)\}$ where $\mathbf{1}_m$ is a vector of m ones and τ_r^2 is a heterogeneity variance for the replication effect sizes (not to be confused with the heterogeneity variance τ^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 $H_0\colon \theta = 0$ relative to the alternative hypothesis $H_1\colon \theta \sim \mathrm{N}(\hat{\theta}_o, \sigma_o^2)$. The marginal density of the replication data under the null hypothesis is simply $\hat{\boldsymbol{\theta}}_r \mid H_0 \sim \mathrm{N}_m\{0\, \mathbf{1}_m, \mathrm{diag}(\boldsymbol{\sigma}^2 + \tau_r^2\, \mathbf{1}_m)\}$, whereas the marginal likelihood under the alternative H_1 is obtained from integrating the likelihood with respect to the prior distribution of θ under the alternative H_1. Let $\mathrm{N}(x; m, v)$ denote the normal density function mean m and variance v evaluated at x. Define also $\hat{\theta}_{r*} = \left\{\sum_{i=1}^n \hat{\theta}_{ri}/(\sigma_{ri}^2 + \tau_r^2)\right\}\sigma_{r*}^2$ and $\sigma_{r*}^2 = 1/\left\{\sum_{i=1}^n 1/(\sigma_{ri}^2 + \tau_r^2)\right\}$, i.e., the weighted average of the replication effect estimates based

on the heterogeneity τ_r^2 and its variance. The marginal density is then

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

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

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

The multisite replication Bayes factor is therefore equivalent to the singlesite replication Bayes factor from (1) but using the weighted average $\hat{\theta}_{r*}$ and its standard error σ_{r*} as the replication effect estimate $\hat{\theta}_r$ and standard error σ_r.

6   The skeptical *p*-value

Held (2020) proposed a reverse-Bayes approach for assessing replicability. One assumes again the data model $\hat{\theta}_i \mid \theta \sim \mathrm{N}(\theta, \sigma_i^2)$ with $i \in \{o, r\}$, along with a zero-mean "skeptical" prior $\theta \sim \mathrm{N}(0, \sigma_s^2)$ 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 skeptical" prior variance σ_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 skeptical prior variance σ_s^2 has the property that it renders the resulting posterior of θ no longer "credible" at level α, that is, the posterior tail probability is fixed to $\Pr(\theta \geq 0 \mid \hat{\theta}_o, \sigma_o, \sigma_s) = 1 - \alpha$ for positive estimates and $\Pr(\theta \leq 0 \mid \hat{\theta}_o, \sigma_o, \sigma_s) = 1 - \alpha$ for negative estimates. In a second step, the conflict between the skeptical 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

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

is computed and replication success at level α is declared if $p_{\text{Box}} \leq \alpha$. The smallest level α at which replication success is achieved is called the *the skeptical p-value* p_{S} and replication success at level α is equivalent with $p_{\text{S}} \leq \alpha$ (see Held, 2020; Held et al., 2022, 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

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

7 The skeptical 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 $\hat{\theta}_i \,|\, \theta \sim \mathrm{N}(\theta, \sigma_i^2)$ for study $i \in \{o, r\}$. In the first step the original data are used to contrast the evidence for the point null hypothesis $H_0 \colon \theta = 0$ relative to the "skeptical" alternative $H_S \colon \theta \sim \mathrm{N}(0, \sigma_s^2)$ with the Bayes factor

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

where $z_o = \hat{\theta}/\sigma_o^2$. One then determines the sufficiently skeptical prior variance σ_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 γ. The sufficiently skeptical prior variance can be computed by

$$
\sigma_s^2 = \begin{cases} -\dfrac{\hat{\theta}_o^2}{q} - \sigma_o^2 & \text{if } -\dfrac{\hat{\theta}_o^2}{q} \geq \sigma_o^2 \\ \text{undefined} & \text{else} \end{cases} \tag{2}
$$

$$
\text{where } q = \mathrm{W}_{-1}\left\{ -\frac{z_o^2}{\gamma^2} \exp\left(-z_o^2 \right) \right\} \tag{3}
$$

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

In a second step the conflict between the skeptical prior and the replication data is quantified. To do so, the skeptic is contrasted to the "advocacy" alternative $H_A \colon \theta \sim \mathrm{N}(\hat{\theta}_o, \sigma_o^2)$ 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 θ. This is done by computing the Bayes factor

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

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

condition for replication success at level γ, which leads to

$$
S_{\text{BF}_\text{S}} = \begin{cases}
(-\infty, -\sqrt{B} - M] \bigcup [\sqrt{B} - M, \infty) & \text{for } \sigma_s^2 < \sigma_o^2 \\
[\hat{\theta}_o - \{(\sigma_o^2 + \sigma_r^2) \log \gamma\}/\hat{\theta}_o, \infty) & \text{for } \sigma_s^2 = \sigma_o^2 \\
[-\sqrt{B} - M, \sqrt{B} - M] & \text{for } \sigma_s^2 > \sigma_o^2
\end{cases}
\tag{4}
$$

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 skeptical prior variance σ_s^2 computed by (2).

8 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

$$
m = K(K_c n_r + K_s)^{-1}.
\tag{5}
$$

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

$$
\sigma_{\hat{\theta}_{r*}}^2 = \frac{\sigma_r^2 + \tau^2}{m} + \frac{\tau^2 + \sigma_o^2}{1 + 1/g}.
\tag{6}
$$

Plugging in (5) into (6) 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.

9 Sample size determination with adaptive shrinkage prior

Figure 3 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 θ 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 θ 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.
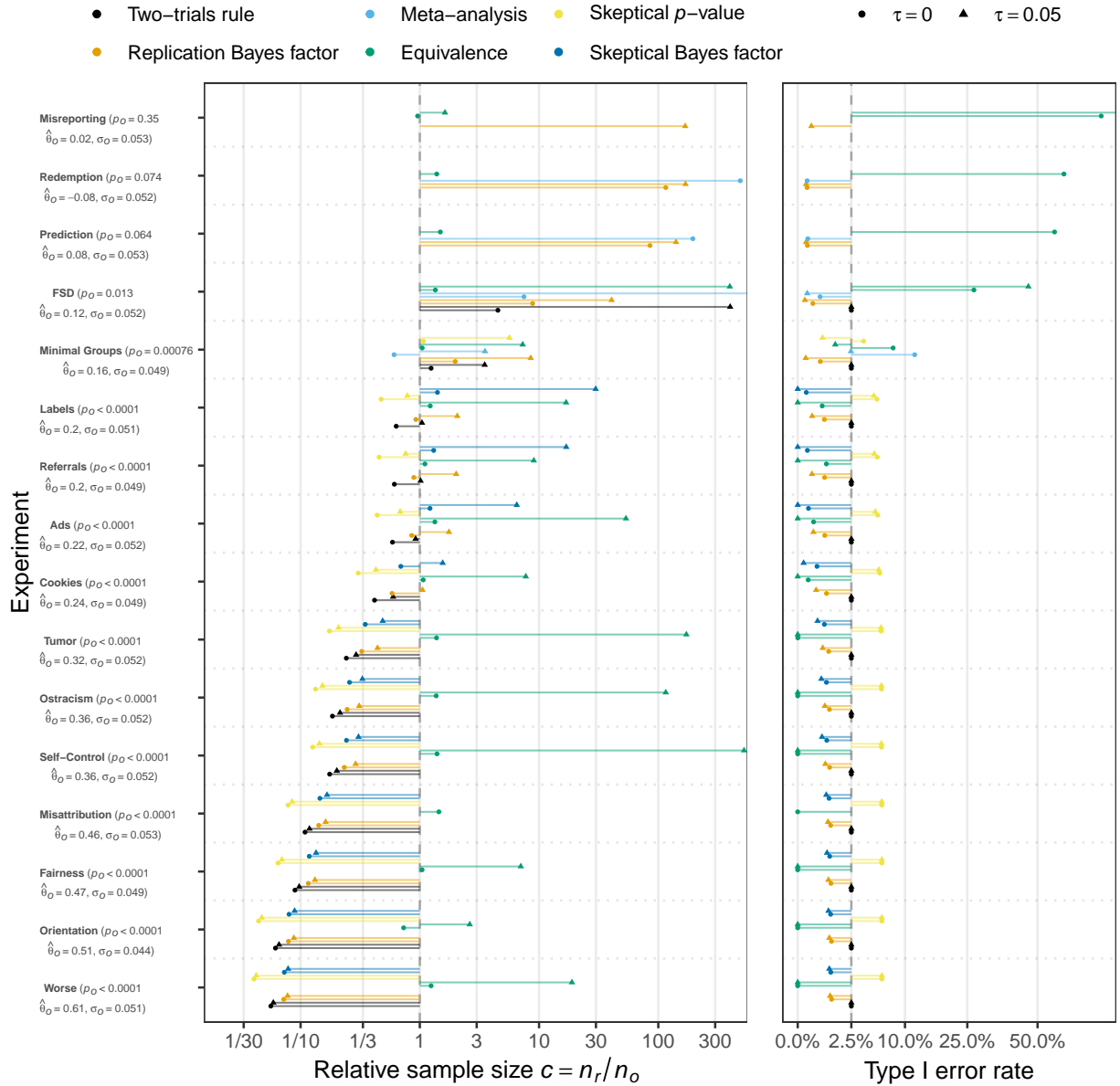
Figure 3: The left plot shows the required relative sample size $c = n_r/n_o$ to achieve a target probability of replication success of $1 - \beta = 80\%$ (if possible). Replication success is defined through the two-trials rule at level $\alpha = 0.025$, replication Bayes factor at level $\gamma = 1/10$, fixed effects-meta analysis at level $\alpha = 0.025^2$, effect size equivalence at level $\alpha = 0.1$ with margin $\Delta = 0.2$, skeptical p-value at level $\alpha = 0.062$, and skeptical 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 θ 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 $p_o = 1 - \Phi(|\hat{\theta}_o|/\sigma_o)$.

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.

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., 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.

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

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795. doi:10.1080/01621459.1995.10476572.

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.

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.

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.

Computational details

```
sessionInfo()

## R version 4.2.1 (2022-06-23)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.5 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.9.0
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.9.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=de_CH.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=de_CH.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=de_CH.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=de_CH.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] BayesRepDesign_0.42   BayesRep_0.1          ggpubr_0.4.0
##  [4] colorspace_2.0-3      scales_1.2.1          ggplot2_3.3.6
##  [7] tidyr_1.2.0           dplyr_1.0.9           ReplicationSuccess_1.2
## [10] knitr_1.39
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.9         pillar_1.8.1      compiler_4.2.1     tools_4.2.1
##  [5] digest_0.6.29      viridisLite_0.4.1 evaluate_0.15      lifecycle_1.0.3
##  [9] tibble_3.1.8       gtable_0.3.1      pkgconfig_2.0.3    rlang_1.0.6
## [13] DBI_1.1.2          cli_3.4.1         xfun_0.31          gridExtra_2.3
## [17] withr_2.5.0        stringr_1.4.0     generics_0.1.2     vctrs_0.4.2
## [21] cowplot_1.1.1      grid_4.2.1        tidyselect_1.1.2   glue_1.6.2
## [25] R6_2.5.1           lamW_2.1.1        rstatix_0.7.0      fansi_1.0.3
## [29] carData_3.0-5      farver_2.1.1      purrr_0.3.4        car_3.0-13
## [33] magrittr_2.0.3     ellipsis_0.3.2    codetools_0.2-18   backports_1.4.1
## [37] assertthat_0.2.1   abind_1.4-5       ggsignif_0.6.3     labeling_0.4.2
## [41] utf8_1.2.2         stringi_1.7.6     RcppParallel_5.1.5 munsell_0.5.0
## [45] broom_0.8.0
```