Bayesian approaches to designing replication studies

Samuel Pawel[⋆]**, Guido Consonni**[†]**, and Leonhard Held**[⋆]
⋆ Department of Biostatistics, University of Zurich
† Dipartimento di Scienze Statistiche, Universita Cattolica del Sacro Cuore
E-mail: samuel.pawel@uzh.ch

October 7, 2022
This is a preprint which has not yet been peer reviewed.

Abstract

Replication studies are essential to assess the credibility of claims from original studies. A critical aspect of designing replication studies is determining their sample size; a too small sample size may lead to inconclusive studies whereas a too large sample size may waste resources that could be allocated better in other projects. Here we show how Bayesian approaches help to efficiently determine the replication sample size. The approach allows researchers to combine the original data and external knowledge in a design prior for the underlying parameters. We investigate design priors in the normal normal hierarchical model where analytical results are available. Based on a design prior, predictions about the replication data can be made, and the replication sample size can be chosen to ensure a sufficiently high probability of replication success. Replication success may be defined through Bayesian or non-Bayesian criteria, and different criteria may also be combined to meet distinct stakeholders and allow conclusive inferences based on multiple analysis approaches. An application to data from a multisite replication project illustrates how the approach helps to design informative and cost-effective replication studies. The methods are made available in an R package.

Keywords: Bayesian design, design prior, multisite replication, sample size determination

1 Introduction

The replicability of research findings is a cornerstone for the credibility of science. However, there is growing evidence that the replicability of many scientific findings is lower than expected (Ioannidis, 2005; Open Science Collaboration, 2015; Camerer et al., 2018; Errington et al., 2021). This "replication crisis" has led to methodological reforms in various fields of science, one of which is an increased conduct of replication studies (Munafò et al., 2017). Statistical methodology plays a key role in the evaluation of replication studies, and various methods have been proposed for quantifying how "successful" a replication study was in replicating the original finding (Bayarri and Mayoral, 2002; Verhagen and Wagenmakers, 2014; Simonsohn, 2015; Anderson and Maxwell, 2016; Patil et al., 2016; Johnson et al., 2016; Etz and Vandekerckhove, 2016; van Aert and van Assen, 2017; Ly et al., 2018; Harms, 2019; Hedges and Schauer, 2019; Mathur and VanderWeele, 2020; Held, 2020; Pawel and Held, 2020; Held et al., 2022b; Pawel and Held, 2022, among others). Yet, as with ordinary studies, statistical methodology is not only important for analyzing replication studies but also for designing them, in particular for their *sample size determination* (SSD). Optimal SSD is important since too small sample sizes may lead to inconclusive studies, whereas too large sample sizes may waste resources which could have been allocated better in other research projects.

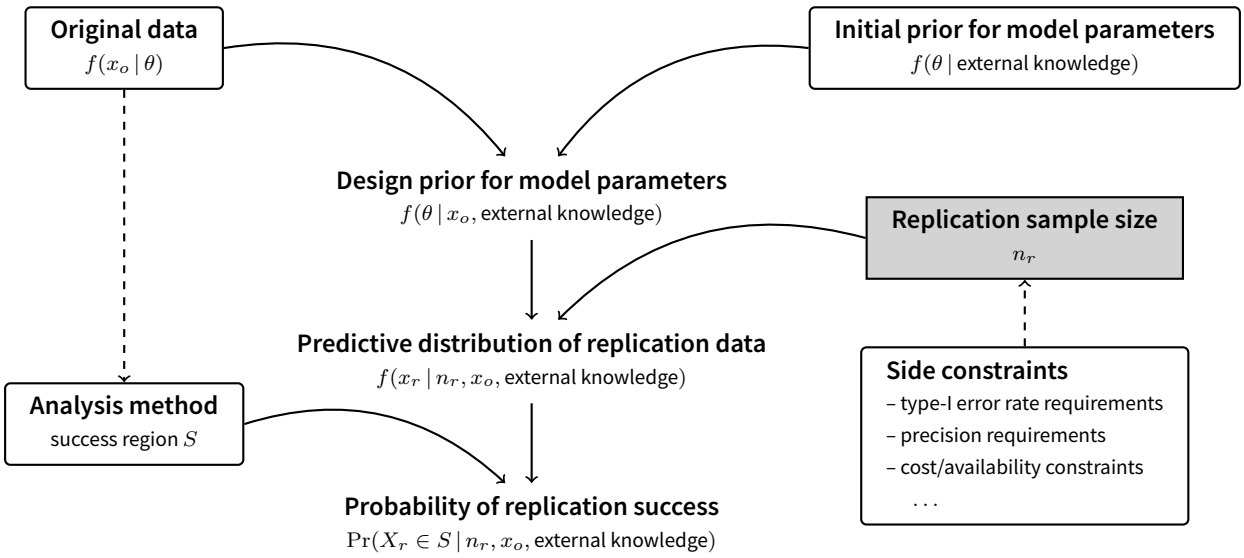


Figure 1: Schematic illustration of the process of Bayesian sample size determination for replication studies.

SSD for replication studies comes with unique opportunities and challenges; the data from the original study can be used to inform SSD, at the same time the analysis of the replication data is often different from an analysis of a study in isolation. For these reasons, a relatively small literature has emerged which specifically deals with replication study SSD for selected analysis methods and data models. For instance, SSD for standardized mean difference effect sizes analyzed with Bayes factors (Bayarri and Mayoral, 2002), SSD for statistical significance assessment of the replication (Goodman, 1992; Senn, 2002; Micheloud and Held, 2022; van Zwet and Goodman, 2022), SSD for reverse-Bayes assessment of the replication (Held, 2020; Pawel and Held, 2022), or SSD for meta-analysis of replication studies (Hedges and Schauer, 2021). The aim of this paper is to unify these methods under a general framework. The proposed framework (schematically illustrated in Figure 1) applies to any kind of data model and analysis method, and is based on principles from Bayesian design analysis (Spiegelhalter, 1986; Spiegelhalter et al., 1986; Spiegelhalter and Freedman, 1986; Weiss, 1997; O'Hagan and Stevens, 2001; De Santis, 2004; Spiegelhalter et al., 2004; Schönbrodt and Wagenmakers, 2017; Grieve, 2022). The design of replication studies is a natural candidate for Bayesian knowledge updating. Specifically, the Bayesian framework allows to combine uncertain information from different sources –for instance, the data from the original study and/or expert knowledge– in a so-called *design prior* for the underlying model parameters. Based on the design prior, predictions about the replication data can be made, and the sample size can be chosen such that the probability of replication success becomes sufficiently high. Importantly, Bayesian design analysis can also be used if the planned analysis of the replication study is non-Bayesian, which is the more common situation in practice.

This paper is structured as follows: We start with presenting a general framework for Bayesian SSD of replication studies (Section 2). We then investigate design priors and sample size determination in the normal normal hierarchical model and for several Bayesian and non-Bayesian analysis methods (Section 3). As a running example, we use data from a cross-laboratory replication project (Protzko et al., 2020). We then close with concluding remarks and practical recommendations (Section 4).

2 General framework

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

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

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

The initial prior of θ may depend on external knowledge (e. g., data from other studies), we will discuss common types of external knowledge in the replication setting in Section 3. The design prior (1) hence represents the state of knowledge and uncertainty about the parameter θ before the replication is conducted, and, along with an assumed replication sample size n_r, it can be used to compute a predictive distribution for the replication data

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

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

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

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

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

3 Sample size determination in the normal normal hierarchical model

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

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

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

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

$$\hat{\theta}_k \,|\, \theta_k \sim \mathrm{N}(\theta_k, \sigma_k^2) \tag{4a}$$
$$\theta_k \,|\, \theta \sim \mathrm{N}(\theta, \tau^2) \tag{4b}$$
$$\theta \sim \mathrm{N}(\mu_\theta, \sigma_\theta^2). \tag{4c}$$

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

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

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

3.1 Design prior and predictive distribution

The observed original data $x_o = \{\hat{\theta}_o, \sigma_o\}$ can be combined with the initial prior by Bayes' theorem (1) to obtain a posterior distribution for the effect size θ

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

where $g = \sigma_\theta^2/(\sigma_o^2+\tau^2)$ is the *relative prior variance*. This posterior serves then as the design prior for predicting the replication data. Specifically, assuming a replication standard error σ_r and integrating the marginal density of the replication effect estimate (24a) with respect to the design prior (6) leads then to the predictive distribution

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

Both the design prior (1) and the predictive distribution (7) depend on the parameters of the initial prior (τ^2, μ_θ, σ_θ^2). In the following, we will explain how these parameters can be specified based on external knowledge.

3.2 Incorporating external knowledge in the initial prior

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

3.2.1 Between-study heterogeneity

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

If heterogeneity is expected, there are different approaches for specifying τ^2; A domain expert may subjectively assess how much heterogeneity is to be expected due to the change in laboratory, study population, and other factors. An alternative is to take an estimate from the literature, e. g., from multisite replication projects or from systematic reviews. Finally, one can also specify an upper limit of "tolerable heterogeneity". This approach is similar to specifying a minimal clinically relevant difference in classical power analysis in the sense that a true replication effect size which is intolerably heterogeneous from the original effect size is not relevant to be detected. An absolute (Spiegelhalter et al., 2004, Chapter 5.7.3) and a relative approach (Held and Pawel, 2020) can be considered. In the absolute approach, a value of τ^2 is chosen such that a suitable range (e. g., the IQR or the range from 2.5% to 97.5% of the distribution (4b)) of study-specific effect sizes is not larger than an effect size difference considered negligible. For example, when 95% of the effect sizes should not vary more than a small effect size e. g., $d = 0.2$ on standardized mean difference scale based on the Cohen (1992) effect size classification, this would lead to $\tau = d/(2 \cdot 1.96) \approx 0.05$. In the relative approach, τ^2 is specified relative to the variance of the original estimate σ_o^2 using field conventions for tolerable relative heterogeneity. For example, in the

Cochrane guidelines for systematic reviews (Deeks et al., 2019) a value of $I^2 = \tau^2/(\tau^2 + \sigma_o^2) = 40\%$ is classified as "negligible", which translates to $\tau^2 = \sigma_o^2/(1/I^2 - 1) = (2\sigma_o^2)/3$.

We note that in principle it is also possible to assign a prior distribution to τ^2 (see the literature from meta-analysis on this issue e. g., Röver et al., 2021). However, for interpretability reasons we will not consider such an approach here as there are no closed-form expressions anymore for the predictive distribution and the probability of replication success.

3.2.2 Knowledge about the effect size

Prior knowledge about the effect size θ can be incorporated via the prior mean μ_θ and prior variance σ_θ^2 in (4c). For instance, the parameters may be specified based on a meta-analysis of related studies or based on expert elicitation. The resulting design prior will then contain more information than what was provided by the original data alone, leading to potentially more efficient designs. If there is no prior knowledge available, one can specify an uninformative initial prior by letting the variance go to infinity ($\sigma_\theta^2 \to \infty$). The resulting design prior will then only contain the information from the original study.

3.2.3 Exaggerated original results

Potentially exaggerated original results can be counteracted by setting $\mu_\theta = 0$ to obtain a shrinkage prior which shrinks the design prior towards less impressive effect sizes than the observed one. For instance, Replicators could believe that the results from the original study are exaggerated because there is no pre-registered study protocol available. Even without such beliefs, weakly informative shrinkage priors may also be motivated from a "regularization" point of view as they will block physically impossible parameter values from taking over the posterior in settings with uninformative data (Gelman, 2009).

The amount of shrinkage is determined via the prior variance σ_θ^2. A diffuse prior ($\sigma_\theta^2 \to \infty$) will lead to no shrinkage, while a highly concentrated prior ($\sigma_\theta^2 \downarrow 0$) will completely shrink the design prior to a point mass the zero. In practice, a pragmatic option with good predictive properties is to use the empirical Bayes estimate based on the original data

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

This choice will lead to adaptive shrinkage (Pawel and Held, 2020) in the sense that shrinkage is large for unconvincing original studies (those with small effect estimates $\hat{\theta}_o$ and/or large standard errors σ_o), but disappears as the data become more convincing (through larger effect estimates $\hat{\theta}_o$ and/or smaller standard errors σ_o). Another option is to use an estimate from a corpus of related studies (e. g., the Cochrane library of systematic reviews as in van Zwet et al., 2021).

3.3 Example: Cross-laboratory replication project

We will now illustrate the construction of design priors based on data from a recently conducted replication project (Protzko et al., 2020), see Figure 2 for a summary of the data. The data were collected in four laboratories. Each of them conducted four original studies and for each original study four replication studies were carried out, one by the same lab and three by the other three labs.

Most studies used simple between-subject designs with two groups and a continuous outcome so that for a study $i \in \{o, r\}$ the standardized mean difference (SMD) effect estimate $\hat{\theta}_i$ can be computed from

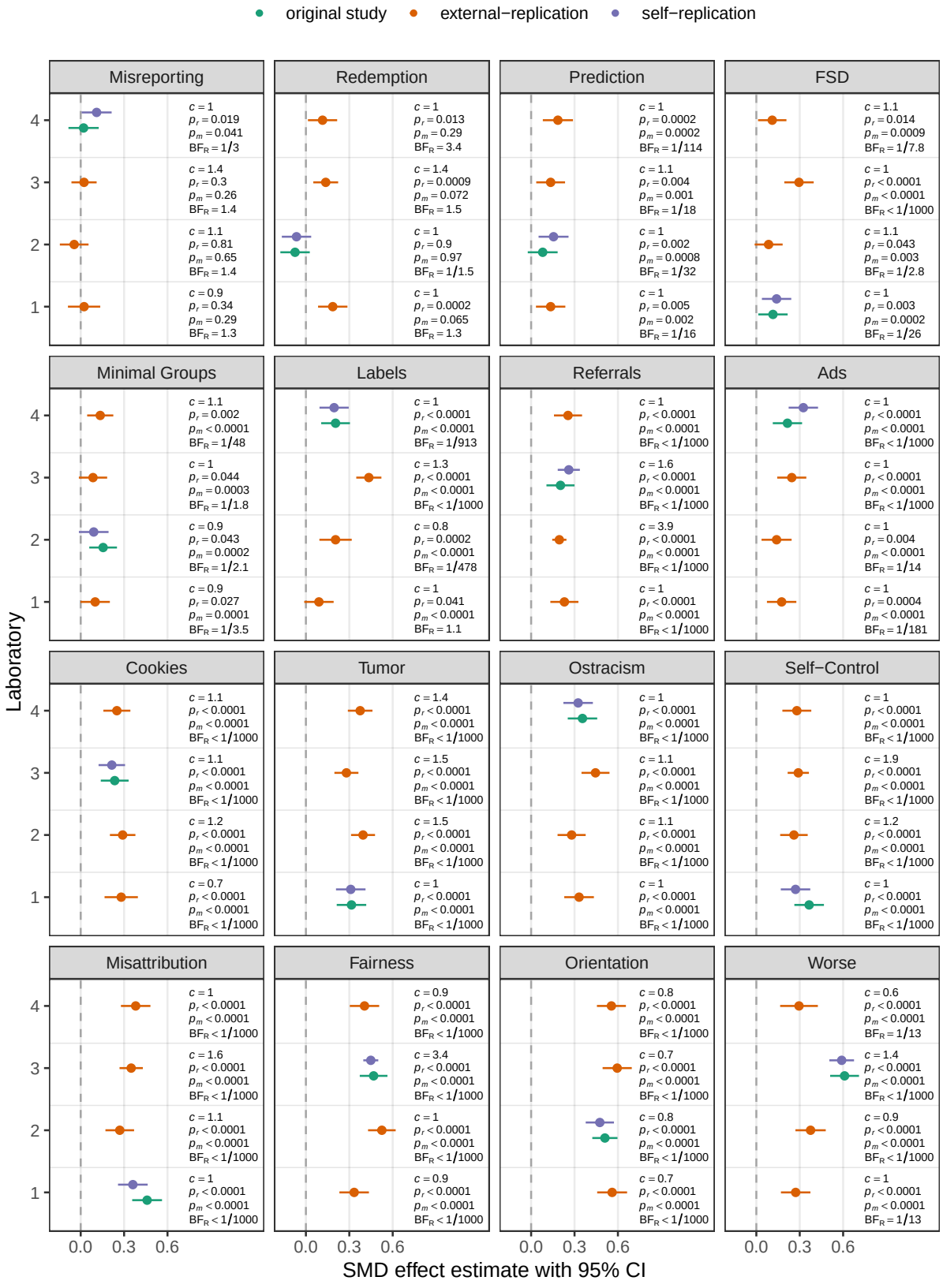


Figure 2: Data from cross-laboratory replication project by Protzko et al. (2020). Shown are standardized mean difference (SMD) effect estimates with 95% confidence intervals stratified by experiment and laboratory. For each replication study, the relative sample size $c = n_r/n_o$, the one-sided replication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

the group means $\bar{y}_{i1}, \bar{y}_{i2}$, group standard deviations s_{i1}, s_{i2}, and group sample sizes n_{i1}, n_{i2} by

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

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

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

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

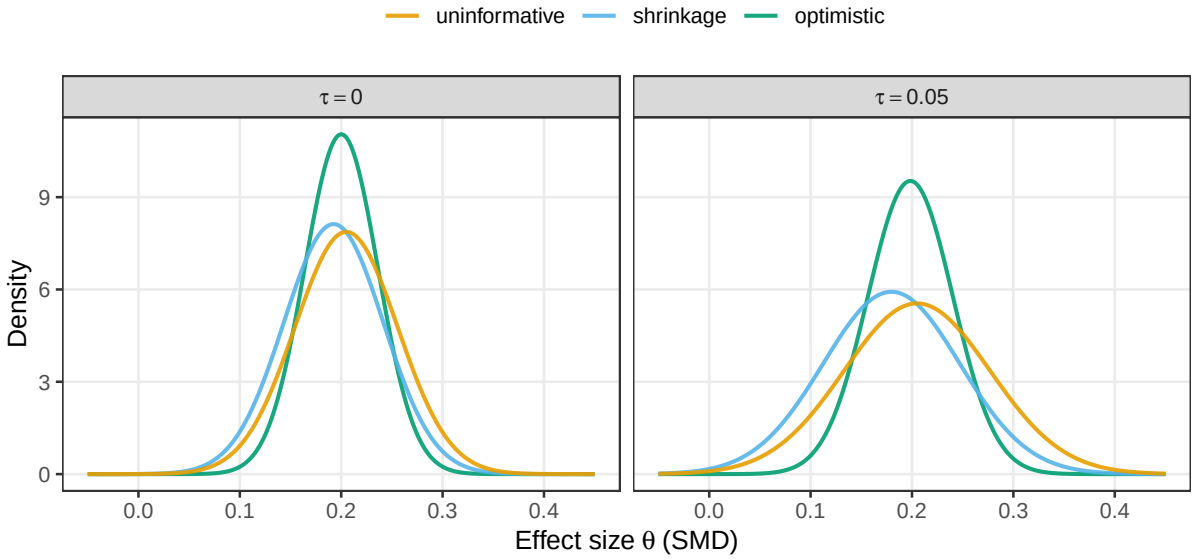


Figure 3: Design priors for the effect size θ (SMD) in the experiment "Labels" based on the original effect estimate $\hat{\theta}_o = 0.2$ with standard error $\sigma_o = 0.05$. Shown are different choices of the initial prior for θ and the between-study heterogeneity τ.

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

Concerning the specification of between-study heterogeneity, a distinction needs to be made between replications which are conducted in the same lab as the original study (*self-replications*) and replications which are conducted in a different lab (*external-replications*). For self-replications it is reasonable to set $\tau^2 = 0$ because we would expect no between-study heterogeneity as the experimental conditions will be nearly identical in both studies. In contrast, one would expect some between-study heterogeneity

for external-replications as the experimental conditions may slightly differ between the labs. In the following, we will use $\tau^2 = 0.05$ elicited via the "absolute" approach as discussed in Section 3.2.1, since it is independent of the sample size of the original study.

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

If there would have been good reason to believe that the original result may have been exaggerated, we might have specified an initial shrinkage prior. For instance, using the empirical Bayes estimate (8) for the prior variance leads to a prior with shrinkage factor $\hat{g}/(1+\hat{g}) = 0.88$ for an external-replication. The mean and variance are then shrunken towards zero by 12% compared to the mean and variance of the design prior based on the uninformative initial prior (the blue lines in Figure 3). Conversely, if we had prior knowledge about the effect size θ from another study, we could have specified an initial optimistic prior. Suppose, for instance, that the self-replication of the experiment "Labels" was a pilot study, and its effect estimate $\hat{\theta}_p = 0.195$ and standard error $\sigma_p = 0.05$ were available to us. This would lead to a design prior centered around the weighted mean of original and pilot study, as well as, a prior precision equal to the sum of the precision of both estimates (green lines in Figure 3). Due to incorporation of the external data, this design prior is much more concentrated than the other two.

3.4 Probability of replication success and required sample size

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

3.4.1 The two-trials rule

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

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

$$S_{\text{2TR}} = [z_\alpha \, \sigma_r, \infty) \,, \tag{10}$$

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

$$\Pr(\hat{\theta} \in S_{\text{2TR}} \,|\, \hat{\theta}_o, \sigma_o, \sigma_r) = \Phi\left(\frac{\mu_{\hat{\theta}_r} - z_\alpha \, \sigma_r}{\sqrt{\sigma_r^2 + \tau^2 + (\sigma_o^2 + \tau^2)/(1 + 1/g)}}\right) \tag{11}$$

with $\Phi(\cdot)$ the standard normal cumulative distribution function and $\mu_{\hat{\theta}_r}$ the mean of the predictive distribution (7). Importantly, by decreasing the standard error σ_r (through increasing the sample size n_r), the probability of replication success (11) cannot become arbitrarily large but is bounded by

$$\text{limP}_{\text{2TR}} = \Phi\left(\frac{\mu_{\hat{\theta}_r}}{\sqrt{\tau^2 + (\sigma_o^2 + \tau^2)/(1 + 1/g)}}\right). \tag{12}$$

The required replication standard error σ_r^* to achieve replication success with probability $1-\beta < \text{limP}_{\text{2TR}}$ can now be obtained by equating (11) to $1 - \beta$ and solving for σ_r. This leads to

$$\sigma_r^* = \frac{\mu_{\hat{\theta}_r} z_\alpha - z_\beta \sqrt{(z_\alpha^2 - z_\beta^2)\left\{\tau^2 + (\sigma_o^2 + \tau^2)/(1 + 1/g)\right\} + \mu_{\hat{\theta}_r}^2}}{z_\alpha^2 - z_\beta^2} \tag{13}$$

for $\alpha < \beta$. The standard error σ_r^* can subsequently be translated in a sample size. The translation depends on the type of effect size, for instance, for SMD effect sizes we can use the approximation $n_r^* \approx \lceil 4/(\sigma_r^*)^2 \rceil$ from Section 3.3.

3.4.2 Fixed effects meta-analysis

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

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

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

$$S_{\text{MA}} = \left[\sigma_r z_\alpha \sqrt{1 + \sigma_r^2/\sigma_o^2} - (\hat{\theta}_o \sigma_r^2)/\sigma_o^2, \infty\right) \tag{14}$$

then corresponds to both replication success defined via a one-sided meta-analytic p-value being smaller than level α, i. e., $p_m = 1 - \Phi(\hat{\theta}_m/\sigma_m) \leq \alpha$, or to replication success defined via a Bayesian posterior probability $\Pr(\theta > 0 \,|\, \hat{\theta}_o, \hat{\theta}_r, \sigma_o, \sigma_r) \geq 1 - \alpha$. From the success region (14) and an assumed standard error σ_r the probability of replication success can be computed by

$$\Pr(\hat{\theta} \in S_{\text{MA}} \,|\, \hat{\theta}_o, \sigma_o, \sigma_r) = \Phi\left(\frac{\mu_{\hat{\theta}_r} - \sigma_r z_\alpha \sqrt{1 + \sigma_r^2/\sigma_o^2} + (\hat{\theta}_o \sigma_r^2)/\sigma_o^2}{\sqrt{\sigma_r^2 + \tau^2 + (\sigma_o^2 + \tau^2)/(1 + 1/g)}}\right). \tag{15}$$

As for the two-trials rule, by decreasing the standard error σ_r the probability (17) can at most become as large as limP_{2TR} from (12).The required standard error σ_r^* to achieve a desired probability of replication success $1 - \beta < \text{limP}_{\text{2TR}}$ can be computed numerically using root finding algorithms.

3.4.3 Effect size equivalence test

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

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

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

$$S_{\text{E}} = \left[\hat{\theta}_o - \Delta + z_{\alpha/2}\sqrt{\sigma_o^2 + \sigma_r^2}, \hat{\theta}_o + \Delta - z_{\alpha/2}\sqrt{\sigma_o^2 + \sigma_r^2}\right]. \tag{16}$$

Wether or not replication success is achievable depends on the standard error of the original estimate σ_o, and in particular when the margin is chosen too small ($\Delta \leq z_{\alpha/2}\sigma_o$), replication success is impossible regardless of how small the replication standard error σ_r will be. Assuming now that the margin is large enough ($\Delta > z_{\alpha/2}\sigma_o$), the probability of replication success can be computed by

$$\begin{aligned}\Pr(\hat{\theta} \in S_{\text{E}} \,|\, \hat{\theta}_o, \sigma_o, \sigma_r) = \Phi&\left(\frac{\hat{\theta}_o + \Delta - z_{\alpha/2}\sqrt{\sigma_o^2 + \sigma_r^2} - \mu_{\hat{\theta}_r}}{\sqrt{\sigma_r^2 + \tau^2 + (\sigma_o^2 + \tau^2)/(1 + 1/g)}}\right) \\ &- \Phi\left(\frac{\hat{\theta}_o - \Delta + z_{\alpha/2}\sqrt{\sigma_o^2 + \sigma_r^2} - \mu_{\hat{\theta}_r}}{\sqrt{\sigma_r^2 + \tau^2 + (\sigma_o^2 + \tau^2)/(1 + 1/g)}}\right).\end{aligned} \tag{17}$$

As with the previous methods, the probability cannot be made arbitrarily large by decreasing the replication standard error σ_r, but is bounded by

$$\text{lim}\text{P}_{\text{E}} = \Phi\left(\frac{\hat{\theta}_o + \Delta - z_{\alpha/2}\sigma_o - \mu_{\hat{\theta}_r}}{\sqrt{\tau^2 + (\sigma_o^2 + \tau^2)/(1 + 1/g)}}\right) - \Phi\left(\frac{\hat{\theta}_o - \Delta + z_{\alpha/2}\sigma_o - \mu_{\hat{\theta}_r}}{\sqrt{\tau^2 + (\sigma_o^2 + \tau^2)/(1 + 1/g)}}\right). \tag{18}$$

The required replication standard error σ_r to achieve a desired power of replication success $1-\beta < \text{lim}\text{P}_{\text{E}}$ can again be computed numerically.

3.4.4 The replication Bayes factor

A Bayesian hypothesis testing approach for assessing replication success was proposed by Verhagen and Wagenmakers (2014). They define a "replication Bayes factor"

$$\text{BF}_{\text{R}} = \frac{f(x_r \,|\, H_0)}{f(x_r \,|\, H_1)}$$

which is ratio of likelihoods of the replication data x_r under the null hypothesis $H_0 \colon \theta = 0$ and under the alternative hypothesis $H_1 \colon \theta \sim f(\theta \,|\, x_o)$, that is the posterior of the effect size θ based on the original data x_o. Bayes factor values $\text{BF}_{\text{R}} < 1$ indicate replication success, the smaller BF_{R} the higher the degree of replication success. Under normality and assuming no heterogeneity the success region for achieving $\text{BF}_{\text{R}} \leq \gamma$ is given by

$$S_{\text{BF}_{\text{R}}} = \left(-\infty, -\sqrt{A} - (\hat{\theta}_o\sigma_r^2)/\sigma_o^2\right] \bigcup \left[\sqrt{A} - (\hat{\theta}_o\sigma_r^2)/\sigma_o^2, \infty\right) \tag{19}$$

with $A = \sigma_r^2(1 + \sigma_r^2/\sigma_o^2)\{\hat{\theta}_o^2/\sigma_o^2 - 2\log\gamma + \log(1 + \sigma_o^2/\sigma_r^2)\}$ (Pawel and Held, 2022). Consequently, the probability of replication success can be computed by

$$\begin{aligned}\Pr(\hat{\theta} \in S_{\text{BF}_\text{R}} \mid \hat{\theta}_o, \sigma_o, \sigma_r) = \Phi\left(\frac{\mu_{\hat{\theta}_r} - \sqrt{A} + (\hat{\theta}_o\sigma_r^2)/\sigma_o^2}{\sqrt{\sigma_r^2 + \tau^2 + (\sigma_o^2 + \tau^2)/(1 + 1/g)}}\right) \\ + \Phi\left(\frac{-\sqrt{A} - (\hat{\theta}_o\sigma_r^2)/\sigma_o^2 - \mu_{\hat{\theta}_r}}{\sqrt{\sigma_r^2 + \tau^2 + (\sigma_o^2 + \tau^2)/(1 + 1/g)}}\right)\end{aligned} \tag{20}$$

which is, as with the other methods, bounded from above by a constant $\text{limP}_{\text{BF}_\text{R}} = \lim_{\sigma_r \downarrow 0} \Pr(\hat{\theta} \in S_{\text{BF}_\text{R}} \mid \hat{\theta}_o, \sigma_o, \sigma_r)$. Root finding algorithms can be used to numerically determine the required standard error for achieving a probability of replication success $1 - \beta < \text{limP}_{\text{BF}_\text{R}}$.

3.4.5 The sceptical *p*-value

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

$$S_{p_\text{S}} = \left[z_\alpha\sqrt{\sigma_r^2 + \frac{\sigma_o^2}{(z_o^2/z_\alpha^2) - 1}}, \infty\right). \tag{21}$$

From the success region (21) the probability of replication success at level α can be computed by

$$\Pr(\hat{\theta} \in S_{p_\text{S}} \mid \hat{\theta}_o, \sigma_o, \sigma_r) = \Phi\left(\frac{\mu_{\hat{\theta}_r} - z_\alpha\sqrt{\sigma_r^2 + \sigma_o^2/\{(z_o^2/z_\alpha^2) - 1\}}}{\sqrt{\sigma_r^2 + \tau^2 + (\sigma_o^2 + \tau^2)/(1 + 1/g)}}\right). \tag{22}$$

As for the two-trials rule, the required standard error σ_r^* to achieve a probability of replication success $1 - \beta < \text{limP}_{p_\text{S}}$ can be computed analytically by

$$\sigma_r^* = \sqrt{x^2 - \frac{\sigma_o^2}{(z_o/z_\alpha)^2 - 1}} \tag{23}$$

with

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

3.4.6 The sceptical Bayes factor

Pawel and Held (2022) modified the sceptical p-value method, using Bayes factors instead of tail probabilities as measures of evidence and prior data conflict. Again, the reverse-Bayes procedure can be summarized with a single measure termed the "sceptical Bayes factor" BF_R, we refer to Pawel and Held (2022) for details, and to the supplementary for the derivation of success region and probability of replication success. The required standard error can again be computed numerically.

3.5 Example: Cross-laboratory replication project (continued)

We will now revisit the experiment "Labels" from Section 3.3 and compute the probability of replication success and replication sample sizes based on the previously discussed analysis methods. The parameters of the methods are specified as follows: For the two-trials rule we use the conventional one-sided significance level $\alpha = 0.025$, while for meta-analysis we use the more stringent level $\alpha = 0.025^2$ as the method is based on two data sets rather than one. Likewise we use a $1 - \alpha = 90\%$ confidence interval conventionally used in equivalence testing along with a margin $\Delta = 0.2$ corresponding to a small SMD effect size according to the classification from Cohen (1992). For the sceptical p-value we use the recommended "golden" level $\alpha = 0.062$ as it guarantees that for original studies which where just significant at $\alpha = 0.025$ replication success is only possible if the replication effect estimate is larger than the original one (Held et al., 2022b). Finally, for the Bayes factor methods we use the "strong evidence" level $\gamma = 1/10$ from Jeffreys (1961).

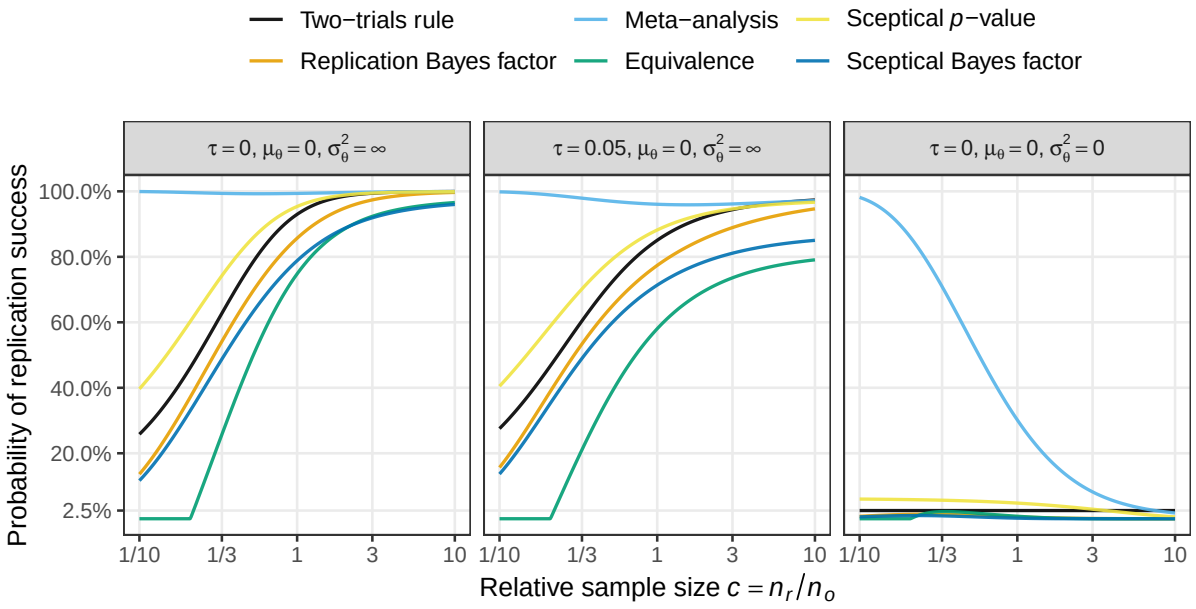


Figure 4: Probability of replication success as a function of relative sample size $c = n_r/n_o$ for experiment "Labels" with original effect estimate $\hat{\theta}_o = 0.2$ and standard error $\sigma_o = 0.05$ for different initial prior parameters $(\tau, \mu_\theta, \sigma_\theta^2)$. Replication 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 at level $\alpha = 0.1$ with margin $\Delta = 0.2$, sceptical p-value at level $\alpha = 0.062$, and sceptical Bayes factor at level $\gamma = 1/10$.

Figure 4 shows the probability of replication success as a function of the relative sample size $c = n_r/n_o$ and for different initial priors. The left and middle plot are based on an uninformative prior for the effect size ($\sigma_\theta^2 = \infty$) without heterogeneity ($\tau^2 = 0$) and with heterogeneity ($\tau^2 = 0.05$), respectively, whereas the right plot shows the prior corresponding to the null hypothesis $H_0 \colon \theta = 0, \tau^2 = 0$, so that the probability of replication success is the type I error rate.

We see from the left and middle plots in Figure 4 that increasing the relative sample size monotonically increases the probability of replication success for all methods but meta-analysis. Meta-analysis shows a non-monotone behavior because the original study was already highly significant so that the pooled effect estimate is significant even for very small replications, and the uncertainty regarding the replication effect estimate $\hat{\theta}_r$ may therefore even reduce the probability of replication success. If heterogeneity is taken into account (middle plot), the probability is closer to 50% compared to if heterogeneity

is not taken into account (left plot). This reflects the larger uncertainty about the effect size θ in the case of heterogeneity. To achieve a probability of replication success of 80% the fewest samples are required with meta-analysis, followed by the sceptical p-value, the two-trials rule, the replication Bayes factor, the sceptical Bayes factor, and lastly the equivalence test. If the chosen sample size should guarantee a sufficiently conclusive replication study with all these methods, the replication sample size has be slightly larger than the original one in the situation of no heterogeneity ($\tau^2 = 0$), while it has to be more than ten-fold increased if there is heterogeneity ($\tau^2 = 0.05$). However, this is mostly due to the equivalence test which requires by far the most samples. If the equivalence test sample size is ignored, $c = 2.5$ leads to at least 80% probability of replication success with the remaining methods.

The right plot in 4 shows that the the type I error rate of the two-trials rule stays constant at $\alpha = 0.025$, as expected by definition of the method. In contrast, the type I error rates of the other methods vary but most of them stay below $\alpha = 0.025$ for all relative sample sizes with the exception of meta-analysis and the sceptical p-value. Meta-analysis has an extremely highly type I error rate as the pooling with the highly significant original data results in replication success unless the replication sample size is drastically increased. The type I error rate of the sceptical p-value is only slightly larger than $\alpha = 0.025$ for small c and it becomes smaller than $\alpha = 0.025$ at approximately $c = 3$.

We now conducted the same analyses for the other studies from the replication project of Protzko et al. (2020). Figure 5 shows the required relative sample size c and the associated type I error rates if a sample size can be computed for a probability of replication success of $1 - \beta = 80\%$ otherwise the space is left blank. We see that for all methods but the equivalence test the required relative sample size c decreases with decreasing original p-value, and original studies with very small p-value typically require much fewer samples in the replication study. For the equivalence test the required c depends instead on the size of the original standard error σ_o, and smaller standard errors require smaller sample sizes in the replication. For instance, the experiment "Orientation" with the original standard error $\sigma_o = 0.045$ requires much less samples than the experiment "Self-Control" with original standard error $\sigma_o = 0.052$. In general taking into account heterogeneity increases the required sample size for all methods. At the same time, a sample size reduces the type I error rate if the original effect estimate $\hat{\theta}_o$ is sufficiently different from zero. If the original effect estimate $\hat{\theta}_o$ is close to zero, the type I error rate of the equivalence test is drastically increased due to this particular definition of replication success being not dependent on the distance to the null hypothesis.

3.6 Sample size determination for multisite replication projects

So far we have considered the situation where one replication is conducted per original study. It is, however, also possible to conduct several replications per original study, so-called *multisite* replication studies. The replication project by Protzko et al. (2020) is an example for this type of design. When the entire ensemble of replications is analyzed jointly, some adaptations of the previously described SSD approach are required.

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

$$\hat{\boldsymbol{\theta}}_r \,|\, \boldsymbol{\theta}_r \sim \mathrm{N}_m \left\{ \boldsymbol{\theta}_r, \mathrm{diag}\left(\boldsymbol{\sigma}_r^2\right) \right\} \tag{24a}$$

$$\boldsymbol{\theta}_r \,|\, \theta \sim \mathrm{N}_m \left\{ \theta\, \mathbf{1}_m, \tau^2 \mathrm{diag}(\mathbf{1}_m) \right\}, \tag{24b}$$

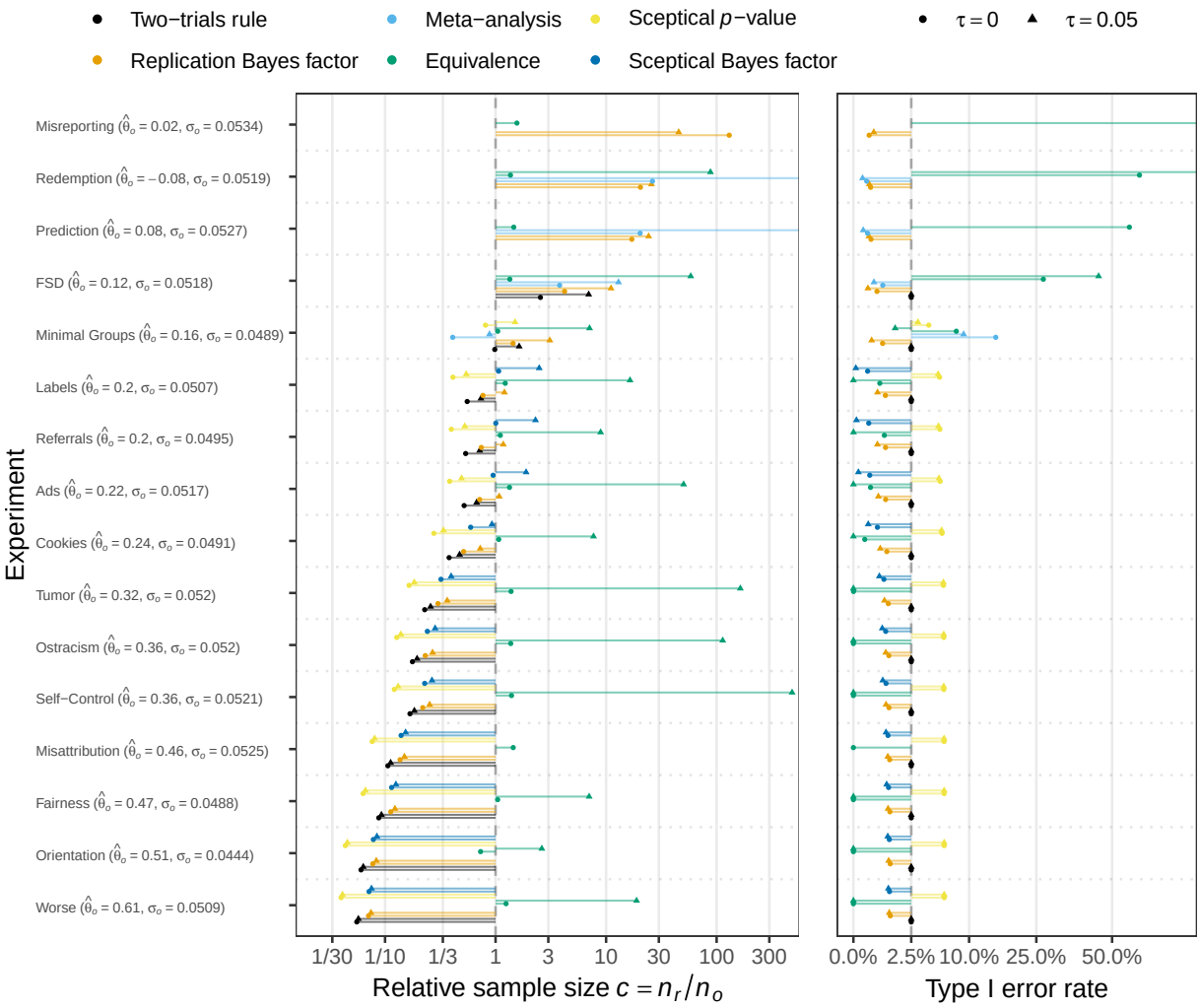


Figure 5: The left plot shows the required relative sample size $c = n_r/n_o$ to achieve replication success with probability of replication success of $1 - \beta = 80\%$ (if possible). Replication success is defined through the two-trials rule at level $\alpha = 0.025$, replication Bayes factor at level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

where θ_r is a vector of m study specific effect sizes and $\mathbf{1}_m$ is a vector of m ones. By marginalizing over the study specific effect size $\boldsymbol{\theta}_k$, the model can alternatively be expressed by

$$\hat{\boldsymbol{\theta}}_r \,|\, \theta \sim \mathrm{N}_m \left\{ \theta \, \mathbf{1}_m, \mathrm{diag} \left(\boldsymbol{\sigma}_r^2 + \tau^2 \mathbf{1}_m \right) \right\}, \tag{25}$$

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

$$\hat{\boldsymbol{\theta}}_r \,|\, \hat{\theta}_o, \sigma_o^2, \boldsymbol{\sigma}_r^2 \sim \mathrm{N}_m \left\{ \mu_{\hat{\theta}_r} \mathbf{1}_m, \mathrm{diag} \left(\boldsymbol{\sigma}_r^2 + \tau^2 \mathbf{1}_m \right) + \left(\frac{\tau^2 + \sigma_o^2}{1 + 1/g} \right) \mathbf{1}_m \mathbf{1}_m^\top \right\} \tag{26}$$

with $\mu_{\hat{\theta}_r}$ the mean of the predictive distribution of a single replication effect estimate from (7).

Often the assessment of replication success can be formulated in terms of a weighted average of the replication effect estimates $\hat{\theta}_{r*} = (\sum_{i=1}^m w_i \hat{\theta}_{ri})/(\sum_{i=1}^m w_i)$. For instance, several multisite replication

projects (e. g., Klein et al., 2018) have investigated replicability in terms of the statistical significance of a meta-analytic effect estimate of the effect size θ. Based on (26), the predictive distribution of the weighted average is given by

$$\hat{\theta}_{r*} \,|\, \hat{\theta}_o, \sigma_o^2, \sigma_r^2 \sim \mathrm{N}\left\{\mu_{\hat{\theta}_r}, \sigma^2_{\hat{\theta}_{r*}} = \left(\sum_{i=1}^{m} w_i^2 \sigma^2_{\hat{\theta}_{ri}} + \sum_{i=1}^{m}\sum_{\substack{j=1 \\ j\neq i}}^{m} w_i w_j \frac{\tau^2 + \sigma_o^2}{1 + 1/g}\right) \Big/ \left(\sum_{i=1}^{m} w_i\right)^2\right\} \tag{27}$$

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

$$\sigma^2_{\hat{\theta}_{r*}} = \frac{\sigma_r^2 + \tau^2}{m} + \frac{\tau^2 + \sigma_o^2}{1 + 1/g}. \tag{28}$$

The probability of replication success can now be obtained by integrating (26), respectively (27), over the corresponding success region S.

3.6.1 Optimal allocation within and between sites

A key challenge in SSD for multisite replication studies is the optimal allocation of samples within and between sites, that is, how many sites m and how many samples n_{ri} per site i should be used. A similar problem exists in SSD for cluster randomized trials and we can adapt the common solution based on cost functions (Cochran, 1977, Chapter 9.6). That is, the optimal configuration is determined such that the probability of replication success is maximized subject to a constrained cost function which accounts for the (possibly different) costs of additional samples and additional sites.

For example, assume a balanced design ($n_{ri} = n_r$ for $i = 1, \ldots, m$) and that the standard errors of the replication effect estimates are inversely proportional to the square-root of the sample size $\sigma_{ri} = \lambda/\sqrt{n_r}$ for some unit variance λ^2. Further assume that maximizing the probability of replication success is corresponds to minimizing the variance of the weighted average $\sigma^2_{\hat{\theta}_{r*}}$ from (28). Let K_s denote the cost of an additional site, and K_c the cost of an additional case. The total cost of the project is then $K = m(K_c\, n_r + K_s)$, and constrained minimization of (28) leads to the optimal sample size per site

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

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

3.7 Example: Cross-laboratory replication project (continued)

Figure 6 illustrates multisite SSD for the experiment "Labels" from the Protzko et al. (2020) project for the two-trials rule and the replication Bayes factor (see the supplement for details on the multisite extension of these two methods). The top plots show the probability of replication success as a function

of the total sample size for different number of sites m. We see that for the same total sample size a larger number of sites increases the probability of replication success. For instance, a total sample size of roughly 3000 is required to achieve 80% probability with one site for the two-trials rule, whereas only approximately half as many samples are required for two sites.

However, focusing only on the total sample size ignores the fact that the costs of an additional site are usually larger than the cost of an additional observation. The bottom plot shows the total cost K of a design (relative to the cost of one case K_c) whose sample size is determined so that the probability of replication success is $1 - \beta = 80\%$. We see that if the cost of an additional site K_s is not much larger than the cost of an additional sample K_c, e. g., $K_S/K_c = 30$ the optimal number of sites is $m = 5$ for the two-trials rule and $m = 8$ for the replication Bayes factor. If an additional site is more costly, e. g., $K_S/K_c = 300$, the optimal number of sites is lower $m = 2$ for the two trials rule and $m = 3$ for the replication Bayes factor.

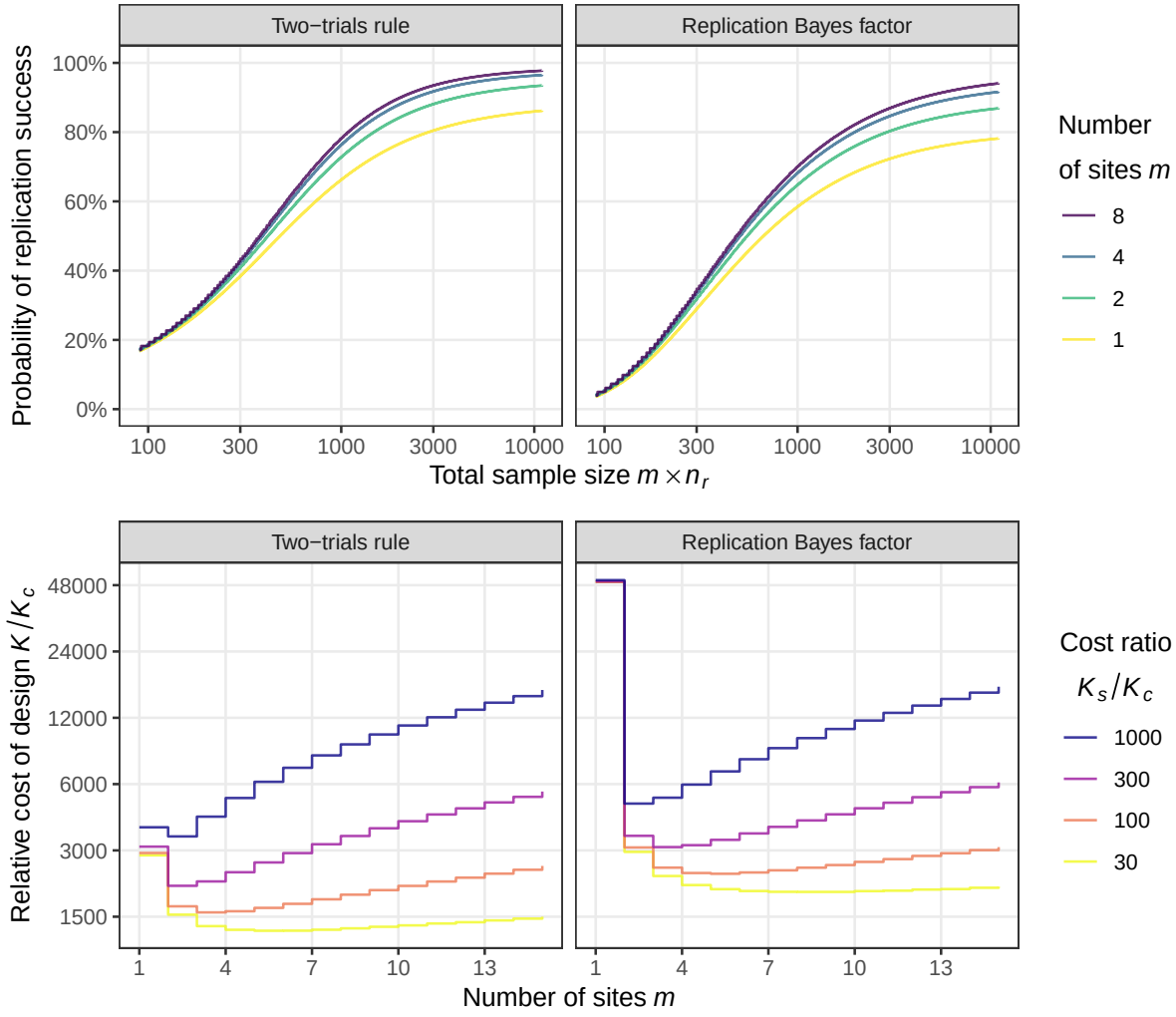


Figure 6: Top plots show the probability of replication success based on the replication Bayes factor at level $\gamma = 1/10$ (left) and the two-trials rule at level $\alpha = 0.025$ (right) as a function of the total sample size and for different number of sites m for data from the experiment "Labels".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

4 Discussion

We showed how Bayesian approaches can be used for determining the sample size of replication studies conducting SSD of replication studies. The Bayesian framework allows to make use of all the available information, and to take into account the associated uncertainty. A key strength is that the approach can also be applied if the analysis of the replication will not be Bayesian, which is the most common situation.

There are some limitations and possible extensions: We have looked at the normal normal hierarchical model with fixed variances in order to obtain closed form expressions for the probability of replication success. Specifying priors on the between-study heterogeneity variances could better reflect the available uncertainty but would come at the price of lower interpretability and higher computational complexity. We did also not considered designs where the replication data are analyzed in a sequential manner. Ideas from the Bayesian sequential design (Schönbrodt and Wagenmakers, 2017) or from the adaptive trials literature (Bretz et al., 2009) could be adapted to the replication setting as in Micheloud and Held (2022). A sequential analysis of the replication data could possibly increase the efficiency of the replication. However, it would also make SSD and practical aspects more challenging. Moreover, we have assumed that the original study has already been finished. One could also consider a scenario where both the original and replication study are planned simultaneously and adopt a "project" perspective as in Held et al. (2022b). However, in this case no information from the original study is available and the design prior needs to be specified entirely based on external knowledge. Finally, researchers have only limited resources and it may happen that they cannot afford a large enough sample size to obtain their desired probability of replication success. In this situation a reverse-Bayes approach (Held et al., 2022a) could be applied in order to determine the prior for the effect size which is required to meet all design requirements based on a fixed sample size. Researchers can then judge whether or not such prior beliefs are scientifically sensible, and decide whether they should conduct the replication study with their limited resources.

Software and data

The data from Protzko et al. (2020) were were downloaded from `https://osf.io/42ef9/`. All analyses were conducted in the R programming language version 4.2.1 (R Core Team, 2022). The code to reproduce this manuscript is available at `https://github.com/SamCH93/BAtDRS`. A snapshot of the Git repository at the time of writing this article is archived at `https://doi.org/10.5281/zenodo.XXXXXX`. Methods for Bayesian SSD of replication studies are implemented in the R package `BayesRepDesign` which is available at `https://github.com/SamCH93/BayesRepDesign`. Appendix A illustrates the basic usage of the package.

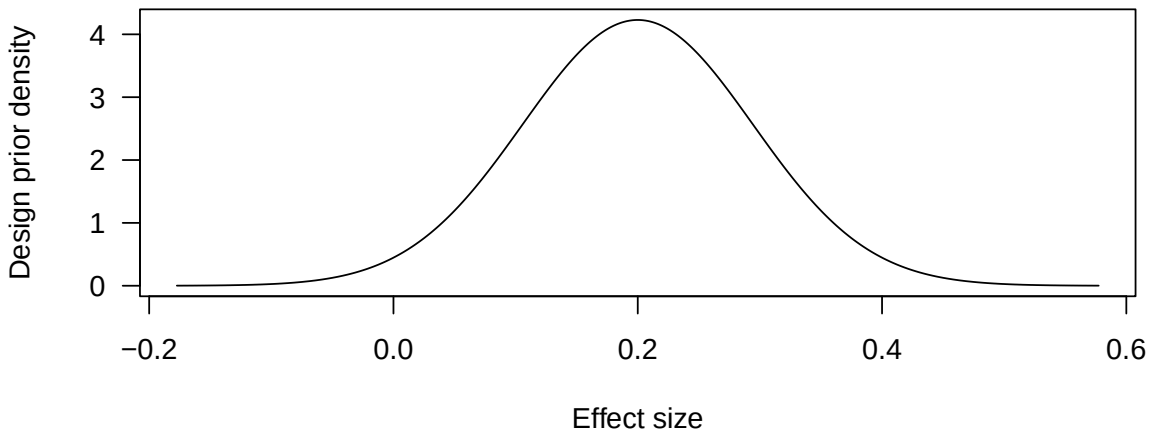
Acknowledgments

This work was supported by the Swiss National Science Foundation(#189295). The funder had no role in study design, data collection, data analysis, data interpretation, decision to publish, or preparation of the manuscript. We thank Protzko et al. (2020) for publicly sharing their data. We thank Charlotte Micheloud for helpful comments on drafts of the manuscript.

Appendix A The BayesRepDesign R package

```r
library("BayesRepDesign")

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




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

##
## Inputs:
## to = 0.2 : original effect estimate
## so = 0.05 : standard error of original effect estimate
## tau = 0.08 : assumed heterogeneity standard deviation of effect sizes
## N(mean = 0, sd = Inf) : initial normal prior for overall effect size
##
## Output:
## N(mean = 0.2, sd = 0.094) : normal design prior for overall effect size
##
##
## power = 0.8 (specified)
## power = 0.8 (recomputed with sr)
## sr = 0.045
## c = so^2/sr^2 ~= nr/no = 1.2
```

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.

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

Bretz, F., Koenig, F., Brannath, W., Glimm, E., and Posch, M. (2009). Adaptive designs for confirmatory clinical trials. *Statistics in Medicine*, 28(8):1181–1217. doi:10.1002/sim.3538.

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.

Cochran, W. (1977). *Sampling Techniques, 3rd ed.* Wiley, New York, NY.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1):155–159.

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

Deeks, J. J., Higgins, J. P., and Altman, D. G. (2019). Analysing data and undertaking meta-analyses. In *Cochrane Handbook for Systematic Reviews of Interventions*, chapter 10, pages 241–284. John Wiley & Sons, Ltd.

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

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

Gelman, A. (2009). Bayes, Jeffreys, prior distributions and the philosophy of statistics. *Statistical Science*, 24(2). doi:10.1214/09-sts284d.

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

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

Grieve, A. P. (2022). *Hybrid frequentist/Bayesian power and Bayesian power in planning clinical trials*. Chapman & Hall/CRC Biostatistics Series. Taylor & Francis, London, England.

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

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2):107–128. doi:10.3102/10769986006002107.

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

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

Held, L., Micheloud, C., and Pawel, S. (2022b). The assessment of replication success based on relative effect size. *The Annals of Applied Statistics*, 16(2):706–720. doi:10.1214/21-aoas1502.

Held, L. and Pawel, S. (2020). Comment on "the role of p-values in judging the strength of evidence and realistic replication expectations". *Statistics in Biopharmaceutical Research*, 13(1):46–48. doi:10.1080/19466315.2020.1828161.

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

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

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

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

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

O'Hagan, A. and Stevens, J. (2001). Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Medical Decision Making*, 21(3):219–230. doi:10.1177/02729890122062514.

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

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

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

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

Raudenbush, S. W. and Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, 5(2):199–213. doi:10.1037/1082-989x.5.2.199.

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.

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

Senn, S. (2002). Letter to the editor: A comment on replication, p-values and evidence by S. N. Goodman, Statistics in Medicine 1992; 11:875–879. *Statistics in Medicine*, 21(16):2437–2444. doi:10.1002/sim.1072.

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

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

Spiegelhalter, D. J. (1986). Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine*, 5:421–433.

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

Spiegelhalter, D. J. and Freedman, L. S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine*, 5(1):1–13. doi:10.1002/sim.4780050103.

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

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

van Zwet, E., Schwab, S., and Senn, S. (2021). The statistical properties of RCTs and a proposal for shrinkage. *Statistics in Medicine*, 40(27):6107–6117. doi:10.1002/sim.9173.

van Zwet, E. W. and Goodman, S. N. (2022). How large should the next study be? predictive power and sample size requirements for replication studies. *Statistics in Medicine*, 41(16):3090–3101. doi:10.1002/sim.9406.

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.

Weiss, R. (1997). Bayesian sample size calculations for hypothesis testing. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(2):185–191. doi:10.1111/1467-9884.00075.