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

September 14, 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This is a preprint which has not yet been peer reviewed.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 a replication study is determining its sample size. Here we show how Bayesian approaches can be used to determine the optimal replication sample size. The Bayesian framework allows both the original study data and external knowledge to be incorporated into a design prior for the underlying parameters. This is particularly useful because external knowledge, such as expected between-study parameter heterogeneity due to differences in the study population of original and replication study, are common in the replication setting. 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; 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.

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. Our proposed framework 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). 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 (O'Hagan and Stevens, 2001). 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 not 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). Finally, the paper ends with concluding remarks and practical recommendations (Section 4).

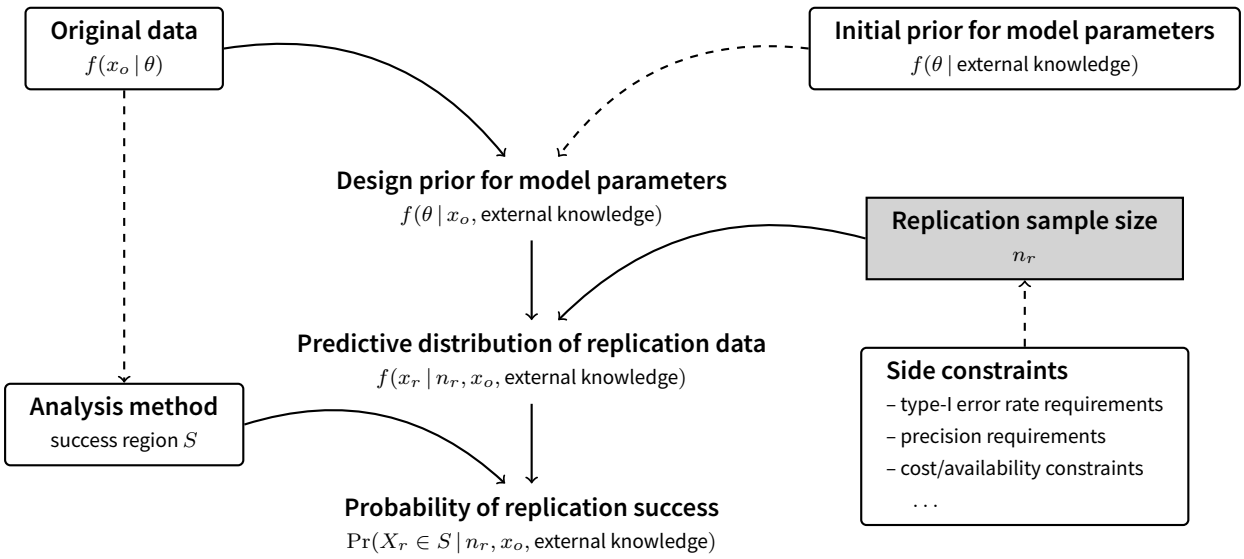

Figure 1: Schematic illustration of the process of Bayesian sample size determination for replication studies.

2 General framework

Figure 1 schematically illustrates the process of Bayesian SSD for replication studies, which we will explain in more detail in the following. 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. In many cases, there will be 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 $1 - \beta$ (in analogy to the desired power in classical SSD). The required sample size n_r^* is then given by

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

Often, replication studies are analyzed using several methods which quantify different aspects of replicability, and which consequently 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 also be side constraints which the replication sample size needs to satisfy. For instance, funders and regulators may require from a method to be *calibrated* (Grieve, 2016), that is, to have appropriate type I error rate control. A possible side constraint may thus be that the sample size n_r has to be chosen such that the probability of replication success under a null hypothesis is not larger than

some desired level. Similarly, it may be reasonable to have *precision* constraints, e. g., to require that a confidence interval for the replication effect estimate has to be at least as tight as the confidence interval estimated in the original study. Finally, in most cases there is an upper limit on the possible sample size due to limited resources and/or availability of samples.

3 Sample size determination in the normal normal hierarchical model

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 as 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 $x_r = \{\hat{\theta}_r, \sigma_r\}$. In the conventional meta-analytic framework it is then further assumed 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 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 inversely proportional to the square root of the sample size n_k, i. e., $\sigma_k = \lambda/\sqrt{n_k}$ with λ^2 some unit variance. 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 size needs to be changed compared to the original study. 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 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 this model provides great flexibility for incorporating the original data and external knowledge into SSD of replication studies.

3.1 Specification of the initial prior

The between-study heterogeneity variance τ^2, the effect size mean μ_θ, and the effect size variance σ_θ^2 of the initial prior provide a means for incorporating external knowledge in the design prior.

A replication study typically tries to follow the original study as closely as possible, yet in practice there will always be deviations from the original protocol. Other factors such as different laboratory equipment or different populations of participants may also increase the dissimilarity between original and replication study. As a result, the effect sizes underlying both studies may differ and one may want to the expected degree of between-study heterogeneity can be incorporated in the design 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 increasingly unrelated. It is also possible to specify an initial prior for τ^2 (see the literature from meta-analysis on this issue e. g., Röver et al., 2021). Here, we consider it as fixed and its value could be informed by external knowledge, for example, by estimates from the literature (e. g., Erp et al., 2017). This approach leads to closed-form expressions for the probability of replication success in many cases, which in turn helps to better understand the effects of the design prior on the resulting replication sample size.

The initial prior of the overall effect size (4c), is centered around μ_θ with variance σ_θ^2. An uninformative initial prior can be obtained by choosing an infinitely large variance ($\sigma_\theta^2 \to \infty$), while for finite variances the design prior will be pulled towards the initial prior mean. If the replicators trust the result from the original study and have no further external knowledge available, they can thus specify an uninformative prior and "let the data speak for themselves". For example, this might be appropriate if the original study was a rigorously conducted confirmatory study with a preregistered study protocol, blinding, large sample sizes, etc. Optimists may even specify μ_θ and σ_θ^2 of the initial prior based on a meta-analysis of related studies or based on expert elicitation so that the resulting design prior contains more information than what provided by the original data alone . In contrast, if the replicators are more sceptical about the original study, an alternative option is to choose $\mu_\theta = 0$ to obtain a shrinkage prior which shrinks the design prior towards less impressive effect sizes than the observed one. The amount of shrinkage is then determined via the variance σ_θ^2. Specifying a diffuse prior ($\sigma_\theta^2 \to \infty$) will lead to no shrinkage, while specifying a highly concentrated prior ($\sigma_\theta^2 \downarrow 0$) will completely shrink the design prior to a point mass. In practice, a pragmatic option with good predictive properties is to use the empirical Bayes estimate based on the original data $\hat{\sigma}_\theta^2 = \max\{(\hat{\theta}_o - \mu_\theta)^2 - \tau^2 - \sigma_o^2, 0\}$. 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).

3.2 Design prior and predictive distribution

Once an initial prior has been specified via the parameters τ^2, μ_θ, and σ_θ^2, it can be combined with the data from the original study to obtain a design prior for the effect size θ. Straightforward application of Bayes theorem (1) leads then to the design prior

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

with *relative prior variance* $g = \sigma_\theta^2/(\sigma_o^2 + \tau^2)$. With (**??**), we obtain the predictive distribution of the replication effect estimate

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

Sometimes it is more convenient to work with the associated predictive distribution of the replication z-value $z_r = \hat{\theta}_r/\sigma_r$, which is given by

$$z_r \,|\, z_o, c \sim \mathrm{N}\left(\mu_{z_r} = \frac{z_o\sqrt{c}}{1+1/g} + \frac{z_\theta\sqrt{cg(1+h)}}{1+g}, \sigma_{z_r}^2 = 1 + ch + \frac{c\,(1+h)}{1+1/g}\right) \tag{8}$$

where $z_o = \hat{\theta}_o/\sigma_o$ is the original z-value, $z_\theta = \mu_\theta/\sigma_\theta$ is the prior z-value, and $h = \tau^2/\sigma_o^2$ is the relative heterogeneity. In the following, we will give several examples for how to specify normal design priors.

3.3 Probability of replication success

3.3.1 The two trials rule

$$S_{\text{2TR}} = [z_{1-\alpha}\,\sigma_r, \infty) \tag{9}$$

3.3.2 Fixed effects meta-analysis

$$S_{\text{MA}} = \left[z_{1-\alpha}\,\sigma_r\left\{\sqrt{\frac{1}{c}+1} - z_o\sqrt{c}\right\}, \infty\right) \tag{10}$$

3.3.3 The replication Bayes factor

$$S_{\text{BF}_\text{R}} = \left(-\infty, -\sqrt{A} - \frac{z_o}{\sqrt{c}}\right] \bigcup \left[\sqrt{A} - \frac{z_o}{\sqrt{c}}, \infty\right) \tag{11}$$

3.3.4 Reverse-Bayes methods

4 Discussion

We have presented a Bayesian approach for SSD of replication studies. The Bayesian framework allows to make use of all the available information, and to take into account the associated uncertainty. We have also discussed how different design requirements can be combined to satisfy different stakeholders, while also enabling conclusive inferences based on several analyses approaches of the replication data. As we showed, the approach helps to design informative and cost-effective replications. We have illustrated the approach for three Bayesian measures of replication success, but in principle our framework can be used for any analysis method, Bayes or non-Bayes, parameter estimation or hypothesis testing.

There are some limitations and possible extensions: We have treated all variances as fixed in order to obtain closed form expressions for the probability of replication success. Also 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 have 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 B 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 Multisite Bayes factors

To determine the multisite version of the replication and the sceptical Bayes factor we need to know the marginal density of the replication effect estimates $\hat{\theta}_r \,|\, \theta \sim \mathrm{N}_n(\theta J_n, \text{diag}\left\{\sigma_r^2 + \tau_r^2 J_n\right\})$ under a normal prior $H_k\colon \theta \sim \mathrm{N}(m, v)$. 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 and its variance. The marginal density is then

given by

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

The marginal density under the sceptical prior H_S is then obtained by setting $m = 0$ and $v = s\sigma_o^2$, whereas the marginal density under the advocacy prior H_A is obtained by setting $m = \hat{\theta}_o$ and $v = \sigma_o^2$. The marginal density under the null hypothesis H_0 is itself a special case of the density under H_S with $s = 0$. Taken together, this leads to the Bayes factor

$$
\mathrm{BF}_{\mathrm{SA}}(\hat{\theta}_r; s) = \frac{f(\hat{\theta}_r \mid H_\mathrm{S})}{f(\hat{\theta}_r \mid H_\mathrm{A})} = \sqrt{\frac{1+\sigma_o^2/\sigma_{r*}^2}{1+s\sigma_o^2/\sigma_{r*}^2}} \exp\left[-\frac{1}{2}\left\{\frac{\hat{\theta}_{r*}^2}{\sigma_{r*}^2+s\sigma_o^2} - \frac{(\hat{\theta}_{r*}-\hat{\theta}_o)^2}{\sigma_{r*}^2+\sigma_o^2}\right\}\right]
$$

which can be further simplified to (**??**).

Appendix B The `BayesRepDesign` R package

References

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.

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.

Erp, S. V., Verhagen, J., Grasman, R. P. P. P., and Wagenmakers, E.-J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in *psychological bulletin* from 1990-2013. *Journal of Open Psychology Data*, 5(1):4. doi:10.5334/jopd.33.

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.

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.

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

Hedges, L. V. and Schauer, J. M. (2019). 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.

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

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

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/`.

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.

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.