Power priors for design and analysis of replication studies

Samuel Pawel[*], **Frederik Aust**[†], **Eric-Jan Wagenmakers**[†], **Leonhard Held**[*]
[*] Department of Biostatistics, University of Zurich
[†] Department of Psychological Methods, University of Amsterdam
E-mail: samuel.pawel@uzh.ch

March 2, 2022

Abstract

Power priors are a class of prior distributions that allow to incorporate historical data in Bayesian analyses. One domain where historical data are available is the analysis of replication studies. Despite the growing interest in replication studies, no one has yet investigated a power prior modeling approach for the analysis of replication studies. Here, we explore this idea for both parameter estimation and hypothesis testing. We show how power priors help both in estimation and testing of effect sizes via a dynamic borrowing of information from the original study. At the same time, they allow to make inferences about study compatibility through the power parameter, which links the original to the replication study. It is well known that for normal data there is an exact mapping of posterior inferences between power prior and hierarchical models when the power parameter and heterogeneity variance are fixed. We establish that a similar mapping exists also in the case when these parameters are assigned a prior distribution. Our results help to better understand prior distributions on heterogenetiy variances, and also how power prior models can alternatively be estimated through a hierarchical modeling framework.

Key words: Power prior, hierarchical models, replication studies, historical data, Bayesian hypothesis testing, Bayesian parameter estimation

1 Introduction

Power priors are a useful class of informative prior distributions that allow to incorporate historical data in Bayesian analyses (Ibrahim et al., 2015). The most basic version of the power prior is obtained by updating an initial prior with the likelihood of the historical data raised to the power of α, where α is usually restricted to the range between zero and one. As such, the power parameter α specifies how much the historical data are discounted, and thus provides a quantitative and easy to interpret way of incorporating historical data.

One domain where historical data are available is the analysis of replication studies. The question is typically to what extent the replication study replicated the result of an original study. Several methods have been proposed to answer this question (Bayarri and Mayoral, 2002; Verhagen and Wagenmakers, 2014; Johnson et al., 2016; Etz and Vandekerckhove, 2016; van Aert and van Assen, 2017; Ly et al., 2018; Hedges and Schauer, 2019; Mathur and VanderWeele, 2020; Held, 2020; Pawel and Held, 2020, 2022; Held et al., 2022, among others). Constructing a power prior from the original data and using it in the analysis of the replication study seems a natural thing to do. However, no one has yet investigated such an approach.

In this paper we show how power priors can be constructed from data of an original study under a meta-analytic framework (Section 2). We then show, how the power prior can then be used for for both parameter estimation (Section 3) and hypothesis testing (Section 4). Our

results indicate that power priors help both in estimation and testing of effect sizes via a dynamic borrowing of information from the original study. At the same time, the approach enables inferences about study compatibility through the power parameter, which links the original to the replication study. It is well known that for normal data there is an exact mapping of posterior inferences between power prior and hierarchical models when the power parameter and heterogeneity variance are fixed (Chen and Ibrahim, 2006). We establish that a similar mapping exists also in the case when these parameters are random and assigned a prior distribution ...

2 The power prior based on an original study

Let θ denote an unknown effect size and $\hat{\theta}_i$ an estimate thereof obtained from study $i \in \{o, r\}$ where the subscript indicates "original" or "replication", respectively. Assume that the likelihood of the effect estimates can be approximated by a normal distribution

$$\hat{\theta}_i \,|\, \theta \sim \mathrm{N}(\theta, \sigma_i^2)$$

with σ_i the (assumed to be known) standard error of the effect estimate $\hat{\theta}_i$. This is the same framework as typically used in meta-analysis, and it is applicable to many types of data and effect sizes (Spiegelhalter et al., 2004, chapter 2.4). There are, of course, situations where the approximation is inadequate and modified distributional assumptions are required (*e. g.* for data from studies with small sample sizes and/or extreme effect sizes).

The goal is now to construct a power prior for θ based on the data from the original study. Under an (improper) flat initial prior $f(\theta) \propto 1$, normalization of the likelihood of the original data raised to a (fixed) power parameter α leads to the normalized power prior

$$\theta \,|\, \hat{\theta}_o, \sigma_o, \alpha \sim \mathrm{N}\left(\hat{\theta}_o, \frac{\sigma_o^2}{\alpha}\right) \tag{1}$$

as proposed by Neuenschwander et al. (2009). There are different ways to specify α. The simplest approach fixes α to an a priori reasonable value, possibly informed from external knowledge about the similarity of the two studies. Another option is to use the empirical Bayes estimate (Gravestock and Held, 2017), *i. e.* , the value of α that maximizes the likelihood of the replication data marginalized over the power prior

$$\hat{\alpha}_{\mathrm{EB}} = \max\left[0, \min\left\{1, \frac{\sigma_o^2}{(\hat{\theta}_o - \hat{\theta}_r)^2 - \sigma_r^2}\right\}\right].$$

Finally, it is also possible to specify a prior distribution for α, the most common choice being a marginal beta distribution

$$\alpha \,|\, x, y \sim \mathrm{Be}(x, y)$$

for a normalized power prior conditional on α as in (1). The uniform distribution ($x = 1$, $y = 1$) is often recommended as the default choice in the literature (Ibrahim et al., 2015). However, we note that if α is not restricted to the unit interval, another default choice could be a $f(\alpha) \propto \alpha^{-1}$ as this is the reference prior for a precision parameter in the normal model.

3 Parameter estimation

Assuming a beta prior for α and conditioning on the replication data leads to the posterior distribution

$$\begin{aligned} f(\alpha, \theta \mid \hat{\theta}_r, \hat{\theta}_o, \sigma_o, \sigma_r, x, y) =& \frac{\mathrm{N}(\hat{\theta}_r; \theta, \sigma_r^2) \times \mathrm{N}(\theta; \hat{\theta}_o, \sigma_o^2/\alpha) \times \mathrm{Be}(\alpha; x, y)}{f(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_r, \sigma_o, x, y)} \\ \propto& \exp\left[-\frac{1}{2}\left\{\left(\frac{1}{\sigma_r^2} + \frac{\alpha}{\sigma_o^2}\right)\left(\theta - \frac{\hat{\theta}_r/\sigma_r^2 + (\hat{\theta}_r\,\alpha)/\sigma_o^2}{1/\sigma_r^2 + \alpha/\sigma_o^2}\right)^2 + \frac{(\hat{\theta}_o - \hat{\theta}_r)^2}{\sigma_o^2/\alpha + \sigma_r^2}\right\}\right] \\ & \times \alpha^{x-1/2}\,(1-\alpha)^{y-1} \end{aligned} \tag{2}$$

with $\mathrm{N}(z; m, v)$ the density function of a normal distribution with mean m and variance v evaluated at z, and $\mathrm{Be}(u; q, p)$ the density function of a beta distribution with parameters q and p evaluated at u. The normalizing constant

$$\begin{aligned} f(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_r, \sigma_o, x, y) &= \int_0^1 \int_{-\infty}^{\infty} \mathrm{N}(\hat{\theta}_r; \theta, \sigma_r^2) \times \mathrm{N}(\theta; \hat{\theta}_o, \sigma_o^2/\alpha) \times \mathrm{Be}(\alpha; x, y)\,\mathrm{d}\theta\,\mathrm{d}\alpha \\ &= \int_0^1 \mathrm{N}(\hat{\theta}_r; \hat{\theta}_o, \sigma_r^2 + \sigma_o^2/\alpha) \times \mathrm{Be}(\alpha; x, y)\,\mathrm{d}\alpha \end{aligned}$$

is not available in closed form but requires numerical integration with respect to the prior distribution of α. Finally, if inference concerns only one parameter, a marginal posterior distributions for either α or θ can be obtained by integrating out the respective nuisance parameter from (2). In the case of the power parameter α, this leads to

$$f(\alpha, \mid \hat{\theta}_r, \hat{\theta}_o, \sigma_o, \sigma_r, x, y) = \frac{\mathrm{N}(\hat{\theta}_r; \hat{\theta}_o, \sigma_r^2 + \sigma_o^2/\alpha) \times \mathrm{Be}(\alpha; x, y)}{f(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_r, \sigma_o, x, y)} \tag{3}$$

while for the effect size θ, we have

$$f(\theta, \mid \hat{\theta}_r, \hat{\theta}_o, \sigma_o, \sigma_r, x, y) = \frac{\mathrm{N}(\hat{\theta}_r; \theta, \sigma_r^2) \int_0^1 \mathrm{N}(\theta; \hat{\theta}_o, \sigma_o^2/\alpha) \times \mathrm{Be}(\alpha; x, y)\,\mathrm{d}\alpha}{f(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_r, \sigma_o, x, y)}.$$

3.1 Example "Labels"

We will now apply the methodology to data from a large-scale replication project by Protzko et al. (2020). The project featured an experiment called "Labels" for which the original study found the following result: "When a researcher uses a label to describe people who hold a certain opinion, he or she is interpreted as disagreeing with those attributes when a negative label is used and agreeing with those attributes when a positive label is used." This finding came with a standardized mean difference effect estimate $\hat{\theta}_o = 0.2$ and standard error $\sigma_o = 0.05$ obtained from 1577 participants. Subsequently, four replication studies were conducted, three of them by a different lab than the original one, and all employing large sample sizes.

Figure 1 shows joint and marginal posterior distributions for effect size θ and power parameter α based on the results of the three external replication studies. We see that one replication found a virtually identical effect estimate as the original study ($\hat{\theta}_r = 0.2$), while the other two replications found either a smaller or larger effect estimate ($\hat{\theta}_r = 0.09$ and $\hat{\theta}_r = 0.44$). This is reflected in the marginal posterior distributions of the power parameter α. That is, the marginal distribution of replication with the very conflicting effect estimate is sharply peaked and has most mass at small values of α. In contrast, the marginal distribution based on the replication with strongly agreeing effect estimate is monotonically increasing, giving the highest support to the value $\alpha = 1$. Yet, the posterior density at $\alpha = 1$ is not much higher than one (the density of prior), suggesting that it is difficult to obtain conclusive posterior inferences about α, even from two highly precise studies.

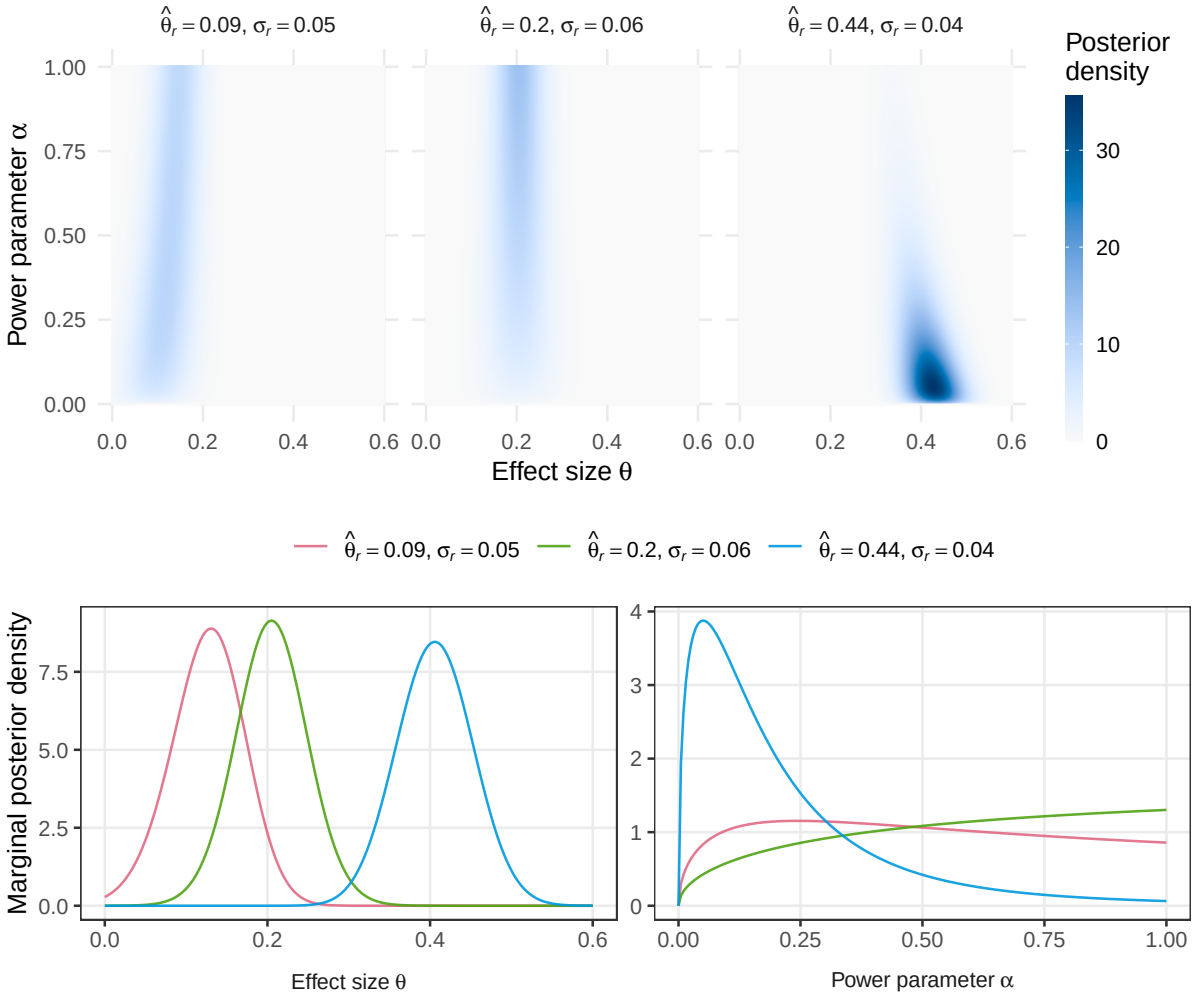

Figure 1: Bayesian analysis of three replication studies from the replication project by Protzko et al. (2020). Shown are joint (top) and marginal (bottom) posterior distributions of effect size θ and power parameter α. A power prior for the effect size θ is constructed from the original effect estimate $\hat{\theta}_o = 0.2$ (with standard error $\sigma_o = 0.05$) and an initial flat prior $f(\theta) \propto 1$. A $\alpha \sim \text{Be}(1, 1)$ marginal prior distribution is used for the power parameter.

The marginal distribution of the effect size θ shows that the degree of compatibility between the two studies, influences how much information is borrowed from the original study. For instance, the marginal posterior density based on the most compatible replication ($\hat{\theta}_r = 0.2$), is the most concentrated among the three replications, despite that this estimate was the least precise. In contrast, the marginal posterior of the most conflicting estimate ($\hat{\theta}_r = 0.44$) borrows fewer information and is the least peaked posterior, despite being the most precise estimate among the three.

3.2 Connection to parameter estimation hierarchical models

Hierarchical modeling is another approach that allows for incorporation of historical data, and hierarchical models have also been used in the replication setting previously (Bayarri and Mayoral, 2002; Pawel and Held, 2020). Assume a hierarchical model

$$\begin{aligned} \hat{\theta}_i \,|\, \theta_i &\sim \text{N}(\theta_i, \sigma_i^2) \\ \theta_i \,|\, \theta_* &\sim \text{N}(\theta_*, \tau^2) \\ \theta_* &\sim \text{N}(0, \infty) \end{aligned}$$

for study $i \in \{o, r\}$. The joint posterior is then proportional to

$$f(\theta_r, \theta_o, \theta_* \,|\, \hat{\theta}_o, \hat{\theta}_r, \tau^2) \propto \prod_{i \in \{o,r\}} \mathrm{N}(\hat{\theta}_i; \theta_i, \sigma_i^2)\ \mathrm{N}(\theta_i; \theta_*, \tau^2). \tag{5}$$

By integrating out θ_o and θ_* from (5), one can show that the marginal posterior distribution of the replication effect size θ_r is

$$\theta_r \,|\, \hat{\theta}_o, \hat{\theta}_r, \tau^2 \sim \mathrm{N}\left(\frac{\hat{\theta}_r/\sigma_r^2 + \hat{\theta}_o/(2\tau^2 + \sigma_o^2)}{1/\sigma_r^2 + 1/(2\tau^2 + \sigma_o^2)}, \frac{1}{1/\sigma_r^2 + 1/(2\tau^2 + \sigma_o^2)} \right). \tag{6}$$

The question is now whether there is a correspondence between the hierarchical and the power prior approach. Under the power prior and for a fixed power parameter α, the posterior of the effect size θ from (2) simplifies to a normal

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

Theorem 2.2 in Chen and Ibrahim (2006) establishes that the two posterior distributions (6) and (7) match if and only if

$$\alpha = \frac{\sigma_o^2}{2\tau^2 + \sigma_o^2},$$

respectively

$$\tau^2 = \left(\frac{1}{\alpha} - 1 \right) \frac{\sigma_o^2}{2}.$$

For instance, a power prior model with $\alpha = 1$ corresponds to a hierarchical model with $\tau^2 = 0$, and a hierarchical model with $\tau \to \infty$ corresponds to a power prior model with $\alpha \downarrow 0$. Interestingly, there is a direct mapping from α to the popular relative heterogeneity measure $I^2 = \tau^2/(\tau^2 + \sigma_o^2)$ (Higgins and Thompson, 2002), that is

$$\alpha = \frac{1 - I^2}{1 + I^2},$$

see also Figure 2. We see that there is an almost linear relationship betwen the two. A useful heuristic to connect power priors to hierarchical models is thus $\alpha \approx 1 - I^2$.

It is unclear whether a mapping exists in cases where α and τ^2 are random. If it would exist, it must hold for any $\theta = \theta_r$ that

$$\int_0^\infty f(\theta_r \,|\, \hat{\theta}_o, \hat{\theta}_r, \tau^2)\, f(\tau^2 \,|\, \hat{\theta}_o, \hat{\theta}_r)\, \mathrm{d}\tau^2 = \int_0^1 f(\theta \,|\, \hat{\theta}_o, \hat{\theta}_r, \alpha)\, f(\alpha \,|\, \hat{\theta}_o, \hat{\theta}_r)\, \mathrm{d}\alpha. \tag{8}$$

By applying a change of variables to the left or right hand side of (8), the marginal posteriors conditional on τ^2 and α match. It is now left to investigate whether there are priors $f(\tau^2)$ and $f(\alpha)$ so that also the marginal posteriors of τ^2 and α match. By replacing the beta prior in (3) with an unspecified prior $f(\alpha)$, we can see that the marginal posterior distribution of α is proportional to

$$f(\alpha \,|\, \hat{\theta}_o, \hat{\theta}_r) \propto f(\alpha) \times \mathrm{N}(\hat{\theta}_r; \hat{\theta}_o, \sigma_r^2 + \sigma_o^2/\alpha).$$

After a change of variables $\tau_*^2 = (1/\alpha - 1)\sigma_o^2/2$ this becomes

$$f(\tau_*^2 \,|\, \hat{\theta}_o, \hat{\theta}_r) \propto f(\alpha = \sigma_o^2/(2\tau_*^2 + \sigma_o^2))\, \frac{2\sigma_o^2}{(2\tau_*^2 + \sigma_o^2)^2} \times (\sigma_o^2 + \sigma_r^2 + 2\tau_*^2)^{-1/2}\, \exp\left\{ -\frac{1}{2} \frac{(\hat{\theta}_r - \hat{\theta}_o)^2}{\sigma_o^2 + \sigma_r^2 + 2\tau_*^2} \right\}.$$

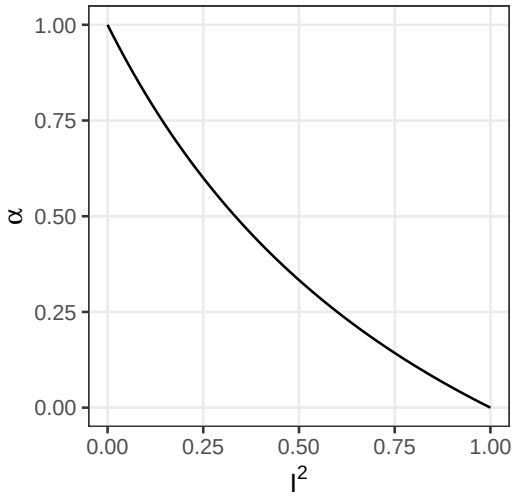

Figure 2: Relative heterogeneity $I^2 = \tau^2/(\tau^2 + \sigma_o^2)$ of hierarchical model and power parameter α from power prior model which lead to matching posteriors for the effect sizes θ and θ_r.

By integrating out θ_r, θ_o, and θ_* from the joint posterior under the hierarchical model, one can show that the marginal posterior of the heterogeneity variance τ^2 is proportional to

$$f(\tau^2 \mid \hat{\theta}_o, \hat{\theta}_r) \propto f(\tau^2) \times (\sigma_o^2 + \sigma_r^2 + 2\tau^2)^{-1/2} \exp\left\{ -\frac{1}{2} \frac{(\hat{\theta}_r - \hat{\theta}_o)^2}{\sigma_o^2 + \sigma_r^2 + 2\tau^2} \right\}.$$

This implies that the marginal posteriors of the effect sizes θ and θ_r match if it holds for every $\tau^2 = \tau_*^2$ that

$$f(\tau^2) = f(\alpha = \sigma_o^2/(2\tau_*^2 + \sigma_o^2)) \, \frac{2\sigma_o^2}{(2\tau_*^2 + \sigma_o^2)^2}. \tag{9}$$

For example, if we assign a $\text{Be}(x, y)$ prior to α, the posteriors will match for a

$$f(\tau^2) \propto \left(\frac{2\tau^2}{\sigma_o^2} + 1 \right)^{1-x} \left(\frac{\sigma_o^2}{2\tau^2} + 1 \right)^{1-y} \frac{2\sigma_o^2}{(2\tau^2 + \sigma_o^2)^2}$$

prior on τ^2.

Figure 3 illustrates several examples of matching priors using the variance of the original effect estimate from the "Labels" experiment. We see that the uniform prior on α corresponds to a $f(\tau^2) \propto \sigma_o^2/(2\tau^2 + \sigma_o^2)^2$ prior which is similar to the "uniform shrinkage" prior $f(\tau^2) \propto \sigma_o^2/(\tau^2 + \sigma_o^2)^2$ (Daniels, 1999). This prior has the highest density at $\tau^2 = 0$, however, it still gives mass to larger values of τ^2. In contrast, the $\alpha \sim \text{Be}(2, 1)$ prior gives most mass to small values of τ^2 relative to σ^2, while the $\alpha \sim \text{Be}(2, 1)$ prior gives no mass to small τ^2 and has zero density at $\tau^2 = 0$.

4 Hypothesis testing

Apart from estimation of θ and α, one may also want to test hypotheses. The standard Bayesian approach is to compute the Bayes factor contrasting the likelihood of the data under two competing hypothesis (Jeffreys, 1961; Kass and Raftery, 1995), to quantify the strength of evidence for either of them.

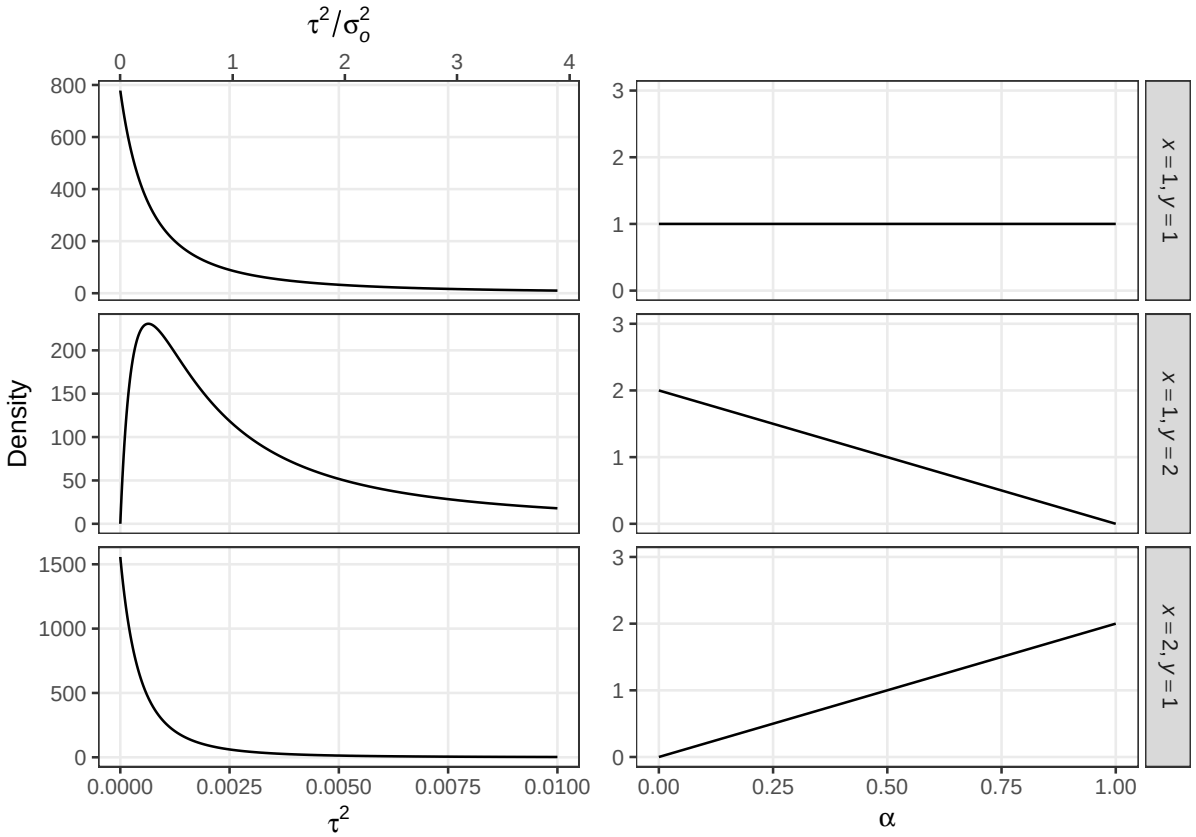

Figure 3: Beta priors on power parameter $\alpha \sim \text{Be}(x, y)$ (right) and corresponding priors on heterogeneity variance τ^2 (left) that lead to matching marginal posteriors for the effect sizes θ and θ_r. The variance of the original effect estimate $\sigma_o = 0.05^2$ from the "Labels" experiment is used for the transformation to the heterogeneity variance scale τ^2.

4.1 Hypotheses about the effect size

We may want to quantify the evidence for a non-zero effect size θ by testing $H_0 \colon \theta = 0$ to $H_1 \colon \theta \neq 0$. This requires specification of a prior distribution under H_1, a natural choice is to use the normalized power prior based on the original data from (1). The respective Bayes factor is then given by

$$\text{BF}_{01}(\hat{\theta}_r \mid x, y) = \frac{f(\hat{\theta}_r \mid H_0)}{f(\hat{\theta}_r \mid H_1)} = \frac{\text{N}(\hat{\theta}_r; 0, \sigma_r^2)}{\int_0^1 \text{N}(\hat{\theta}_r; \hat{\theta}_o, \sigma_o^2/\alpha)\ \text{Be}(\alpha; x, y)\, \text{d}\alpha}. \tag{10}$$

A reasonable choice for the prior of α under H_1 is a uniform $\alpha \sim \text{Be}(1, 1)$ distribution. However, it is worth noting that fixing $\alpha = 1$ leads to the *replication Bayes factor* under normality (Verhagen and Wagenmakers, 2014; Ly et al., 2018; Pawel and Held, 2022), that is, the Bayes factor contrasting a point null hypothesis to the posterior distribution of the effect size based on the original data (and in this case a uniform initial prior). A fixed $\alpha = 1$ can also be seen as the limiting case of a beta prior with $y = 1$ and $x \to \infty$. The power prior version of the replication Bayes factor is thus a generalization of the standard replication Bayes factor where the original data are to some extent discounted.

4.2 Hypotheses about the power parameter

In order to quantify compatibility of original and replication study we may want to test hypotheses regarding the power parameter α as it provides a link between the studies. For example, we might want to test $H_\text{d} \colon \alpha = 0$ ("different") vs. $H_\text{c} \colon \alpha = 1$ ("compatible"). One issue is that for a

flat initial prior $f(\theta) \propto 1$, the power prior with $\alpha = 0$ is not proper and so the resulting Bayes factor is only defined up to an arbitrary constant. Instead of the flat prior, we may thus choose an uninformative but proper initial prior, *e. g.* the unit-information prior (Kass and Wasserman, 1995)

$$\theta \sim \mathrm{N}(0, \kappa^2)$$

with κ^2 the variance from one (effective) observation. This leads to the Bayes factor

$$\mathrm{BF}_{\mathrm{dc}}(\hat{\theta}_r \mid \kappa^2) = \frac{f(\hat{\theta}_r \mid H_\mathrm{d})}{f(\hat{\theta}_r \mid H_\mathrm{c})} = \frac{\mathrm{N}(\hat{\theta}_r; 0, \sigma_r^2 + \kappa^2)}{\mathrm{N}(\hat{\theta}_r; s\hat{\theta}_o, \sigma_r^2 + s\sigma_o^2)} \tag{11}$$

with shrinkage factor $s = (\kappa^2/\sigma_o^2)/\{1 + (\kappa^2/\sigma_o^2)\}$.

An alternative approach that avoids the specification of a proper initial prior for θ is to specify priors for α under H_d and H_c. A suitable class of priors are $H_\mathrm{d}\colon \alpha \sim \mathrm{Be}(1, y)$ and $H_\mathrm{c}\colon \alpha \sim \mathrm{Be}(x, 1)$ with $x, y > 1$. Two examples with $x = 2$ and $y = 2$ are shown in Figure 3. These priors have the highest density at $\alpha = 0$, respectively, $\alpha = 1$, they have zero density at the opposite side, and they are monotonically decreasing, respectively, increasing. Instead of testing two composite hypotheses, one can also contrast the composite hypothesis $H_\mathrm{d}\colon \alpha \sim \mathrm{Be}(1, y_\mathrm{d})$ to the simple hypothesis $H_\mathrm{c}\colon \alpha = 1$, as also under this approach no proper initial prior has to be specified for the effect size θ. The resulting Bayes factor is then given by

$$\mathrm{BF}_{\mathrm{dc}}(\hat{\theta}_r \mid y) = \frac{\int_0^1 \mathrm{N}(\hat{\theta}_r; \hat{\theta}_o, \sigma_r^2 + \sigma_o^2/\alpha)\ \mathrm{Be}(\alpha; 1, y)\, \mathrm{d}\alpha}{\mathrm{N}(\hat{\theta}_r; \hat{\theta}_r, \sigma_r^2 + \sigma_o^2)}.$$

The parameter y determines how much mass small values of α receive under H_d. The simple hypothesis $H_\mathrm{d}\colon \alpha = 0$ can be seen as a limiting case when $y \to \infty$.

4.3 Example "Labels" (continued)

Table 1 displays the results of the proposed hypothesis tests applied to the three replications of the experiment "Labels". We see from the Bayes factors contrasting H_0 to H_1 that the data indicate absence of evidence for either hypothesis in the first replication, but decisive evidence for H_1 in the second and third replication. In all three cases, the Bayes factors are very close to the standard replication Bayes factors obtained from setting $\alpha = 1$.

Table 1: Hypothesis testing for replications of experiment "Labels" with original standardized mean difference effect estimate $\hat{\theta}_o = 0.2$ and standard error $\sigma_o = 0.05$. Shown replication effect estimates $\hat{\theta}_r$ with standard errors σ_r, Bayes factors contrasting $H_0\colon \theta = 0$ to $H_1\colon \theta \neq 0$ for different priors of α under H_1, and Bayes factor contrasting $H_d\colon \alpha = 0$ to $H_c\colon \alpha = 1$.

$\hat{\theta}_r$	σ_r	$\mathrm{BF}_{01}(\hat{\theta}_r \mid x = 1, y = 1)$	$\mathrm{BF}_{01}(\hat{\theta}_r \mid \alpha = 1)$	$\mathrm{BF}_{\mathrm{dc}}(\hat{\theta}_r \mid \kappa^2 = 2)$	$\mathrm{BF}_{\mathrm{dc}}(\hat{\theta}_r \mid y = 2)$
0.09	0.05	1/1.1	1.1	1/5.6	1.2
0.20	0.06	1/367	1/478	1/19	1/1.5
0.44	0.04	$< 1/1000$	$< 1/1000$	16	25

In order to compute the Bayes factor for testing H_d vs. H_c we need to specify a unit variance for the unit-information prior. A crude approximation for the variance of a standardized mean difference effect estimate is given by $\mathrm{Var}(\hat{\theta}_i) = 4/n_i$ with n_i the total sample size of the study, and assuming equal sample size in both groups (Hedges and Schauer, 2021, p. 5). We may thus set the variance of the unit-information prior to $\kappa^2 = 2$ since at least one observation from each group is required to estimate a standardized mean difference (assuming the variance is known). Based on this choice, the data provide strong, respectively substantial evidence H_c in the third and first replication study, while they indicate strong evidence for H_d in the second replication study.

To conclude, our analysis suggests that only the third replication was successful in the sense that it is compatible with the original study while also providing evidence against a null effect. The first replication is compatible but does not provide any evidence for a non-zero effect, whereas the second replication provides much evidence for an effect but is incompatible with the original study.

4.4 Asymptotics

It is of interest to investigate the asymptotic behavior of the Bayes factors (10) and (11). For instance, we may want to understand what happens when the sample size of the replication study n_r becomes larger. Assume that $\hat{\theta}_r$ is a consistent estimator of the true underlying effect size $\tilde{\theta}$, and that the standard error is inversely proportional to the square root of the sample size $\sigma_r = \kappa/\sqrt{n_r}$ with κ^2 some unit variance. As the replication sample size goes to infinity ($n_r \to \infty$), the estimate will converge in probability to the true effect size, and the standard error will go to zero.

Assuming that the original estimate was not zero ($\hat{\theta}_o \neq 0$), the Bayes factor (10) with $\alpha = 1$ (the replication Bayes factor) is information consistent, meaning that it will increasingly favor the correct hypothesis as the replication data accumulate. SP: also true for priors on α? Probably yes, can maybe do a Laplace approximation to prove it In contrast, the Bayes factor (11) does not grow unboundedly but converges to a constant

$$\text{BF}^*_{\text{dc}} = \sqrt{s\sigma_o^2/\kappa^2}\, \exp\left[-\frac{1}{2}\left\{\frac{\tilde{\theta}^2}{\kappa^2} - \frac{(\tilde{\theta} - \hat{\theta}_o)^2}{s\sigma_o^2}\right\}\right].$$

The amount of evidence one can find for either hypothesis thus depends on the original effect estimate $\hat{\theta}_o$, standard error σ_o, and the true effect size $\tilde{\theta}$. For instance, in the example from before we have an original effect estimate $\hat{\theta}_o = 0.2$ and standard error $\sigma_o = 0.05$. The bound is minimized for a true effect size of $\tilde{\theta} = (\hat{\theta}_o)/\{1 - (s\sigma_o^2)/\kappa^2\} = 0.21$, so the most extreme level we can obtain is $\text{BF}^*_{\text{dc}} = 1/28$. Even in an infinitely precise replication study, we cannot find more evidence for H_c.

4.5 Design

Assume now that the replication study has not yet been conducted and we want to determine its sample size. In the case of the replication Bayes factor under normality, Pawel and Held (2022) derived the probability of replication success in closed form under H_0 and H_1. Based on their result, standard Bayesian design analysis (Weiss, 1997; De Santis, 2004; Schönbrodt and Wagenmakers, 2017) can be conducted to determine the appropriate replication sample size. For the generalized replication Bayes factor (10), numerical integration or simulation is required to compute the probability of replication success as the marginal likelihood is not available in closed form under H_1.

It is also possible to derive the probability of replication success analytically for the power parameter Bayes factor (11). With some algebra, one can show that $\text{BF}_{\text{dc}} \leq \gamma$ is equivalent to

$$\left(\hat{\theta}_r - \frac{s\hat{\theta}_o(\sigma_r^2 + \kappa^2)}{\kappa^2 - s\sigma_o^2}\right)^2 \leq X = \frac{(\sigma_r^2 + \kappa^2)(\sigma_r^2 + s\sigma_o^2)}{\kappa^2 - s\sigma_o^2}\left\{\log\gamma^2 - \log\left(\frac{\sigma_r^2 + s\sigma_o^2}{\sigma_r^2 + \kappa^2}\right) - \frac{s^2\hat{\theta}_o^2}{s\sigma_o^2 - \kappa^2}\right\} \tag{12}$$

for $\kappa^2 > s\sigma_o^2$. Denote by m_i and v_i the mean and variance of $\hat{\theta}_r$ under hypothesis $i \in \{d, c\}$. The left hand side of (12) then follows scaled non-central chi-squared distribution under both hypotheses. The probability of replication success is hence given by

$$\Pr(\text{BF}_{\text{dc}} \leq \gamma \mid H_i) = \Pr\left(\chi^2_{1,\lambda_i} \leq X/v_i\right) \tag{13}$$

with non-centrality parameter

$$\lambda_i = \left(m_i - \frac{s\hat{\theta}_o(\sigma_r^2 + \kappa^2)}{\kappa^2 - s\sigma_o^2} \right)^2 / v_i.$$

To determine the replication sample size, we can now use (13) to compute the probability of replication success at a desired level γ over a grid of replication standard errors σ_r, and under either hypothesis H_d and H_c. The appropriate standard error is then chosen so that the probability for finding correct evidence is sufficiently high under the respective hypothesis, and sufficiently low under the wrong hypothesis. Subsequently, the standard error σ_r needs to be translated into a sample size, *e. g.* for standardized mean differences via the aforementioned approximation $n_r \approx 4/\sigma_r^2$.

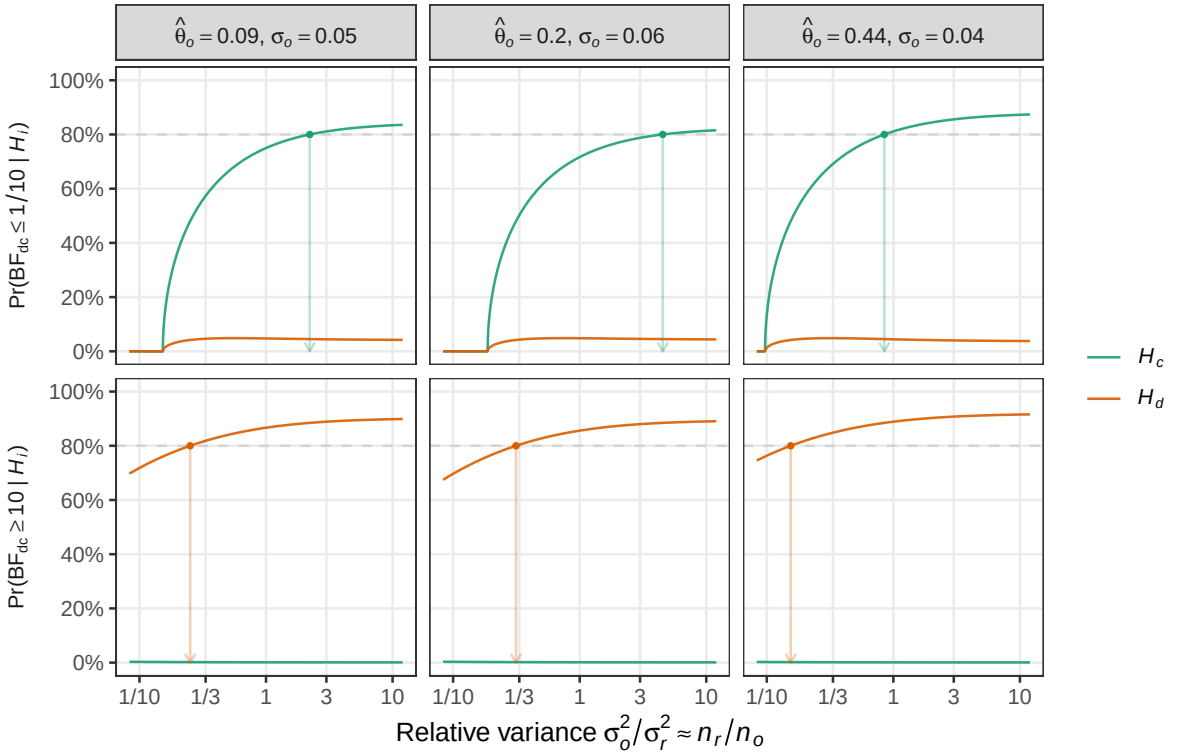

Figure 4: Probability of replication success as a function of relative variance for the three replications of experiment "Labels" regarded as original study. Relative sample size that correspond to a probability of 80% under the respective hypothesis are indicated by arrows.

4.6 Example "Labels" (continued)

Figure 4 illustrates Bayesian design analysis based on the power parameter Bayes factor for the three replication studies from the experiment "Labels" which are now each regarded as original studies. Shown is the probability of replication success as a function of the relative sample size. The curves look more or less similar for all three studies. We see from the lower panels that the probability for finding strong evidence for H_c is not much affected by the sample size of the replication study staying at almost zero under H_c, while under H_d it increases from about 75% to about 90%. In contrast, the top panels show that the probability for finding strong evidence for H_c rapidly increases under H_c and seems to level off at an asymptote. Under H_d the probability stays below 10% across the whole range.

The plots also display the required replication sample size to obtain strong evidence with probability of 80% under the correct hypothesis. We see that original studies with smaller standard errors require smaller replication sample sizes to achieve the same probability of replication

success. Under H_c the required sample sizes are larger than under H_d. However, while the probability of misleading evidence under H_c seems to be well controlled under the determined sample size, under H_d it stays roughly 5% for all three studies, and even for very large replication sample sizes.

For all three studies choosing the sample size based on finding strong evidence for H_c assuming H_c is true also guarantees appropriate error probabilities for finding strong evidence for H_d. At the same time, it seems that the probability for finding misleading evidence for H_c cannot be reduced below around 5% which might undesirably high for certain applications.

4.7 Connection to hypothesis testing in hierarchical models

As with parameter estimation, it is also of interest to know whether there is a correspondence between hypothesis tests in the power prior and the hierarchical modeling frameworks. Concerning the generalized replication Bayes factor from (10) testing $H_0\colon \theta = 0$ vs. $H_1\colon \theta \sim \mathrm{N}(\hat{\theta}_o, \sigma_o^2/\alpha)$, it is straightforward to show that it matches with the Bayes factor contrasting

$$\begin{aligned} H_0\colon \theta_* &= 0 \\ \tau^2 &= 0 \end{aligned} \qquad \text{to} \qquad \begin{aligned} H_1\colon \theta_* \,|\, \tau^2 &\sim \mathrm{N}(\hat{\theta}_o, \sigma_o^2 + \tau^2) \\ \tau^2 &= (1/\alpha - 1)\,(\sigma_o^2/2) \end{aligned}$$

in the hierachical framework. Under H_1 one may also assign a prior distribution to τ^2 which must satisfy the conditions (9) as for the matching posteriors. The Bayes factor thus quantifies the plausibility of the hypothesis H_0 that the global effect size θ_* is zero and that there is no effect size heterogeneity, relative to the hypothesis H_1 that θ_* follows the posterior based on the original data and an initial flat prior for θ_*.

The Bayes factor (11) testing $H_c\colon \alpha = 1$ to $H_d\colon \alpha = 0$ can be obtained in the hierarchical framework by contrasting

$$H_d\colon \theta_* \sim \mathrm{N}(0, \kappa^2), \tau^2 = 0 \qquad \text{to} \qquad H_c\colon \theta_* \sim \mathrm{N}(s\,\hat{\theta}_o, s\,\sigma_o^2), \tau^2 = 0$$

with $s = (\kappa^2/\sigma_o^2)/\{1 + (\kappa^2/\sigma_o^2)\}$.

5 Discussion

We showed how the power prior framework can be used for design and analysis of replication studies. The approach supplies analysts with a suite of methods for assessing effect sizes and study compatibility. We also show how the approach is connected to hierarchical modeling, and gave the conditions under which posterior distributions and hypothesis tests can be mapped from normal power prior models to the normal-normal hierarchical models. This connection helps explaining why even with two highly precise studies we were unable to make conclusive posterior inferences about the power parameter. Just as it is difficult to learn about a heterogeneity variance from two studies, it seems difficult to do learn about the power parameter from two studies alone. Failed replication efforts (in terms of null hypothesis testing) have led many to argue that researchers should shift their focus to effect size comparison. Our results show that in order to obtain conclusive evidence for effect size compatibility, extremely price estimates are required, which likely is more resource intensive than obtaining conclusive evidence for testing a null hypothesis.

References

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

Chen, M.-H. and Ibrahim, J. G. (2006). The relationship between the power prior and hierarchical models. *Bayesian Analysis*, 1(3). doi:10.1214/06-ba118.

Daniels, M. J. (1999). A prior for the variance in hierarchical models. *Canadian Journal of Statistics*, 27(3):567–578. doi:10.2307/3316112.

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.

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.

Gravestock, I. and Held, L. (2017). Adaptive power priors with empirical Bayes for clinical trials. *Pharmaceutical Statistics*, 16(5):349–360. doi:10.1002/pst.1814.

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:https://doi.org/10.1111/rssa.12688.

Held, L. (2020). A new standard for the analysis and design of replication studies (with discussion). *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(2):431–448. doi:10.1111/rssa.12493.

Held, L., Micheloud, C., and Pawel, S. (2022). The assessment of replication success based on relative effect size. URL `https://www.e-publications.org/ims/submission/AOAS/user/submissionFile/47896?confirm=532335fe`. to appear in *The Annals of Applied Statistics*.

Higgins, J. P. T. and Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11):1539–1558. doi:10.1002/sim.1186.

Ibrahim, J. G., Chen, M.-H., Gwon, Y., and Chen, F. (2015). The power prior: theory and applications. *Statistics in Medicine*, 34(28):3724–3749. doi:10.1002/sim.6728.

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.

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

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90(431):928–934. doi:10.1080/01621459.1995.10476592.

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.

Neuenschwander, B., Branson, M., and Spiegelhalter, D. J. (2009). A note on the power prior. *Statistics in Medicine*, 28(28):3562–3566. doi:10.1002/sim.3722.

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

Pawel, S. and Held, L. (2022). The sceptical Bayes factor for the assessment of replication success. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. doi:10.1111/rssb.12491.

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.

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.

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

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.

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.