Power Priors for Replication Studies

Samuel Pawel***, Frederik Aust**†**, Leonhard Held*****, Eric-Jan Wagenmakers**†

$\star$ Department of Biostatistics, University of Zurich

† Department of Psychological Methods, University of Amsterdam

E-mail: samuel.pawel@uzh.ch

January 30, 2023

This is a preprint which has not yet been peer reviewed.

Abstract

The ongoing replication crisis in science has increased interest in the methodology of replication studies. We propose a novel Bayesian analysis approach using power priors: The likelihood of the original study's data is raised to the power of α, and then used as the prior distribution in the analysis of the replication data. Posterior distribution and Bayes factor hypothesis tests related to the power parameter α quantify the degree of compatibility between the original and replication study. Inferences for other parameters, such as effect sizes, dynamically borrow information from the original study. The degree of borrowing depends on the conflict between the two studies. The practical value of the approach is illustrated on data from three replication studies, and the connection to hierarchical modeling approaches explored. We generalize the known connection between normal power priors and normal hierarchical models for fixed parameters and show that normal power prior inferences with a beta prior on the power parameter α align with normal hierarchical model inferences using a generalized beta prior on the relative heterogeneity variance I^2. The connection illustrates that power prior modeling is unnatural from the perspective of hierarchical modeling since it corresponds to specifying priors on a relative rather than an absolute heterogeneity scale.

Keywords: Bayes factor, Bayesian hypothesis testing, Bayesian parameter estimation, hierarchical models, historical data

1 Introduction

Power priors form a class of informative prior distributions that allow data analysts to incorporate historical data into a Bayesian analysis (Ibrahim et al., 2015). The most basic version of the power prior is obtained by updating an initial prior distribution with the likelihood of the historical data raised to the power of α, where α is usually restricted to the range from zero (i.e., complete discounting) to one (i.e., complete pooling). As such, the power parameter α specifies the degree to which historical data are discounted, thereby providing a quantitative compromise between the extreme positions of completely ignoring and fully trusting the historical data.

One domain where historical data are per definition available is the analysis of replication studies. One pertinent question in this domain is the extent to which a replication study has successfully replicated the result of an original study. Many methods have been proposed to address this question (Bayarri and Mayoral, 2002b; 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). Here we propose a new and conceptually straightforward approach, namely to construct a power prior for the data from the original study, and to use that prior to draw inferences from the data of the replication study.

Below we first show how power priors can be constructed from data of an original study under a meta-analytic framework (Section 2). We then shown how the power prior can be used for parameter estimation (Section 2.1), Bayes factor hypothesis testing (Section 2.2), and for the design of new replication studies (Section 2.3). Throughout, the methodology is illustrated by application to data from three replication studies which were part of a large-scale replication project (Protzko et al., 2020). In Section 3, we explore the connection to the alternative hierarchical modeling approach for incorporating the original data (Bayarri and Mayoral, 2002b,a; Pawel and Held, 2020), which has been previously used for evidence synthesis and compatibility assessment in replication settings. In doing so, we identify explicit conditions under which posterior distributions and tests can be reverse-engineered from one framework to the other. Essentially, power prior inferences using the commonly assigned beta prior on the power parameter α align with normal hierarchical model inferences if either a generalized F prior is assigned to the between-study heterogeneity variance τ^2 which scales with the variance of the original data, or if a generalized beta prior is assigned to the relative heterogeneity I^2. This perspective also explains the observed difficulty of making conclusive inferences about the power parameter α as it is difficult to make inferences about a variance from two observations alone, and also because the commonly assigned beta prior on α is entangled with the variance from the data.

2 Power prior modeling of replication studies

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" and "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 $\hat{\theta}_i$. The effect size may be adjusted for confounding variables, and depending on the outcome variable, a transformation may be required for the normal approximation to be accurate (e. g., a log-transformation for an odds ratio effect size). This is the same framework that is 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.

Updating of an (improper) flat initial prior $f(\theta) \propto 1$ by the likelihood of the original data raised to a (fixed) power parameter α leads to the normalized power prior

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

as first proposed by Duan et al. (2005), see also Neuenschwander et al. (2009). There are different ways to specify α. The simplest approach fixes α to an *a priori* reasonable value, possibly informed by background knowledge about the similarity of the two studies. Another option is to use the empirical Bayes estimate (Gravestock and Held, 2017), that is, the value of α that maximizes the likelihood of the replication data marginalized over the power prior. Finally, it is also possible to specify a prior distribution for α, the most common choice being a beta distribution $\alpha \mid x, y \sim \mathrm{Be}(x, y)$ for a normalized power prior conditional on α as in (1). This approach leads to a joint prior for the effect size θ and power parameter α with density

$$f(\theta, \alpha \mid \hat{\theta}_o, x, y) = \mathrm{N}(\theta \mid \hat{\theta}_o, \sigma_o^2/\alpha)\ \mathrm{Be}(\alpha \mid x, y) \tag{2}$$

where $\mathrm{N}(\cdot \mid m, v)$ is the normal density function with mean m and variance v, and $\mathrm{Be}(\cdot \mid x, y)$ is the beta density with parameters x and y. The uniform distribution ($x = 1$, $y = 1$) is often recommended as the default choice (Ibrahim et al., 2015). We note that α does not have to be restricted to the unit interval but could also be treated as a relative precision parameter (Held and Sauter, 2017). We will, however, not consider such an approach since power parameters $\alpha > 1$ lead to priors with more information than what was actually supplied by the original study.

2.1 Parameter estimation

Updating the prior (2) with the likelihood of the replication data leads to the posterior distribution

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

The normalizing constant

$$f(\hat{\theta}_r \mid \hat{\theta}_o, x, y) = \int_0^1 \mathrm{N}(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_r^2 + \sigma_o^2/\alpha)\ \mathrm{Be}(\alpha \mid x, y)\, \mathrm{d}\alpha \tag{4}$$

is generally not available in closed form but requires numerical integration with respect to α. If inference concerns only one parameter, a marginal posterior distribution for either α or θ can be obtained by integrating out the respective nuisance parameter from (3). In the case of the power parameter α, this leads to

$$f(\alpha \mid \hat{\theta}_r, \hat{\theta}_o, x, y) = \frac{\mathrm{N}(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_r^2 + \sigma_o^2/\alpha)\ \mathrm{Be}(\alpha \mid x, y)}{f(\hat{\theta}_r \mid \hat{\theta}_o, x, y)} \tag{5}$$

whereas for the effect size θ, this gives

$$f(\theta \,|\, \hat{\theta}_r, \hat{\theta}_o, x, y) = \frac{\mathrm{N}(\hat{\theta}_r \,|\, \theta, \sigma_r^2)\, \mathrm{B}(x + 1/2, y)}{f(\hat{\theta}_r \,|\, \hat{\theta}_o, x, y)\, \sqrt{2\pi\sigma_o^2}\, \mathrm{B}(x, y)}\, M\left\{x + 1/2, x + y + 1/2, -\frac{(\hat{\theta}_o - \theta)^2}{2\sigma_o^2}\right\}$$

with $\mathrm{B}(z, w) = \int_0^1 t^{z-1}(1 - t)^{w-1}\, \mathrm{d}t = \{\Gamma(z)\Gamma(w)\}/\Gamma(z + w)$ the beta function and $M(a, b, z) = \{\int_0^1 \exp(zt)t^{a-1}(1 - t)^{b-a-1}\, \mathrm{d}t\}/\mathrm{B}(b - a, a)$ the confluent hypergeometric function (Abramowitz and Stegun, 1965, chapters 6 and 13).

2.1.1 Example "Labels"

We now illustrate the methodology on data from the large-scale replication project by Protzko et al. (2020). The project featured an experiment called "Labels" for which the original study reported the following conclusion: "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" (Protzko et al., 2020, p. 17). This conclusion was based on a standardized mean difference effect estimate $\hat{\theta}_o = 0.21$ and standard error $\sigma_o = 0.05$ obtained from 1577 participants. Subsequently, four replication studies were conducted, three of them by a different laboratory than the original one, and all employing large sample sizes. Since the same original study was replicated by three independent laboratories, this is an instance of a "multisite" replication design (Mathur and VanderWeele, 2020). While in principle it would be possible to analyze all of these studies jointly, we will show separate analyses for each pair of original and replication study as it reflects the typical situation of only one replication study being conducted per original study. Section 4 discusses possible extensions of the power prior approach for joint analyses in multisite designs.

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. The first replication found an effect estimate which was smaller than the original one ($\hat{\theta}_{r1} = 0.09$), whereas the other two replications found effect estimates that were either identical ($\hat{\theta}_{r2} = 0.21$) or larger ($\hat{\theta}_{r3} = 0.44$) than that reported in the original study. This is reflected in the marginal posterior distributions of the power parameter α, shown in the bottom right panel of Figure 1. That is, the marginal distribution of the first replication (yellow) is slightly peaked around $\alpha = 0.2$ suggesting some incompatibility with the original study. In contrast, the second replication shows a marginal distribution (green) which is monotonically increasing so that the value $\alpha = 1$ receives the highest support, thereby indicating compatibility of the two studies. Finally, the marginal distribution of the third replication (blue) is sharply peaked around $\alpha = 0.05$ indicating strong conflict between this replication and the original study. The sharply peaked posterior is in stark contrast to the relatively diffuse posteriors of the first and second replications which hardly changed from the uniform prior. This is consistent with the asymptotic behavior of normalized power priors identified in Pawel et al. (2022); In case of data incompatibility, normalized power priors with beta prior assigned to α permit arbitrarily peaked posteriors for small values of α, whereas in case of data compatibility there is a limiting posterior for α that hardly differs from the prior. The limiting posterior is in this case a $\mathrm{Be}(3/2, 1)$ distribution, whose density is indicated by the

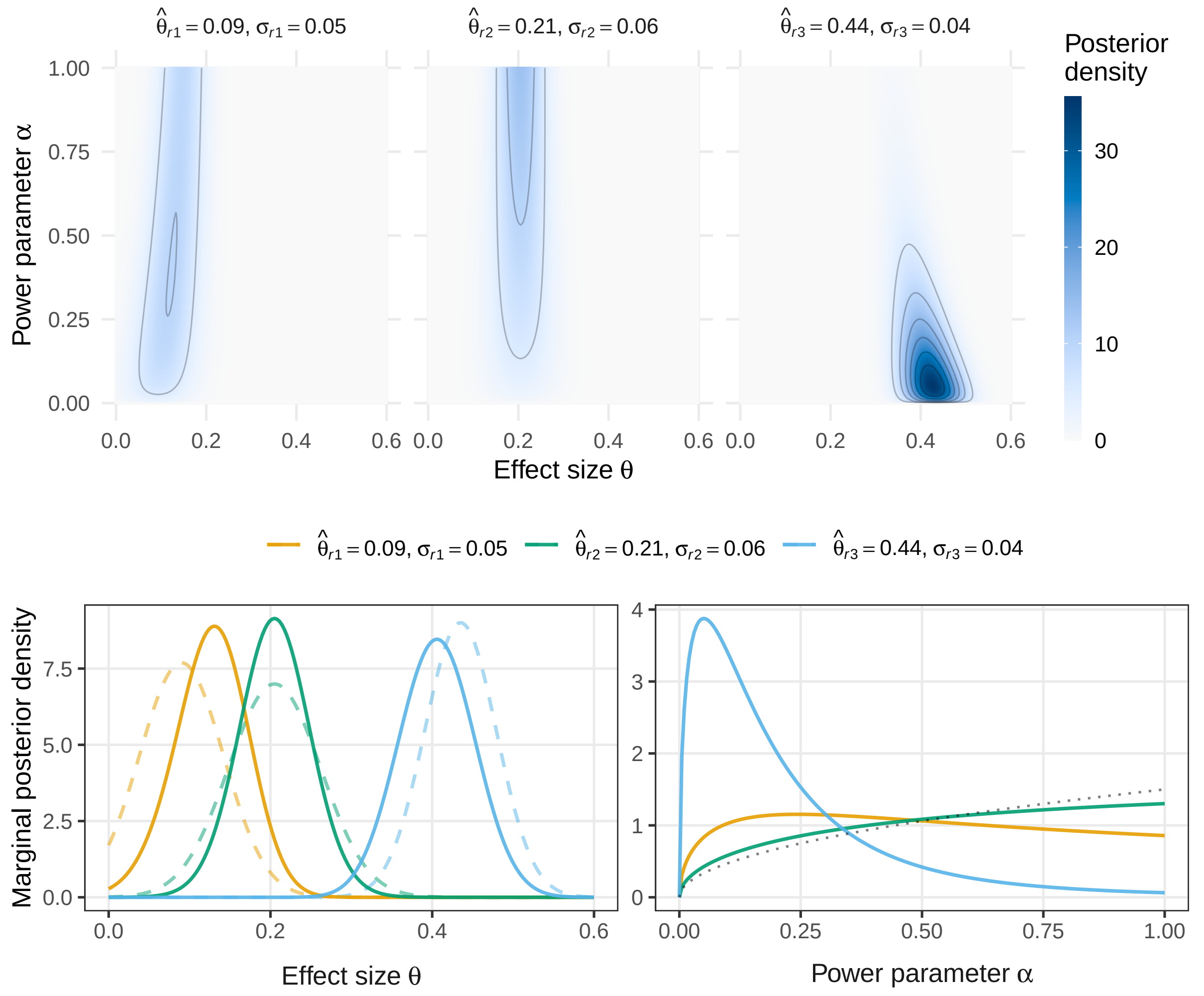

Figure 1: Analysis of three replication studies of the "Labels" experiment from Protzko et al. (2020). Shown are joint (top) and marginal (bottom) posterior distributions of effect size θ and power parameter α. A power prior for the effect size θ is constructed from the original effect estimate $\hat{\theta}_o = 0.21$ (with standard error $\sigma_o = 0.05$) and an initial flat prior $f(\theta) \propto 1$. The power parameter α is assigned a uniform $\text{Be}(1, 1)$ prior distribution. The dashed lines depict the posterior density for the effect size θ when the replication data are analysed in isolation without incorporation of the original data. The dotted line represents the limiting posterior density of the power parameter α for perfectly agreeing original and replication studies.

dotted line. One can see, that the (green) posterior from the second replication is relatively close to the limiting posterior, despite its finite sample size.

The bottom left panel of Figure 1 shows the marginal posterior distribution of the effect size θ. Shown is also the posterior distribution of θ when the replication data are analyzed in isolation (dashed line), to see the information gain from incorporating the original data via a power prior. The degree of compatibility with the replication study influences how much information is borrowed from the original study. For instance, the (green) marginal posterior density based on the most compatible replication ($\hat{\theta}_{r2} = 0.21$) is the most concentrated among the three replications, despite the standard error being the largest (i. e., $\sigma_{r2} = 0.06$). In contrast, the (blue) marginal posterior of the most conflicting estimate (i. e., $\hat{\theta}_{r3} = 0.44$) borrows less information and consequently yields the least peaked posterior, despite the standard error being the smallest (i. e., $\sigma_{r3} = 0.04$). In this case, the conflict with the original study even inflates the variance of posterior compared to the isolated replication posterior given by dashed blue line.

2.2 Hypothesis testing

Apart from the estimation of θ and α, one may also wish to test hypotheses regarding these parameters. A principled Bayesian approach is to quantify the strength of evidence that the data provide for two competing hypotheses $\mathcal{H}_j$ and $\mathcal{H}_k$ about the parameters, say, by computing the Bayes factor $\mathrm{BF}_{jk}(\hat{\theta}_r) = f(\hat{\theta}_r \,|\, \mathcal{H}_j)/f(\hat{\theta}_r \,|\, \mathcal{H}_k)$, with

$$f(\hat{\theta}_r \,|\, \mathcal{H}_i) = \int \mathrm{N}(\hat{\theta}_r \,|\, \theta, \sigma_r^2)\, f(\theta, \alpha \,|\, \mathcal{H}_i)\, \mathrm{d}\theta\, \mathrm{d}\alpha \tag{6}$$

the marginal likelihood of the replication data $\hat{\theta}_r$ under hypothesis $i \in \{k, j\}$. The Bayes factor can either be interpreted as the updating factor of the prior odds to the posterior odds of the hypotheses $\mathcal{H}_j$ and $\mathcal{H}_k$, or as the relative accuracy with which $\mathcal{H}_j$ and $\mathcal{H}_k$ predict the data (Good, 1958; Jeffreys, 1961; Kass and Raftery, 1995).

2.2.1 Hypotheses about the effect size θ

One is often interested in quantifying the evidence for a non-zero effect size θ by testing $\mathcal{H}_0\colon \theta = 0$ against $\mathcal{H}_1\colon \theta \neq 0$. This requires the specification of a prior distribution for θ under $\mathcal{H}_1$, and a natural choice is to use the normalized power prior based on the original data along with a beta prior for the power parameter as in (2). The associated Bayes factor is then given by

$$\mathrm{BF}_{01}(\hat{\theta}_r \,|\, x, y) = \frac{f(\hat{\theta}_r \,|\, \mathcal{H}_0)}{f(\hat{\theta}_r \,|\, \mathcal{H}_1)} = \frac{\mathrm{N}(\hat{\theta}_r \,|\, 0, \sigma_r^2)}{\int_0^1 \mathrm{N}(\hat{\theta}_r \,|\, \hat{\theta}_o, \sigma_r^2 + \sigma_o^2/\alpha)\, \mathrm{Be}(\alpha \,|\, x, y)\, \mathrm{d}\alpha}. \tag{7}$$

An intuitively reasonable choice for the prior of α under $\mathcal{H}_1$ is a uniform $\alpha \sim \mathrm{Be}(1, 1)$ distribution. However, it is worth noting that assigning a point mass $\alpha = 1$ leads to

$$\mathrm{BF}_{01}(\hat{\theta}_r \,|\, \alpha = 1) = \frac{\mathrm{N}(\hat{\theta}_r \,|\, 0, \sigma_r^2)}{\mathrm{N}(\hat{\theta}_r \,|\, \hat{\theta}_o, \sigma_o^2 + \sigma_r^2)}, \tag{8}$$

which is 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 $y > 0$ and $x \to \infty$. The power prior version of the replication Bayes factor is thus a generalization of the standard replication Bayes factor, one that allows the original data to be discounted to some degree.

2.2.2 Hypotheses about the power parameter α

In order to quantify the compatibility between the original and the replication study we may be interested in testing hypotheses regarding the power parameter α. For example, we may wish to test $\mathcal{H}_{\mathrm{c}} \colon \alpha = 1$ ("compatible") versus $\mathcal{H}_{\mathrm{d}} \colon \alpha < 1$ ("different"). One approach is to assign a point prior $\mathcal{H}_{\mathrm{d}} \colon \alpha = 0$. This leads to the issue 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 assign an uninformative but proper initial prior to θ, for instance, a unit-information prior $\theta \sim \mathrm{N}(0, \kappa^2)$ with κ^2 the variance from one (effective) observation (Kass and Wasserman, 1995) as it encodes minimal prior information about the direction or magnitude of the effect size (Best et al., 2021). This leads to the Bayes factor

$$\mathrm{BF}_{\mathrm{dc}}(\hat{\theta}_r \,|\, \kappa^2) = \frac{f(\hat{\theta}_r \,|\, \mathcal{H}_{\mathrm{d}})}{f(\hat{\theta}_r \,|\, \mathcal{H}_{\mathrm{c}})} = \frac{\mathrm{N}(\hat{\theta}_r \,|\, 0, \sigma_r^2 + \kappa^2)}{\mathrm{N}(\hat{\theta}_r \,|\, s\hat{\theta}_o, \sigma_r^2 + s\sigma_o^2)} \tag{9}$$

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

An alternative approach that avoids the specification of a proper initial prior for θ is to assign a prior to α under $\mathcal{H}_{\mathrm{d}}$. A suitable class of priors is given by $\alpha \,|\, \mathcal{H}_{\mathrm{d}} \sim \mathrm{Be}(1, y)$ with $y > 1$. The $\mathrm{Be}(1, y)$ prior has its highest density at $\alpha = 0$ and is monotonically decreasing. The parameter y determines how much mass is assigned to small values of α. The resulting Bayes factor is then given by

$$\mathrm{BF}_{\mathrm{dc}}(\hat{\theta}_r \,|\, y) = \frac{\int_0^1 \mathrm{N}(\hat{\theta}_r \,|\, \hat{\theta}_o, \sigma_r^2 + \sigma_o^2/\alpha)\ \mathrm{Be}(\alpha \,|\, 1, y)\, \mathrm{d}\alpha}{\mathrm{N}(\hat{\theta}_r \,|\, \hat{\theta}_o, \sigma_r^2 + \sigma_o^2)}, \tag{10}$$

and the simple hypothesis $\mathcal{H}_{\mathrm{d}} \colon \alpha = 0$ can be seen as a limiting case when $y \to \infty$.

2.2.3 Example "Labels" (continued)

Table 1 displays the results of the proposed hypothesis tests applied to the three replications of the "Labels" experiment. The Bayes factors contrasting $\mathcal{H}_0 \colon \theta = 0$ to $\mathcal{H}_1 \colon \theta \neq 0$ with normalized power prior with uniform prior for the power parameter α under the alternative (column $\mathrm{BF}_{01}(\hat{\theta}_r \,|\, x = 1, y = 1)$) indicate absence of evidence for either hypothesis in the first replication, but decisive evidence for $\mathcal{H}_1$ in the second and third replication. In all three cases, the Bayes factors are close to the standard replication Bayes factors with $\alpha = 1$ under the alternative (column $\mathrm{BF}_{01}(\hat{\theta}_r \,|\, \alpha = 1)$).

In order to compute the Bayes factor for testing $\mathcal{H}_{\mathrm{d}} \colon \alpha = 0$ versus $\mathcal{H}_{\mathrm{c}} \colon \alpha = 1$ 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 Bayes factors $\mathrm{BF}_{\mathrm{dc}}(\hat{\theta}_r \,|\, \kappa^2 = 2)$ in Table 1 indicate that the data provide substantial and strong evidence for $\mathcal{H}_{\mathrm{c}}$ in the first

Table 1: Hypothesis tests for replications of experiment "Labels" with original standardized mean difference effect estimate $\hat{\theta}_o = 0.21$ and standard error $\sigma_o = 0.05$. Shown are replication effect estimates $\hat{\theta}_r$ with standard errors σ_r, Bayes factors contrasting $\mathcal{H}_0 \colon \theta = 0$ to $\mathcal{H}_1 \colon \theta \neq 0$ with either uniform prior ($x = 1$, $y = 1$) assigned to α or fixed $\alpha = 1$ under $\mathcal{H}_1$, and Bayes factors contrasting $\mathcal{H}_{\mathrm{d}} \colon \alpha < 1$ to $\mathcal{H}_{\mathrm{c}} \colon \alpha = 1$ with either initial unit-information prior $\theta \sim \mathrm{N}(0, \kappa^2)$ and $\mathcal{H}_{\mathrm{d}} \colon \alpha = 0$ or $\alpha \,|\, \mathcal{H}_{\mathrm{d}} \sim \mathrm{Be}(1, y)$ prior under $\mathcal{H}_{\mathrm{d}}$.

	$\hat{\theta}_r$	σ_r	$\mathrm{BF}_{01}(\hat{\theta}_r \,\|\, x = 1, y = 1)$	$\mathrm{BF}_{01}(\hat{\theta}_r \,\|\, \alpha = 1)$	$\mathrm{BF}_{\mathrm{dc}}(\hat{\theta}_r \,\|\, \kappa^2 = 2)$	$\mathrm{BF}_{\mathrm{dc}}(\hat{\theta}_r \,\|\, y = 2)$
1	0.09	0.05	1/1.1	1.1	1/5.6	1.2
2	0.21	0.06	1/367	1/478	1/19	1/1.5
3	0.44	0.04	< 1/1000	< 1/1000	16	25

and second replication study, respectively, whereas the data indicate strong evidence for $\mathcal{H}_{\mathrm{d}}$ in the third replication study. The Bayes factor $\mathrm{BF}_{\mathrm{dc}}(\hat{\theta}_r \,|\, y = 2)$ in the right-most column with $\alpha \,|\, \mathcal{H}_{\mathrm{d}} \sim \mathrm{Be}(1, 2)$ prior assigned under the hypothesis $\mathcal{H}_{\mathrm{d}}$ indicates absence of evidence for either hypothesis in the first and second replication, but strong evidence for $\mathcal{H}_{\mathrm{d}}$ in the third replication. Compared to the Bayes factor with $\mathcal{H}_{\mathrm{c}} \colon \alpha = 0$, the results are thus more ambiguous for the first two replications but more compelling for the third replication.

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

2.2.4 Bayes factor asymptotics

Some of the Bayes factors in the previous example provided only modest evidence for the test-relevant hypotheses despite the large sample sizes in original and replication study. It is therefore of interest to understand the asymptotic behavior of the proposed Bayes factors. For instance, we may wish to understand what happens when the standard error of the replication study σ_r becomes arbitrarily small (through an increase in sample size). Assume that $\hat{\theta}_r$ is a consistent estimator of its true underlying effect size θ_r, so that as the standard error σ_r goes to zero, the estimate will converge in probability to the true effect size θ_r.

The limiting Bayes factors for testing the effect size θ from (7) and (8) are then given by

$$\lim_{\sigma_r \downarrow 0} \mathrm{BF}_{01}(\hat{\theta}_r \,|\, x, y) = \frac{\delta(\theta_r) \sqrt{2\pi}\, \mathrm{B}(x, y)}{\mathrm{B}(x + 1/2, y)}\, M \left\{ x + 1/2, x + y + 1/2, -\frac{(\theta_r - \hat{\theta}_o)^2}{2\sigma_o^2} \right\}^{-1}$$

and

$$\lim_{\sigma_r \downarrow 0} \mathrm{BF}_{01}(\hat{\theta}_r \,|\, \alpha = 1) = \frac{\delta(\theta_r)}{\mathrm{N}(\theta_r \,|\, \hat{\theta}_o, \sigma_o^2)},$$

with $\delta(\cdot)$ the Dirac delta function. Both Bayes factors are hence consistent (Bayarri et al., 2012) in the sense that they indicate overwhelming evidence for the correct hypothesis (i. e., the Bayes factors go to infinity/zero if the true effect size θ_r is zero/non-zero). In contrast, the Bayes factors for testing the power parameter α from (9) and (10) converge to positive constants

$$\lim_{\sigma_r \downarrow 0} \mathrm{BF}_{\mathrm{dc}}(\theta_r \,|\, \kappa^2) = \sqrt{1-s}\, \exp\left[-\frac{1}{2}\left\{\frac{\theta_r^2}{\kappa^2} - \frac{(\theta_r - s\hat{\theta}_o)^2}{s\sigma_o^2}\right\}\right] \tag{11}$$

and

$$\lim_{\sigma_r \downarrow 0} \mathrm{BF}_{\mathrm{dc}}(\theta_r \,|\, y) = \frac{\mathrm{B}(3/2, y)}{\mathrm{B}(1, y)}\, M\left\{y, y + 3/2, \frac{(\theta_r - \hat{\theta}_o)^2}{2\sigma_o^2}\right\}. \tag{12}$$

The amount of evidence one can find for either hypothesis thus depends on the original effect estimate $\hat{\theta}_o$, the standard error σ_o, and the true effect size θ_r. For instance, in the "Labels" experiment we have an original effect estimate $\hat{\theta}_o = 0.21$ and standard error $\sigma_o = 0.05$. The bound (11) is minimized for a true effect size equal to the original effect estimate $\theta_r = \hat{\theta}_o = 0.21$, so the most extreme level we can obtain is $\lim_{\sigma_r \downarrow 0} \mathrm{BF}_{\mathrm{dc}}(\theta_r \,|\, \kappa^2 = 2) = 1/28$. Similarly, the bound (12) is minimized for $\theta_r = \hat{\theta}_o = 0.21$ since then the confluent hypergeometric function term becomes one, leading to $\lim_{\sigma_r \downarrow 0} \mathrm{BF}_{\mathrm{dc}}(\theta_r \,|\, y = 2) = \mathrm{B}(3/2, y)/\mathrm{B}(1, y) = 1/1.9$. Even in a perfectly precise replication study we cannot find more evidence.

While the Bayes factors (9) and (10) are inconsistent if the replication data become arbitrarily informative, the situation is different when also the original data become arbitrarily informative (reflected by also the standard error σ_o going to zero and the original effect estimate $\hat{\theta}_o$ converging to its true effect size θ_o). The Bayes factor with $\mathcal{H}_{\mathrm{d}} \colon \alpha = 0$ from (9) is then consistent as the the limit (11) goes correctly to infinity/zero if the true effect size of the replication study θ_r is different/equivalent from the true effect size of the original study θ_o. In contrast, the Bayes factor with $\alpha \,|\, \mathcal{H}_{\mathrm{d}} \sim \mathrm{Be}(1, y)$ from (10) is still inconsistent since it only shows the correct asymptotic behavior when the true effect sizes are unequal (i. e., the Bayes factor goes to infinity) but not when the effect sizes are equivalent, in which case it is still bounded by $\mathrm{B}(3/2, y)/\mathrm{B}(1, y)$.

2.3 Design of replication studies

Now assume that the replication study has not yet been conducted and we wish to plan for a suitable sample size for a hypothesis test as described previously. In the case of the replication Bayes factor under normality (8), Pawel and Held (2022) derived the probability of replication success in closed form under $\mathcal{H}_0$ and $\mathcal{H}_1$. 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 (7), numerical integration or simulation is required to compute the probability of replication success as the marginal likelihood is not available in closed form under $\mathcal{H}_1$ in general.

It is also possible to derive the probability of replication success at some level γ analytically for the Bayes factor (9). With some algebra, one can show that $\mathrm{BF}_{\mathrm{dc}} \leq \gamma$ is equivalent to

$$\left\{\hat{\theta}_r - \frac{\hat{\theta}_o\,(\sigma_r^2 + \kappa^2)}{\kappa^2}\right\}^2 \leq X \tag{13}$$

with

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

and $s = \kappa^2/(\sigma_o^2 + \kappa^2)$. Denote by m_i and v_i the mean and variance of $\hat{\theta}_r$ under hypothesis $i \in \{\mathrm{d}, \mathrm{c}\}$. The left hand side of (13) then follows a scaled non-central chi-squared distribution under both hypotheses. Hence the probability of replication success is given by

$$\Pr(\mathrm{BF}_{\mathrm{dc}} \leq \gamma \mid \mathcal{H}_i) = \Pr\left(\chi^2_{1,\lambda_i} \leq X/v_i\right) \tag{14}$$

with non-centrality parameter

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

To determine the replication sample size, we can now use (14) 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 $\mathcal{H}_{\mathrm{d}}$ and $\mathcal{H}_{\mathrm{c}}$. The appropriate standard error σ_r 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 $n_r \approx 4/\sigma_r^2$.

2.3.1 Example "Labels" (continued)

Figure 2 illustrates Bayesian design analysis based on the power parameter Bayes factor $\mathrm{BF}_{\mathrm{dc}}(\hat{\theta}_r \mid \kappa^2)$ from (9). The three replication studies from the experiment "Labels" are now regarded as original studies, and each column of the figure shows the corresponding design analyses for future replications. In each plot, the probability for finding strong evidence for $\mathcal{H}_{\mathrm{c}}\colon \alpha = 1$ (top) or $\mathcal{H}_{\mathrm{d}}\colon \alpha = 0$ (bottom) is shown as a function of the relative sample size. In both cases, the probability is computed assuming that either $\mathcal{H}_{\mathrm{c}}$ (blue) or $\mathcal{H}_{\mathrm{d}}$ (yellow) is true.

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 $\mathcal{H}_{\mathrm{d}}$ is not much affected by the sample size of the replication study; it stays at almost zero under $\mathcal{H}_{\mathrm{c}}$, while under $\mathcal{H}_{\mathrm{d}}$ it increases from about 75% to about 90%. In contrast, the top panels show that the probability for finding strong evidence for $\mathcal{H}_{\mathrm{c}}$ rapidly increases under $\mathcal{H}_{\mathrm{c}}$ and seems to level off at an asymptote. Under $\mathcal{H}_{\mathrm{d}}$ the probability stays below 5% across the whole range.

The plots also display the required relative 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 relative sample sizes in the replication to achieve the same probability of replication success. Under $\mathcal{H}_{\mathrm{c}}$ the required relative sample sizes are larger than under $\mathcal{H}_{\mathrm{d}}$. However, while the probability of misleading evidence under $\mathcal{H}_{\mathrm{c}}$ seems to be well controlled under the determined sample size, under $\mathcal{H}_{\mathrm{d}}$ it stays roughly 5% for all three studies, and even for very

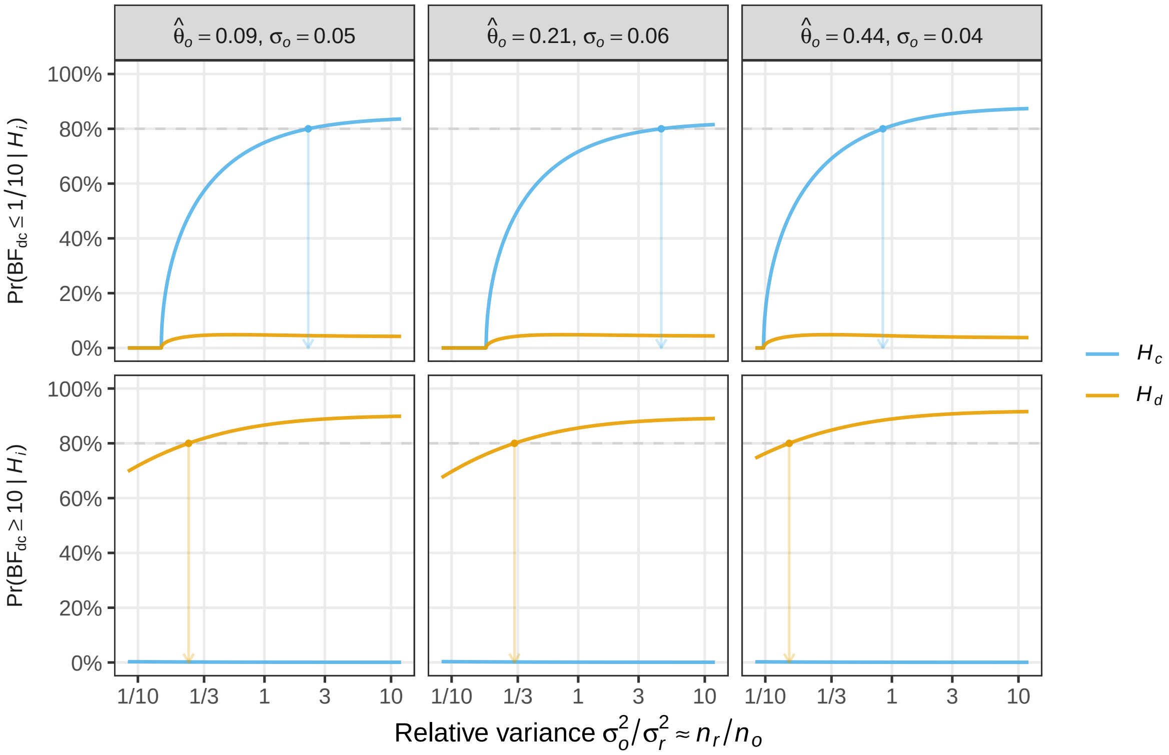

Figure 2: Probability of replication success as a function of relative variance for the three replications of experiment “Labels” regarded as original study. The arrows point to the relative variance associated with an 80% probability under the respective hypotheses.

large replication sample sizes. Choosing the sample size based on finding strong evidence for $\mathcal{H}_\mathrm{c}$ assuming $\mathcal{H}_\mathrm{c}$ is true thus guarantees appropriate error probabilities for finding strong evidence for $\mathcal{H}_\mathrm{d}$ in all three studies. At the same time, it seems that the probability for finding misleading evidence for $\mathcal{H}_\mathrm{c}$ cannot be reduced below around 5% which might undesirably high for certain applications.

3 Connection to hierarchical modeling of replication studies

Hierarchical modeling is another approach that allows for the incorporation of historical data in Bayesian analyses; moreover, hierarchical models have previously been used in the replication setting (Bayarri and Mayoral, 2002b,a; Pawel and Held, 2020). We will now investigate how the hierarchical modeling approach is related to the power prior approach in the analysis of replication studies, both in parameter estimation and hypothesis testing.

3.1 Connection to parameter estimation in hierarchical models

Assume a hierarchical model

$$\hat{\theta}_i \,|\, \theta_i \sim \mathrm{N}(\theta_i, \sigma_i^2) \tag{15a}$$

$$\theta_i \,|\, \theta_* \sim \mathrm{N}(\theta_*, \tau^2) \tag{15b}$$

$$f(\theta_*) \propto k \tag{15c}$$

where for study $i \in \{o, r\}$ the effect estimate $\hat{\theta}_i$ is normally distributed around a study specific effect size θ_i which itself is normally distributed around an overall effect size θ_*. The heterogeneity variance τ^2 determines the similarity of the study specific effect sizes θ_i. The overall effect size θ_* is assigned an (improper) flat prior $f(\theta_*) \propto k$, for some $k > 0$, which is a common approach in hierarchical modeling of effect estimates (Röver et al., 2021).

We show in Appendix A that under the hierarchical model (15) the marginal posterior distribution of the replication specific effect size θ_r is given by

$$\theta_r \,|\, \hat{\theta}_o, \hat{\theta}_r, \tau^2 \sim \mathrm{N}\left(\frac{\hat{\theta}_r/\sigma_r^2 + \hat{\theta}_o/(2\tau^2 + \sigma_o^2)}{1/\sigma_r^2 + 1/(2\tau^2 + \sigma_o^2)}, \frac{1}{1/\sigma_r^2 + 1/(2\tau^2 + \sigma_o^2)}\right), \tag{16}$$

that is, a normal distribution whose mean is a weighted average of the replication effect estimate $\hat{\theta}_r$ and the original effect estimate $\hat{\theta}_o$. The amount of shrinkage of the replication towards the original effect estimate depends on how large the replication standard error σ_r is relative to the heterogeneity variance τ^2 and the original standard error σ_o. There exists a correspondence between the posterior for the replication effect size θ_r from hierarchical model (16) and the posterior for the effect size θ under the power prior approach. Specifically, note that under the power prior and for a fixed power parameter α, the posterior of the effect size θ is given by

$$\theta \,|\, \hat{\theta}_o, \hat{\theta}_r, \alpha \sim \mathrm{N}\left(\frac{\hat{\theta}_r/\sigma_r^2 + (\hat{\theta}_o\alpha)/\sigma_o^2}{1/\sigma_r^2 + \alpha/\sigma_o^2}, \frac{1}{1/\sigma_r^2 + \alpha/\sigma_o^2}\right). \tag{17}$$

The hierarchical posterior (16) and the power prior posterior (17) thus match if and only if

$$\alpha = \frac{\sigma_o^2}{2\tau^2 + \sigma_o^2}, \tag{18}$$

respectively

$$\tau^2 = \left(\frac{1}{\alpha} - 1\right)\frac{\sigma_o^2}{2}, \tag{19}$$

which was first shown by Chen and Ibrahim (2006). For instance, a power prior model with $\alpha = 1$ corresponds to a hierarchical model with $\tau^2 = 0$, and a hierarchical model with $\tau^2 \to \infty$ corresponds to a power prior model with $\alpha \downarrow 0$. In between these two extremes, however, α has to be interpreted as a relative measure of heterogeneity since the transformation to τ^2 involves a scaling by the variance σ_o^2 of the original effect estimate. For this reason, there is a direct correspondence between α and the popular relative heterogeneity measure $I^2 = \tau^2/(\tau^2 + \sigma_o^2)$

(Higgins and Thompson, 2002) computed from τ^2 and the variance of the original estimate σ_o^2, that is,

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

with inverse of the same functional form. Figure 3 shows α and the corresponding τ^2 and I^2 values which lead to matching posteriors. The relationship between I^2 and α seems not too far off from linear. Therefore, a rough and ready heuristic to connect power priors to hierarchical models is $\alpha \approx 1 - I^2$.

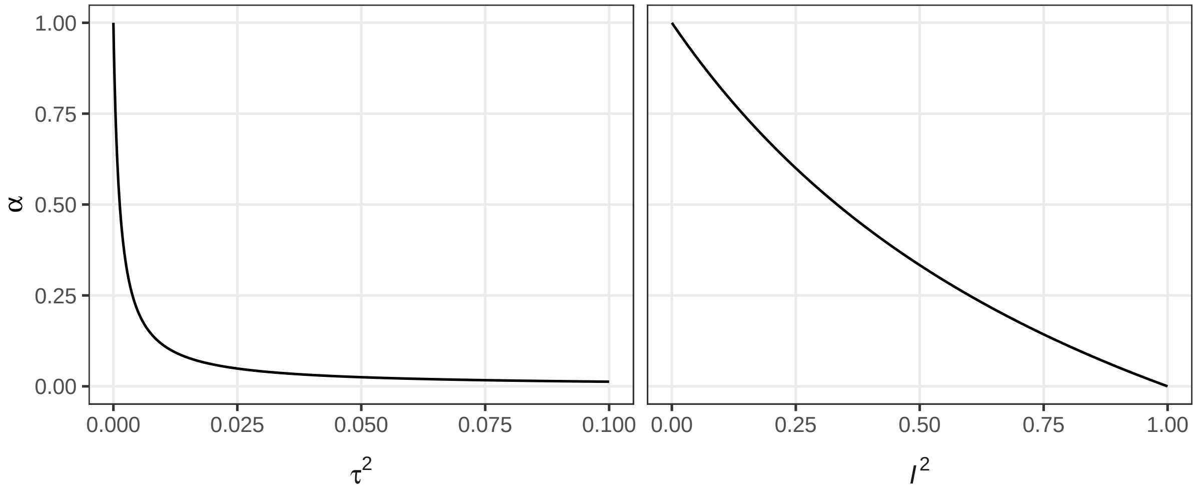

Figure 3: The heterogeneity τ^2 and relative heterogeneity $I^2 = \tau^2/(\tau^2 + \sigma_o^2)$ of a hierarchical model versus the power parameter α from a power prior model which lead to matching posteriors for the effect sizes θ and θ_r. The variance of the original effect estimate $\sigma_o^2 = 0.05^2$ from the "Labels" experiment is used for the transformation to the heterogeneity scale τ^2.

It has remained unclear whether or not a similar correspondence exists in cases where α and τ^2 are random and assigned prior distributions. Here we confirm that there is indeed such a correspondence. Specifically, the marginal posterior of the replication effect size θ_r from the hierarchical model matches with the marginal posterior of the effect size θ from the power prior model if the prior density functions $f_{\tau^2}(\cdot)$ and $f_\alpha(\cdot)$ of τ^2 and α satisfy

$$f_{\tau^2}(\tau^2) = f_\alpha\left(\frac{\sigma_o^2}{2\tau^2 + \sigma_o^2}\right) \frac{2\sigma_o^2}{(2\tau^2 + \sigma_o^2)^2} \tag{20}$$

for every $\tau^2 \geq 0$, see Appendix B for details. Importantly, the correspondence condition (20) involves a scaling by the variance from the original effect estimate σ_o^2, meaning that also in this case α acts similar to a relative heterogeneity parameter. This can also be seen from the correspondence condition between α and $I^2 = \tau^2/(\sigma_o^2 + \tau^2)$, which can be derived in exactly the same way as the correspondence between α and τ^2. That is, the marginal posteriors of θ and θ_r

match if the prior density functions $f_{I^2}(\cdot)$ and $f_\alpha(\cdot)$ of I^2 and α satisfy

$$f_{I^2}(I^2) = f_\alpha\left(\frac{1 - I^2}{1 + I^2}\right) \frac{2}{(1 + I^2)^2} \tag{21}$$

for every $0 \leq I^2 \leq 1$.

Interestingly, conditions (21) and (20) imply that a beta prior on the power parameter $\alpha \sim \text{Be}(x, y)$ corresponds to a generalized F prior on the heterogeneity $\tau^2 \sim \text{GF}(y, x, 2/\sigma_o^2)$ and a generalized beta prior on the relative heterogeneity $I^2 \sim \text{GBe}(y, x, 2)$, see Appendix C for details on both distributions. This connection provides a convenient analytical link between hierarchical modeling and power prior framework, as beta priors for α are almost universally used in applications of power priors. The result also illustrates that the power prior framework seems unnatural from the perspective of hierarchical modeling since it corresponds to specifying priors on the I^2 scale rather than on the τ^2 scale. The same prior on I^2 will imply different degrees of informativeness on the τ^2 scale for original effect estimates $\hat{\theta}_o$ with different variances σ_o^2 since I^2 is entangled with the variance of the original effect estimate.

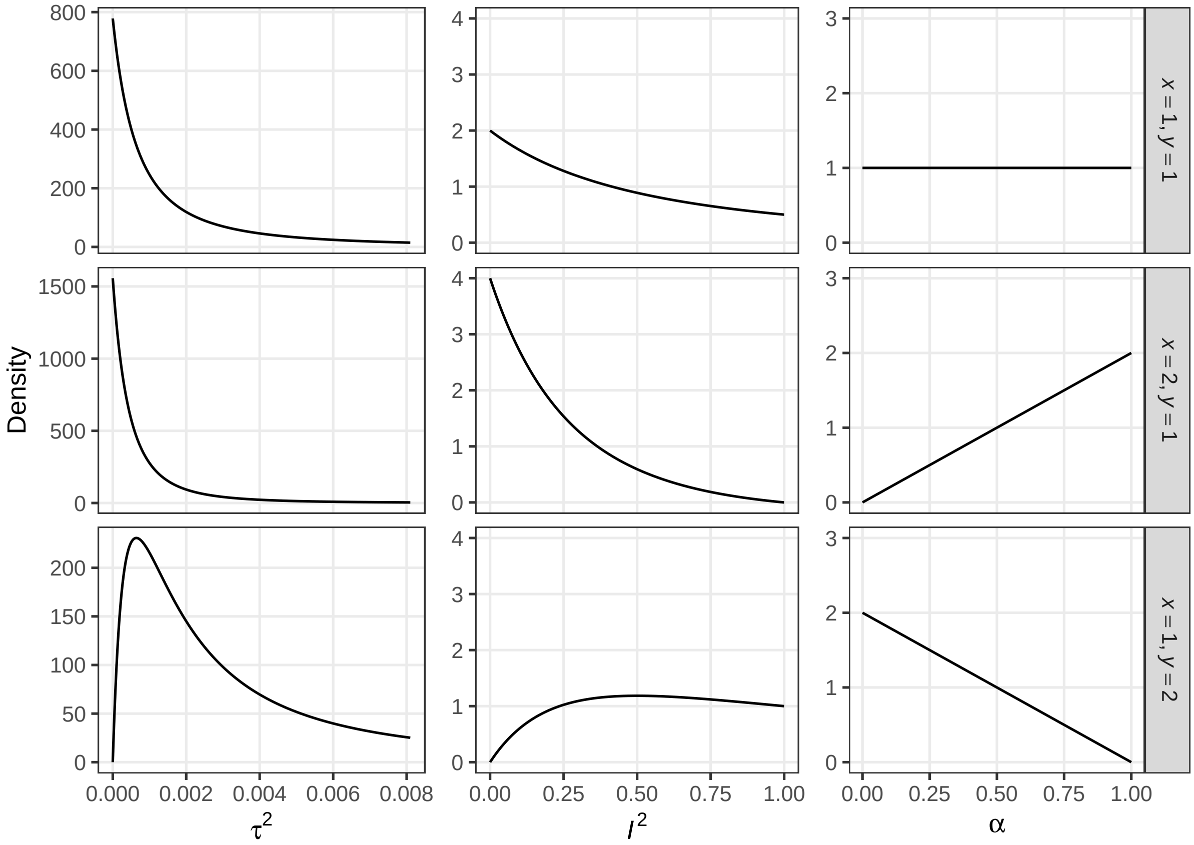

Figure 4: Priors on the heterogeneity $\tau^2 \sim \text{GF}(y, x, 2/\sigma_o^2)$ (left), the relative heterogeneity $I^2 = \tau^2/(\sigma_o^2 + \tau^2) \sim \text{GBe}(y, x, 2)$ (middle) and the power parameter $\alpha \sim \text{Be}(x, y)$ (right) that lead to matching marginal posteriors for effect sizes θ and θ_r. The variance of the original effect estimate $\sigma_o^2 = 0.05^2$ from the “Labels” experiment is used for the transformation to the heterogeneity scale τ^2.

Figure 4 provides three examples of matching priors using the variance of the original effect estimate from the "Labels" experiment for the transformation to the heterogeneity scale τ^2. The top row of Figure 4 shows 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 $\tau^2 = 0$ but still gives some mass to larger values of τ^2. Similarly, on the scale of I^2 the prior slightly favors smaller values. The middle row of Figure 4 shows that the $\alpha \sim \text{Be}(2, 1)$ prior–indicating more compatibility between original and replication than the uniform prior–gives even more mass to small values of τ^2 and I^2, and also has the highest density at $\tau^2 = 0$ and $I^2 = 0$. In contrast, the bottom row of Figure 4 shows that the $\alpha \sim \text{Be}(1, 2)$ prior–indicating less compatibility between original and replication than the uniform prior–gives less mass to small τ^2 and I^2, and has zero density at $\tau^2 = 0$ and $I^2 = 0$.

3.2 Connection to hypothesis testing in hierarchical models

Two types of hypothesis tests can be distinguished in the hierarchical model; tests for the overall effect size θ_* and tests for the heterogeneity variance τ^2. In all cases, computations of marginal likelihoods of the form

$$f(\hat{\theta}_r \mid \mathcal{H}_i) = \int \text{N}(\hat{\theta}_r \mid \theta_*, \sigma_r^2 + \tau^2)\, f(\theta_*, \tau^2 \mid \mathcal{H}_i)\, \mathrm{d}\theta_*\, \mathrm{d}\tau^2 \tag{22}$$

with $i \in \{j, k\}$ are required for obtaining Bayes factors $\text{BF}_{jk}(\hat{\theta}_r) = f(\hat{\theta}_r \mid \mathcal{H}_j)/f(\hat{\theta}_r \mid \mathcal{H}_k)$ which quantify the evidence that the replication data $\hat{\theta}_r$ provide for a hypothesis $\mathcal{H}_k$ over a competing hypothesis $\mathcal{H}_j$. Under each hypothesis a joint prior for τ^2 and θ_* needs to be assigned.

As with parameter estimation, it is of interest to investigate whether there is a correspondence with hypothesis tests from the power prior framework from Section 2.2. For two tests to match, one needs to assign priors to τ^2 and θ_*, respectively, to α and θ so that the marginal likelihood (22) equals the marginal likelihood from the power prior model (6) under both test-relevant hypotheses.

Concerning the generalized replication Bayes factor from (7) testing $\mathcal{H}_0\colon \theta = 0$ versus $\mathcal{H}_1\colon \theta \neq 0$, one can show that it matches with the Bayes factor contrasting $\mathcal{H}_0\colon \theta_* = 0$ versus $\mathcal{H}_1\colon \theta_* \neq 0$ with

$$\begin{aligned} \mathcal{H}_0\colon \theta_* &= 0 & \qquad \text{versus} \qquad & & \theta_* \mid \tau^2, \mathcal{H}_1 &\sim \text{N}(\hat{\theta}_o, \sigma_o^2 + \tau^2) \\ \tau^2 &= 0 & & & \tau^2 \mid \mathcal{H}_1 &\sim \text{GF}(y, x, \sigma_o^2/2) \end{aligned}$$

for the replication data in in the hierarchical framework. The Bayes factor thus compares the likelihood of the replication data under the hypothesis $\mathcal{H}_0$ postulating that the global effect size θ_* is zero and that there is no effect size heterogeneity, relative to the likelihood of the data under the hypothesis $\mathcal{H}_1$ postulating that θ_* follows the posterior based on the original data and an initial flat prior for θ_* along with a generalized F prior on the heterogeneity τ^2. Setting the heterogeneity to $\tau^2 = 0$ under $\mathcal{H}_1$ instead produces the replication Bayes factor under normality from (8).

The Bayes factor (9) that tests $\mathcal{H}_\text{d}\colon \alpha = 0$ to $\mathcal{H}_\text{c}\colon \alpha = 1$ can be obtained in the hierarchical framework by contrasting

$$\begin{aligned} &\mathcal{H}_{\mathrm{d}}\colon \theta_* \sim \mathrm{N}(0, \kappa^2) && \text{versus} && \mathcal{H}_{\mathrm{c}}\colon \theta_* \sim \mathrm{N}(s\,\hat{\theta}_o, s\,\sigma_o^2) \\ &\tau^2 = 0 && && \tau^2 = 0 \end{aligned}$$

with $s = \kappa^2/(\sigma_o^2 + \kappa^2)$. Hence, the Bayes factor compares the likelihood of the replication data under the initial unit-information prior relative to the likelihood of the replication data under the unit-information prior updated by the original data, assuming no heterogeneity under either hypothesis (so that the hierarchical model collapses to a fixed effects model). Although this particular test relates to the power parameter α in the power prior model, it is surprisingly unrelated to testing the heterogeneity variance τ^2 in the hierarchical model.

The Bayes factor (10) testing $\mathcal{H}_{\mathrm{d}}\colon \alpha < 1$ versus $\mathcal{H}_{\mathrm{c}}\colon \alpha = 1$ with $\alpha \mid \mathcal{H}_{\mathrm{d}} \sim \mathrm{Be}(1, y)$ corresponds to testing $\mathcal{H}_{\mathrm{d}}\colon \tau^2 > 0$ versus $\mathcal{H}_{\mathrm{c}}\colon \tau^2 = 0$ with

$$\begin{aligned} &\theta_* \mid \tau^2, \mathcal{H}_{\mathrm{d}} \sim \mathrm{N}(\hat{\theta}_o, \sigma_o^2 + \tau^2) && \text{versus} && \theta_* \mid \tau^2, \mathcal{H}_{\mathrm{c}} \sim \mathrm{N}(\hat{\theta}_o, \sigma_o^2 + \tau^2) \\ &\tau^2 \mid \mathcal{H}_{\mathrm{d}} \sim \mathrm{GF}(y, 1, \sigma_o^2/2) && && \mathcal{H}_{\mathrm{c}}\colon \tau^2 = 0 \end{aligned}$$

The test for compatibility via the power parameter α is thus equivalent to a test for compatibility via the heterogeneity τ^2 (to which a generalized F prior is assigned) after updating of a flat prior for θ_* with the data from the original study.

3.3 Bayes factor asymptotics in the hierarchical model

Like the original test of $\mathcal{H}_{\mathrm{c}}\colon \alpha = 1$ versus $\mathcal{H}_{\mathrm{d}}\colon \alpha < 1$ with $\alpha \mid \mathcal{H}_{\mathrm{d}} \sim \mathrm{Be}(1, y)$, the corresponding test of τ^2 is inconsistent in the sense that when the standard errors from both studies go to zero ($\sigma_o \downarrow 0$ and $\sigma_r \downarrow 0$) and their true effect sizes are equivalent ($\theta_o = \theta_r$), the Bayes factor $\mathrm{BF}_{\mathrm{dc}}$ does not go to zero (to indicate overwhelming evidence for $\mathcal{H}_{\mathrm{c}}\colon \tau^2 = 0$) but converges to a positive constant. It is, however, possible to construct a consistent test for $\mathcal{H}_{\mathrm{c}}\colon \tau^2 = 0$ when we assign a different prior to τ^2 under $\mathcal{H}_{\mathrm{d}}\colon \tau^2 > 0$. For instance, when we assign an inverse gamma prior $\tau^2 \mid \mathcal{H}_{\mathrm{d}} \sim \mathrm{IG}(q, r)$ with shape q and scale r, the Bayes factor is given by

$$\mathrm{BF}_{\mathrm{dc}}(\hat{\theta}_r \mid q, r) = \frac{\int \mathrm{N}(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_r^2 + \sigma_o^2 + 2\tau^2)\,\mathrm{IG}(\tau^2 \mid q, r)\,\mathrm{d}\tau^2}{\mathrm{N}(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_r^2 + \sigma_o^2)}$$

with $\mathrm{IG}(\cdot \mid q, r)$ the density function of the inverse gamma distribution. The limiting Bayes factor is therefore

$$\lim_{\sigma_o, \sigma_r \downarrow 0} \mathrm{BF}_{\mathrm{dc}}(\hat{\theta}_r \mid q, r) = \frac{\Gamma(q + 1/2)\{r + (\theta_r - \theta_o)^2/4\}^{-(q+1/2)}}{\delta(\theta_r - \theta_o)\,\sqrt{4\pi}},$$

so it correctly goes to zero/infinity when the effect sizes θ_r and θ_o are equivalent/different. To understand why the test with $\tau^2 \mid \mathcal{H}_{\mathrm{d}} \sim \mathrm{IG}(q, r)$ is consistent, but the original test with $\alpha \mid \mathcal{H}_{\mathrm{d}} \sim \mathrm{Be}(1, y)$ is not, one can transform the consistent test on τ^2 to the corresponding test on α. The inverse gamma prior for τ^2 implies a prior for α with density

$$f(\alpha \,|\, q, r) = \frac{r^q}{\Gamma(q)} \, \frac{\alpha^{q-1}}{(1-\alpha)^{q+1}} \left(\frac{2}{\sigma_o^2}\right)^q \exp\left\{-\frac{2\,r\,\alpha}{\sigma_o^2(1-\alpha)}\right\}. \tag{23}$$

The Bayes factor contrasting $\mathcal{H}_{\mathrm{c}}\colon \alpha = 1$ versus $\mathcal{H}_{\mathrm{d}}\colon \alpha < 1$ with prior (23) assigned to α under $\mathcal{H}_{\mathrm{d}}$ will thus produce a consistent test. Importantly, the prior (23) depends on the variance of the original effect estimate σ_o^2, so that original studies with different variances will result in different priors on α, even when the parameters s and t from the prior stay the same. The prior thus "unscales" α from the original variance σ_o^2, thereby leading to a consistent test for study compatibility and resolving the undesirable property of the beta prior.

4 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 showed how the power prior approach is connected to hierarchical modeling, and gave conditions under which posterior distributions and hypothesis tests correspond between normal power prior models and normal hierarchical models. This connection provides an intuition for why even with highly precise and compatible original and replication study one can hardly draw conclusive inferences about the power parameter α; the power parameter α has a direct correspondence to the relative heterogeneity variance I^2, and an indirect correspondence to the heterogeneity variance τ^2 in a hierarchical model. Making inferences about a heterogeneity variance from two studies alone seems like a virtually impossible task since the "unit of information" is the number of studies and not the number of samples within a study. Moreover, Bayes factor hypothesis tests related to α have the undesirable asymptotic property of inconsistency if a beta prior is assigned to α. This is because the prior scales with the variance of the original data, just as a beta prior for I^2 would in a hierarchical model.

Which of the two approaches should data analysts use in practice? We believe that the choice should be primarily guided by whether the hierarchical or the power prior model is *scientifically* more suitable for the studies at hand. If data analysts deem it scientifically plausible that the studies' underlying effect sizes are connected via an overarching distribution then the hierarchical model may be more suitable, particularly because the approach naturally generalizes to more than two studies. On the other hand, if data analysts simply want to downweight the original studies' contribution depending on the observed conflict, the power prior approach might be more suitable. The identified limitations for inferences related to the power parameter α should, however, be kept in mind when beta priors are assigned to the power parameter α.

There are also situations where the hierarchical and power prior frameworks can be combined, for example, when multiple replications of a single original study are conducted (multisite replications). In that case, one may model the replication effect estimates in a hierarchical fashion but link their overall effect size to the original study via a power prior. Multisite replications are thus the opposite of the usual situation in clinical trials where several historical "original" studies but only one current "replication" study is available (Gravestock and Held, 2019).

Another commonly used Bayesian approach for incorporating historical data are *robust mixture priors*, i. e., priors which are mixtures of the posterior based on the historical data and an uninformative prior distribution (Schmidli et al., 2014). We conjecture that inferences based on robust mixture priors can be reverse-engineered within the framework of power priors through Bayesian model averaging over two hypotheses about the power parameter; however, more research is needed to explore the relationship between the two approaches.

The proposed methods rely on the standard meta-analytic assumption of approximate normality of effect estimates. This assumption might be inadequate in some situations, for example, when studies have small sample sizes. In this case, the methods could be modified to use the exact likelihood of the data (e. g., binomial or t). However, using the exact likelihood would require numerical methods for the evaluation of integrals which can be evaluated analytically under normality.

Software and data

The CC-By Attribution 4.0 International licensed data were downloaded from `https://osf.io/42ef9/`. All analyses were conducted in the R programming language version 4.2.2 (R Core Team, 2020). The code and data to reproduce this manuscript is available at `https://github.com/SamCH93/ppReplication`. A snapshot of the GitHub repository at the time of writing this article is archived at `https://doi.org/10.5281/zenodo.6940238`. We also provide an R package for estimation and testing under the power prior framework with documentation and example code at `https://github.com/SamCH93/ppRep`.

Acknowledgments

We thank Protzko et al. (2020) for publicly sharing their data. We thank Małgorzata Roos for helpful comments on a draft of the manuscript. This work was supported in part by an NWO Vici grant (016.Vici.170.083) to EJW, an Advanced ERC grant (743086 UNIFY) to EJW, and a Swiss National Science Foundation mobility grant (189295) to LH and SP.

Conflict of interest

The authors have no conflicts of interest to declare.

Appendix A Posterior distribution under the hierarchical model

Under the hierarchical model from (15), the joint posterior conditional on a heterogeneity τ^2 is given by

$$f(\theta_r, \theta_o, \theta_* \,|\, \hat{\theta}_o, \hat{\theta}_r, \tau^2) = \frac{\prod_{i \in \{o,r\}} \mathrm{N}(\hat{\theta}_i \,|\, \theta_i, \sigma_i^2)\, \mathrm{N}(\theta_i \,|\, \theta_*, \tau^2)\, k}{f(\hat{\theta}_o, \hat{\theta}_r \,|\, \tau^2)} \tag{24}$$

with normalizing constant

$$\begin{aligned} f(\hat{\theta}_o, \hat{\theta}_r \,|\, \tau^2) &= \int \prod_{i \in \{o,r\}} \mathrm{N}(\hat{\theta}_i \,|\, \theta_i, \sigma_i^2)\ \mathrm{N}(\theta_i \,|\, \theta_*, \tau^2)\, k \,\mathrm{d}\theta_o \,\mathrm{d}\theta_r \,\mathrm{d}\theta_* \\ &= \int \prod_{i \in \{o,r\}} \mathrm{N}(\hat{\theta}_i \,|\, \theta_*, \sigma_i^2 + \tau^2) k \,\mathrm{d}\theta_* \\ &= k\ \mathrm{N}(\hat{\theta}_r \,|\, \hat{\theta}_o, \sigma_o^2 + \sigma_r^2 + 2\tau^2). \end{aligned} \tag{25}$$

To obtain the marginal posterior distribution of the replication effect size θ_r we need to integrate out θ_o and θ_* from (24). This leads to

$$\begin{aligned} f(\theta_r \,|\, \hat{\theta}_o, \hat{\theta}_r, \tau^2) &= \frac{\int \prod_{i \in \{o,r\}} \mathrm{N}(\hat{\theta}_i \,|\, \theta_i, \sigma_i^2)\ \mathrm{N}(\theta_i \,|\, \theta_*, \tau^2)\, k \,\mathrm{d}\theta_o \,\mathrm{d}\theta_*}{f(\hat{\theta}_o, \hat{\theta}_r \,|\, \tau^2)} \\ &= \frac{\mathrm{N}(\hat{\theta}_r \,|\, \theta_r, \sigma_r^2) \int \mathrm{N}(\theta_r \,|\, \theta_*, \tau^2)\ \mathrm{N}(\hat{\theta}_o \,|\, \theta_*, \sigma_o^2 + \tau^2) \,\mathrm{d}\theta_*}{\mathrm{N}(\hat{\theta}_r \,|\, \hat{\theta}_o, \sigma_o^2 + \sigma_r^2 + 2\tau^2)} \\ &= \frac{\mathrm{N}(\hat{\theta}_r \,|\, \theta_r, \sigma_r^2)\ \mathrm{N}(\theta_r \,|\, \hat{\theta}_o, \sigma_o^2 + 2\tau^2)}{\mathrm{N}(\hat{\theta}_r \,|\, \hat{\theta}_o, \sigma_o^2 + \sigma_r^2 + 2\tau^2)} \end{aligned}$$

which can be further simplified to identify the posterior given in (16).

When the heterogeneity τ^2 is also assigned a prior distribution, the posterior distribution can be factorized in the posterior conditional on τ^2 from (24) and the marginal posterior of τ^2

$$f(\tau^2, \theta_r, \theta_o, \theta_* \,|\, \hat{\theta}_o, \hat{\theta}_r) = f(\theta_r, \theta_o, \theta_* \,|\, \hat{\theta}_o, \hat{\theta}_r, \tau^2)\, f(\tau^2 \,|\, \hat{\theta}_o, \hat{\theta}_r).$$

Integrating out θ_r, θ_o, and θ_* from the joint posterior and using the previous results (25), the marginal posterior of τ^2 can be derived to be

$$\begin{aligned} f(\tau^2 \,|\, \hat{\theta}_o, \hat{\theta}_r) &= \frac{\int \prod_{i \in \{o,r\}} \mathrm{N}(\hat{\theta}_i \,|\, \theta_i, \sigma_i^2)\ \mathrm{N}(\theta_i \,|\, \theta_*, \tau^2)\, k\, f(\tau^2) \,\mathrm{d}\theta_o \,\mathrm{d}\theta_r \,\mathrm{d}\theta_*}{f(\hat{\theta}_o, \hat{\theta}_r)} \\ &= \frac{f(\hat{\theta}_r, \hat{\theta}_o \,|\, \tau^2)\, f(\tau^2)}{\int f(\hat{\theta}_r, \hat{\theta}_o \,|\, \tau^2)\, f(\tau^2) \,\mathrm{d}\tau^2} \\ &= \frac{\mathrm{N}(\hat{\theta}_r \,|\, \hat{\theta}_o, \sigma_o^2 + \sigma_r^2 + 2\tau^2)\, f(\tau^2)}{\int \mathrm{N}(\hat{\theta}_r \,|\, \hat{\theta}_o, \sigma_o^2 + \sigma_r^2 + 2\tau^2)\, f(\tau^2) \,\mathrm{d}\tau^2}. \end{aligned}$$

Appendix B Conditions for matching posteriors

For the marginal posteriors of θ_r and θ to match it must hold for every $\theta = \theta_r$ that

$$\begin{aligned} f(\theta_r \,|\, \hat{\theta}_o, \hat{\theta}_r) &= f(\theta \,|\, \hat{\theta}_o, \hat{\theta}_r) \\ \int_0^\infty f(\theta_r \,|\, \hat{\theta}_o, \hat{\theta}_r, \tau^2)\, f(\tau^2 \,|\, \hat{\theta}_o, \hat{\theta}_r) \,\mathrm{d}\tau^2 &= \int_0^1 f(\theta \,|\, \hat{\theta}_o, \hat{\theta}_r, \alpha)\, f(\alpha \,|\, \hat{\theta}_o, \hat{\theta}_r) \,\mathrm{d}\alpha. \end{aligned} \tag{26}$$

By applying a change of variables (18) or (19) to the left or right hand side of (26), the marginal posteriors conditional on τ^2 and α match. It is now left to investigate whether there are priors for τ^2 and α so that also the marginal posteriors of τ^2 and α match. The marginal posterior distribution of α is proportional to

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

After a change of variables $\tau^2 = (1/\alpha - 1)\,(\sigma_o^2/2)$ the marginal posterior becomes

$$f(\tau^2 \,|\, \hat{\theta}_o, \hat{\theta}_r) \propto f_\alpha\left(\frac{\sigma_o^2}{2\tau^2 + \sigma_o^2}\right) \frac{2\sigma_o^2}{(2\tau^2 + \sigma_o^2)^2} \; \mathrm{N}(\hat{\theta}_r \,|\, \hat{\theta}_o, \sigma_r^2 + \sigma_o^2 + 2\tau^2),$$

Since, as shown in Appendix A, the marginal posterior of τ^2 under the hierarchical model is proportional to

$$f(\tau^2 \,|\, \hat{\theta}_o, \hat{\theta}_r) \propto f_{\tau^2}(\tau^2) \; \mathrm{N}(\hat{\theta}_r \,|\, \hat{\theta}_o, \sigma_r^2 + \sigma_o^2 + 2\tau^2),$$

the marginal posteriors of the effect sizes θ and θ_r match if

$$f_{\tau^2}(\tau^2) = f_\alpha\left(\frac{\sigma_o^2}{2\tau^2 + \sigma_o^2}\right) \frac{2\sigma_o^2}{(2\tau^2 + \sigma_o^2)^2}$$

holds for every $\tau^2 \geq 0$.

Appendix C The generalized beta and F distributions

A random variable $X \sim \mathrm{GBe}(a, b, \lambda)$ with density function

$$f(x \,|\, a, b, \lambda) = \frac{\lambda^a \, x^{a-1} \, (1 - x)^{b-1}}{\mathrm{B}(a, b) \, \{1 - (1 - \lambda)x\}^{a+b}} \, \mathbf{1}_{[0,1]}(x) \tag{27}$$

follows a generalized beta distribution (in the parametrization of Libby and Novick, 1982) with $\mathbf{1}_S(x)$ denoting the indicator function that x is in the set S. A random variable $X \sim \mathrm{GF}(a, b, \lambda)$ with density function

$$f(x \,|\, a, b, \lambda) = \frac{\lambda^a \, x^{a-1}}{\mathrm{B}(a, b) \, (1 + \lambda x)^{a+b}} \, \mathbf{1}_{[0,\infty)}(x) \tag{28}$$

follows a generalized F distribution (in the parametrization of Pham-Gia and Duong, 1989).

References

Abramowitz, M. and Stegun, I. A., editors (1965). *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. Dover Publications, Inc., New York.

Bayarri, M. and Mayoral, A. (2002a). Bayesian analysis and design for comparison of effect-sizes. *Journal of Statistical Planning and Inference*, 103(1-2):225–243. doi:10.1016/s0378-3758(01)00223-3.

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, 40(3):1550–1577. doi:10.1214/12-aos1013.

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

Best, N., Price, R. G., Pouliquen, I. J., and Keene, O. N. (2021). Assessing efficacy in important subgroups in confirmatory trials: An example using Bayesian dynamic borrowing. *Pharmaceutical statistics*, 20(3):551–562. doi:10.1002/pst.2093.

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.

Duan, Y., Ye, K., and Smith, E. P. (2005). Evaluating water quality using power priors to incorporate historical information. *Environmetrics*, 17(1):95–106. doi:10.1002/env.752.

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.

Good, I. J. (1958). Significance tests in parallel and in series. *Journal of the American Statistical Association*, 53(284):799–813.

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.

Gravestock, I. and Held, L. (2019). Power priors based on multiple historical studies for binary outcomes. *Biometrical Journal*, 61(5):1201–1218. doi:10.1002/bimj.201700246.

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. *The Annals of Applied Statistics*, 16(2). doi:10.1214/21-AOAS1502.

Held, L. and Sauter, R. (2017). Adaptive prior weighting in generalized regression. *Biometrics*, 73(1):242–251. doi:10.1111/biom.12541.

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.

Libby, D. L. and Novick, M. R. (1982). Multivariate generalized beta distributions with applications to utility assessment. *Journal of Educational Statistics*, 7(4):271–294. doi:10.3102/10769986007004271.

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., Aust, F., Held, L., and Wagenmakers, E.-J. (2022). Normalized power priors always discount historical data. doi:10.48550/ARXIV.2206.04379. Preprint.

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.

Pham-Gia, T. and Duong, Q. (1989). The generalized beta- and F-distributions in statistical modelling. *Mathematical and Computer Modelling*, 12(12):1613–1625. doi:10.1016/0895-7177(89)90337-3.

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

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

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.

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D., and Neuenschwander, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032. doi:10.1111/biom.12242.

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.

Computational details

```
cat(paste(Sys.time(), Sys.timezone(), "\n"))

## 2023-01-30 08:31:47 Europe/Zurich

sessionInfo()

## R version 4.2.2 Patched (2022-11-10 r83330)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.5 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.9.0
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.9.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=de_CH.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=de_CH.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=de_CH.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=de_CH.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] gridExtra_2.3          ReplicationSuccess_1.2 hypergeo_1.2-13
## [4] dplyr_1.0.10           xtable_1.8-4           colorspace_2.0-3
## [7] ggplot2_3.4.0          ppRep_0.42             knitr_1.41
##
## loaded via a namespace (and not attached):
##  [1] ggpubr_0.5.0     pillar_1.8.1     compiler_4.2.2   highr_0.10
##  [5] tools_4.2.2      evaluate_0.20    lifecycle_1.0.3  tibble_3.1.8
##  [9] gtable_0.3.1     pkgconfig_2.0.3  rlang_1.0.6      DBI_1.1.3
## [13] cli_3.6.0        xfun_0.36        withr_2.5.0      stringr_1.5.0
## [17] generics_0.1.3   vctrs_0.5.1      contfrac_1.1-12  elliptic_1.4-0
## [21] cowplot_1.1.1    isoband_0.2.7    grid_4.2.2       tidyselect_1.2.0
## [25] glue_1.6.2       deSolve_1.34     R6_2.5.1         rstatix_0.7.1
## [29] fansi_1.0.3      carData_3.0-5    car_3.1-1        tidyr_1.2.1
## [33] purrr_1.0.1      farver_2.1.1     magrittr_2.0.3   backports_1.4.1
## [37] scales_1.2.1     MASS_7.3-58.1    abind_1.4-5      assertthat_0.2.1
## [41] ggsignif_0.6.4   labeling_0.4.2   utf8_1.2.2       stringi_1.7.12
## [45] munsell_0.5.0    broom_1.0.2
```