Power Priors for Replication Studies

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

May 2, 2022

Abstract

Power priors are used to incorporate historical data in Bayesian data analysis by taking the likelihood of the historical data raised to the power of α as the prior distribution. Here we propose a power prior modeling approach for the analysis of replication studies. The power parameter α quantifies the similarity between the original study and a replication attempt. We show how power priors help effect size parameter estimation and Bayes factor hypothesis testing by dynamic borrowing of information from the original study. Power priors also enable inferences about study compatibility through estimates and tests for the power parameter α. We give new asympotic results on power prior inferences, showing that a complete discounting of the original data ($\alpha = 0$) is possible, whereas a complete pooling of original and replication ($\alpha = 1$) can never be achieved, even when the replication perfectly mirrors the original study and the sample sizes from both studies become arbitrarily large. We also generalize the known connection between power priors and hierarchical models for fixed parameters to the situation in which the power parameter α and the between-study heterogeneity τ^2 are assigned a prior distribution. Our result implies that the commonly assigned beta prior on α corresponds to a generalized F prior on τ^2 which scales with the variance of the original data. This 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, leading to undesirable finite sample and asymptotic properties due to the scaling of the prior with the variance from the data.

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

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, 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). 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 show how the power prior can be used for parameter estimation (Section 3) and Bayes factor hypothesis testing (Section 4). In both cases, the connection to an alternative approach for incorporating historical data—hierarchical modeling—is explored. We give explicit conditions under which posterior distributions and tests can be reverse-engineered from one framework to the other. Moreover, we study posterior distributions and tests related to the power parameter α from an asymptotic point of view. This perspective shows an inherent asymmetry in inferences related to α: by increasing the sample size of both the original and the replication study, one can obtain arbitrarily peaked posteriors for α at $\alpha = 0$ when original and replication study are incompatible, whereas for perfectly compatible studies the posterior of α will hardly change from the prior. The implications of our results are then discussed in Section 5. 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).

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" 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

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

as 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

$$\hat{\alpha}_{\mathrm{EB}} = \begin{cases} 1 & \text{if } (\hat{\theta}_r - \hat{\theta}_o)^2 - \sigma_r^2 \leq 0 \\ \min\left[1, \sigma_o^2 \,/\, \left\{(\hat{\theta}_r - \hat{\theta}_o)^2 - \sigma_r^2\right\}\right] & \text{else.} \end{cases}$$

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 (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. 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.

3 Parameter estimation

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

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

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

$$f(\hat{\theta}_r \,|\, \hat{\theta}_o, x, y) = \int_0^1 \int_{-\infty}^{\infty} \mathrm{N}(\hat{\theta}_r \,|\, \theta, \sigma_r^2) \times \mathrm{N}(\theta \,|\, \hat{\theta}_o, \sigma_o^2/\alpha) \times \mathrm{Be}(\alpha \,|\, x, y)\, \mathrm{d}\theta\, \mathrm{d}\alpha \tag{3}$$

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

is generally not available in closed form but requires numerical integration with respect to the prior distribution of α. However, in Section 3.2 we show some situations where closed form solu-

tions exist. Finally, 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 (2). In the case of the power parameter α, this leads to

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

whereas for effect size θ we have

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

3.1 Example "Labels"

We 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 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 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. 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 (red) 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. This 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 suggests that it is easier to obtain conclusive posterior inferences about α for conflicting replication studies but more difficult for compatible replications, even with the high sample sizes employed in these three replications.

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,

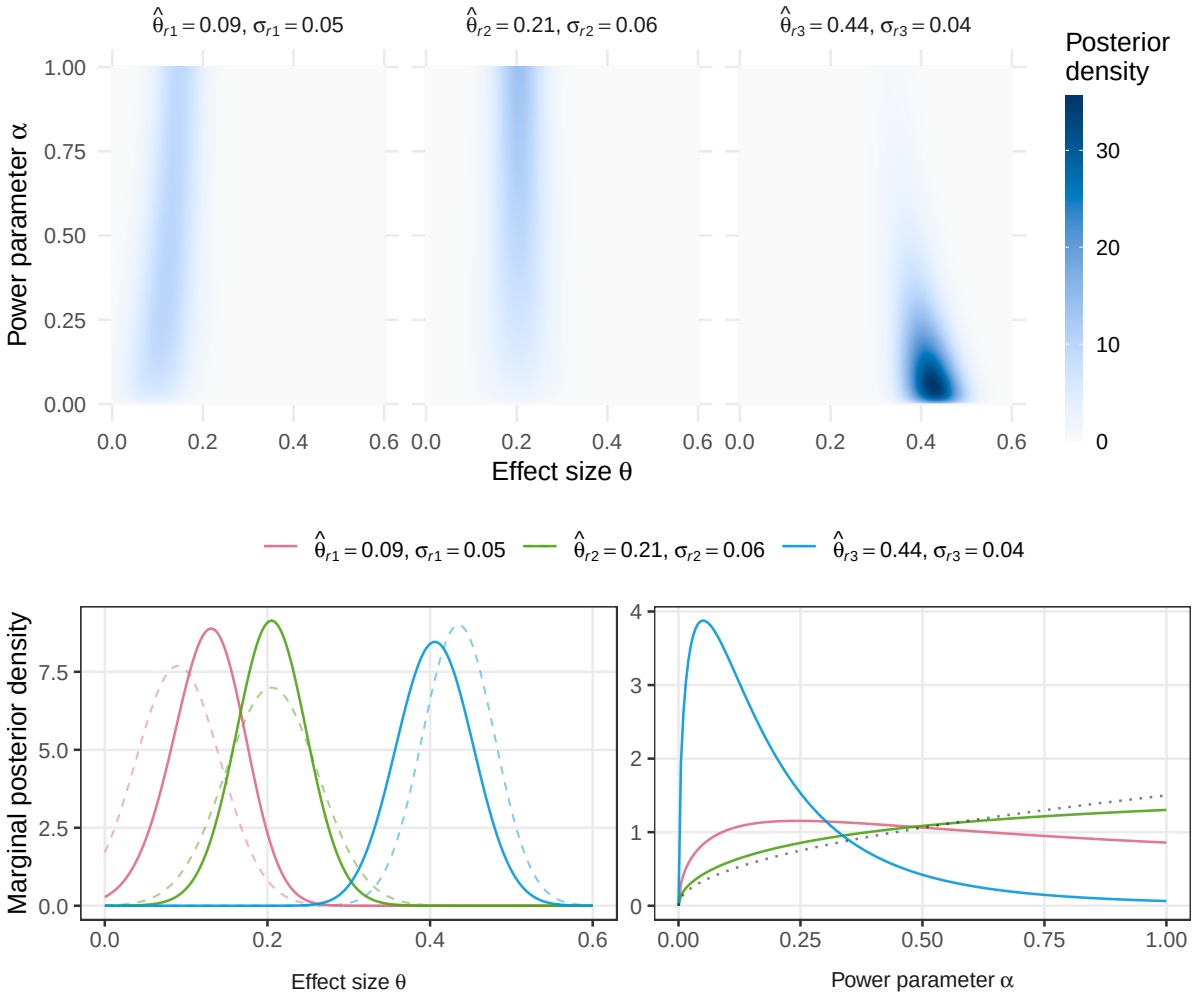

Figure 1: Bayesian analysis of three replication studies from the replication project by Protzko et al. (2020). Shown are joint (top) and marginal (bottom) posterior distributions of effect size θ and power parameter α. A power prior for the effect size θ is constructed from the original effect estimate $\hat{\theta}_o = 0.21$ (with standard error $\sigma_o = 0.05$) and an initial flat prior $f(\theta) \propto 1$.The power parameter α is assigned a uniform $\alpha \sim \text{Be}(1, 1)$marginal 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 through a power prior. The dotted line represents the limiting posterior density of the power parameter α for perfectly agreeing original and replication studies.

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.

3.2 Power parameter asymptotics

It is counterintuitive that for studies with exactly equivalent effect estimates the marginal posterior of α barely changes compared to the prior. A possible explanation could be that the studies had a too small sample size for the posterior to become sufficiently peaked. In the following, we thus investigate this phenomenon from an asymptotic point of view, looking at the situation where the standard errors become arbitrarily small which reflects an arbitrarily large increase of the sample size.

When original and replication effect estimate are equivalent ($\hat{\theta}_r = \hat{\theta}_o$), several terms cancel in the marginal posterior (5) such that it simplifies to

$$f(\alpha \,|\, \hat{\theta}_o = \hat{\theta}_r, x, y) = \frac{(1/c + 1/\alpha)^{-1/2} \times \text{Be}(\alpha \,|\, x, y)}{\int_0^1 (1/c + 1/\alpha')^{-1/2} \times \text{Be}(\alpha' \,|\, x, y)\, \text{d}\alpha'} \tag{6}$$

where $c = \sigma_o^2/\sigma_r^2$ is the relative variance. Importantly, the marginal posterior (6) does not depend on the actual value of the standard errors σ_o and σ_r but only the variance ratio c. This means that (6) holds for finite standard errors but also in the idealized mathematical situation where both standard errors go equally fast to zero (i. e., infinite sample size), but with possibly different starting values ($c \neq 1$). Furthermore, the integral in the denominator of (6) can be represented in terms of the hypergeometric function ${}_2F_1(a, b, c; z) = \{\int_0^1 t^{b-1}(1-t)^{c-b-1}(1 - tz)^{-a}\, \text{d}t\}/\text{B}(b, c-b)$ with $\text{B}(x, y)$ the beta function (Abramowitz and Stegun, 1965, chapter 15). Using this representation, the marginal posterior is given by

$$f(\alpha \,|\, \hat{\theta}_o = \hat{\theta}_r, x, y) = \frac{(1/c + 1/\alpha)^{-1/2}\, \alpha^{x-1}\, (1-\alpha)^{y-1}}{{}_2F_1(1/2, x + 1/2, y + x + 1/2; -1/c)\, \text{B}(x + 1/2, y)}. \tag{7}$$

Typically, the original data are predetermined and only the standard error of the replication study can be changed. It is therefore interesting to study the behavior of (7) for $c \to \infty$, i. e., the replication standard error σ_r goes to zero while the original standard error σ_o remains fixed, reflecting an arbitrary increase of the replication sample size. In that case it is straightforward to see from the power series representation of the hypergeometric function that

$$\lim_{c \to \infty} {}_2F_1(1/2, x + 1/2, y + x + 1/2; -1/c) = \lim_{c \to \infty} 1 + \mathcal{O}(1/c) = 1.$$

Hence, the limiting posterior density is

$$\lim_{c \to \infty} f(\alpha \,|\, \hat{\theta}_o = \hat{\theta}_r, x, y) = \text{Be}(\alpha \,|\, x + 1/2, y), \tag{8}$$

that is, again a beta density but with success parameter $x + 1/2$, so just slightly more mass for larger values of α compared to the prior. In case of the "Labels" experiment, the marginal posterior density from the replication with perfectly agreeing effect estimate (the green line in Figure 1) is close to the limiting $\text{Be}(1 + 1/2, 1)$ density (dotted line).

It is also possible to derive the limiting marginal posterior distribution of α when the effect estimates are not the same and the standard error of the replication estimate σ_r goes to zero while the original standard error σ_o remains fixed. In this case, the integral in the normalizing constant (4) can be represented by the confluent hypergeometric function $M(a, b, z) = \{\int_0^1 \exp(zt) t^{a-1}(1-t)^{b-a-1}\, \text{d}t\}/\text{B}(b-a, a)$ (Abramowitz and Stegun, 1965, chapter 13) so that the marginal posterior is given by

$$\lim_{\sigma_r \downarrow 0} f(\alpha \,|\, \hat{\theta}_o, \hat{\theta}_r, x, y) = \text{Be}(\alpha \,|\, x + 1/2, y) \times \frac{\exp\left\{-\alpha\, (\hat{\theta}_o - \hat{\theta}_r)^2/(2\sigma_o^2)\right\}}{M\{x + 1/2, x + 1/2 + y, -(\hat{\theta}_o - \hat{\theta}_r)^2/(2\sigma_o^2)\}}. \tag{9}$$

The distribution (9) reduces to (8) when the effect estimates are equal ($\hat{\theta}_o = \hat{\theta}_r$) since then the right fraction becomes one, which can be shown using the power series representation of the confluent hypergeometric function. However, when the effect estimates become more different

(larger $|\hat{\theta}_o - \hat{\theta}_r|$) the limiting distribution (9) will be increasingly shifted towards smaller values of α indicating more incompatibility, see Figure 2. Smaller original standard errors σ_o will amplify this shift, meaning that the posterior can become arbitrarily peaked by increasing the sample size of the original study. In contrast, when the effect estimates ($\hat{\theta}_o = \hat{\theta}_r$) are the same the original standard error σ_o does not influence the posterior.

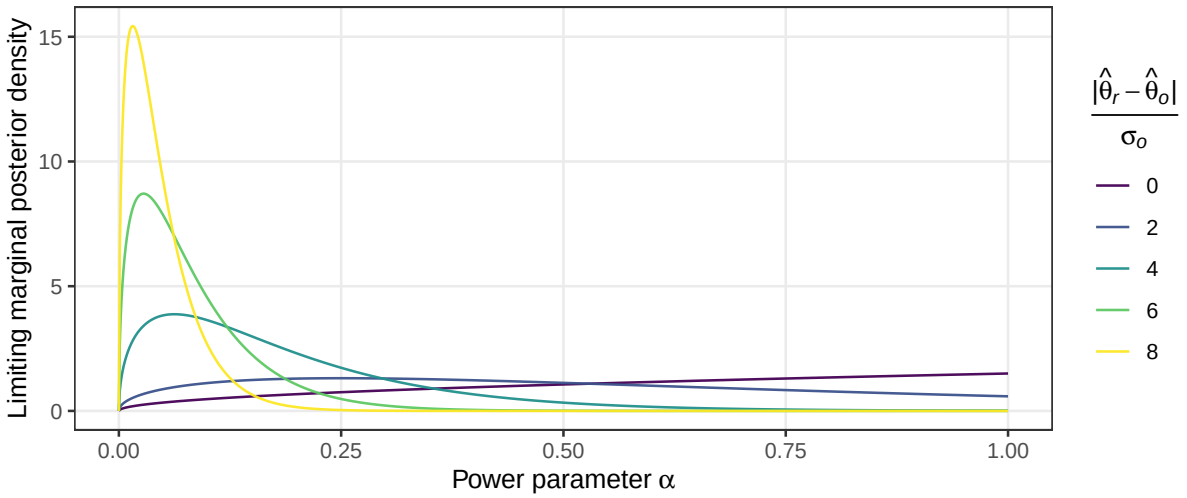

Figure 2: Limiting marginal posterior distribution of power parameter α when the replication standard error goes to zero ($\sigma_r \downarrow 0$) and a $\alpha \sim \text{Be}(1, 1)$ prior is chosen, for different values of the effect difference standardized by the original standard error $|\hat{\theta}_r - \hat{\theta}_o|/\sigma_o$.

These results show that it is possible to obtain arbitrarily peaked posteriors for α when the underlying effect sizes are not equivalent, whereas the posterior hardly changes from the prior when the underlying effect sizes are equivalent, even in the limit of infinitely large sample sizes. This implies that a complete pooling of original and replication ($\alpha = 1$) can never be achieved, whereas a complete discounting ($\alpha = 0$) is possible, assuming the initial prior for the effect size θ is proper. While mathematically unambiguous, these results may appear counterintuitive. To understand the phenomenon better, we will now view it from the perspective of hierarchical modeling.

3.3 Connection to parameter estimation in hierarchical models

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

$$\hat{\theta}_i \,|\, \theta_i \sim \text{N}(\theta_i, \sigma_i^2) \tag{10a}$$

$$\theta_i \,|\, \theta_* \sim \text{N}(\theta_*, \tau^2) \tag{10b}$$

$$f(\theta_*) \propto k \tag{10c}$$

where for study $i \in \{o, r\}$ the effect estimates $\hat{\theta}_i$ are normally distributed around study-specific effect sizes θ_i which themselves are 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 the

default approach in hierarchical modeling of effect estimates (Röver et al., 2021).

As shown in Appendix A, the marginal posterior distribution of the replication effect size θ_r is given by

$$\theta_r \,|\, \hat{\theta}_o, \hat{\theta}_r, \tau^2 \sim \mathrm{N}\left(\frac{\hat{\theta}_r/\sigma_r^2 + \hat{\theta}_o/(2\tau^2 + \sigma_o^2)}{1/\sigma_r^2 + 1/(2\tau^2 + \sigma_o^2)}, \frac{1}{1/\sigma_r^2 + 1/(2\tau^2 + \sigma_o^2)}\right). \tag{11}$$

There exists a correspondence between the hierarchical and 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 from (2) simplifies to a normal

$$\theta \,|\, \hat{\theta}_o, \hat{\theta}_r, \alpha \sim \mathrm{N}\left(\frac{\hat{\theta}_r/\sigma_r^2 + (\hat{\theta}_o\alpha)/\sigma_o^2}{1/\sigma_r^2 + \alpha/\sigma_o^2}, \frac{1}{1/\sigma_r^2 + \alpha/\sigma_o^2}\right). \tag{12}$$

Theorem 2.2 in Chen and Ibrahim (2006) then establishes that the two posterior distributions (11) and (12) 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 $\tau^2 \to \infty$ 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 mapping from α to 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 that the relationship between α and I is approximately linear. Therefore, a rough and ready heuristic to connect power priors to hierarchical models is $\alpha \approx 1 - I$.

It has remained unclear whether or not a mapping exists in cases where α and τ^2 are random. If it were to exist, it must hold for any $\theta = \theta_r$ that

$$\begin{gathered} f(\theta_r \,|\, \hat{\theta}_o, \hat{\theta}_r) = f(\theta \,|\, \hat{\theta}_o, \hat{\theta}_r) \\ \int_0^\infty f(\theta_r \,|\, \hat{\theta}_o, \hat{\theta}_r, \tau^2)\, f(\tau^2 \,|\, \hat{\theta}_o, \hat{\theta}_r)\, \mathrm{d}\tau^2 = \int_0^1 f(\theta \,|\, \hat{\theta}_o, \hat{\theta}_r, \alpha)\, f(\alpha \,|\, \hat{\theta}_o, \hat{\theta}_r)\, \mathrm{d}\alpha. \end{gathered} \tag{13}$$

By applying the change of variables mentioned above either to the left or right hand side of (13), 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

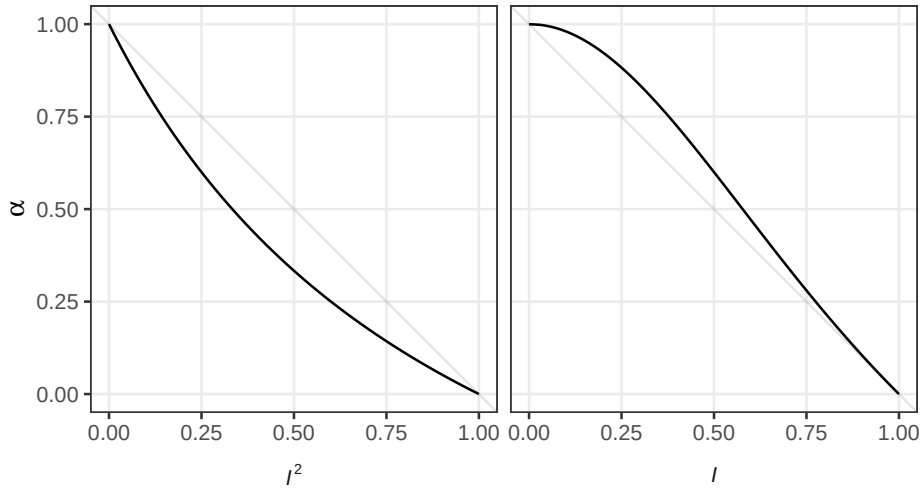

Figure 3: The relative heterogeneity measure $I^2 = \tau^2/(\tau^2 + \sigma_o^2)$, respectively its square root I, of a hierarchical model and the power parameter α from a power prior model which lead to matching posteriors for the effect sizes θ and θ_r.

replacing the beta prior in (5) with an unspecified prior $f(\alpha)$, we observe that the marginal posterior distribution of α is proportional to

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

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

$$f(\tau_*^2 \mid \hat{\theta}_o, \hat{\theta}_r) \propto f\left(\alpha = \frac{\sigma_o^2}{2\tau_*^2 + \sigma_o^2}\right) \frac{2\sigma_o^2}{(2\tau_*^2 + \sigma_o^2)^2} \times \mathrm{N}(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_r^2 + \sigma_o^2 + 2\tau_*^2).$$

Appendix A shows that the marginal posterior of τ^2 is proportional to

$$f(\tau^2 \mid \hat{\theta}_o, \hat{\theta}_r) \propto f(\tau^2) \times \mathrm{N}(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_r^2 + \sigma_o^2 + 2\tau^2). \tag{14}$$

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

$$f(\tau^2) = f\left(\alpha = \frac{\sigma_o^2}{2\tau_*^2 + \sigma_o^2}\right) \frac{2\sigma_o^2}{(2\tau_*^2 + \sigma_o^2)^2}. \tag{15}$$

Condition (15) generalizes the known mapping between power prior and hierarchical models to situations when α and τ^2 are random. Importantly, the mapping involves a scaling by the variance from the original effect estimate σ_o^2, meaning that α acts similar to a relative heterogeneity parameter. This can also be seen from the mapping between α and $I^2 = \tau^2/(\sigma_o^2 + \tau^2)$, which can be derived in exactly the same way as the mapping between α and τ^2. That is, the marginal posterior for the effect size θ and θ_r match if the priors for I^2 and α satisfy

$$f(I^2) = f\left(\alpha = \frac{1 - I^2}{1 + I^2}\right) \frac{2}{(1 + I^2)^2}. \tag{16}$$

Interestingly, conditions (16) and (15) 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 B 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 odd 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 historical data with different variances since I^2 is entangled with the variance of the historical data.

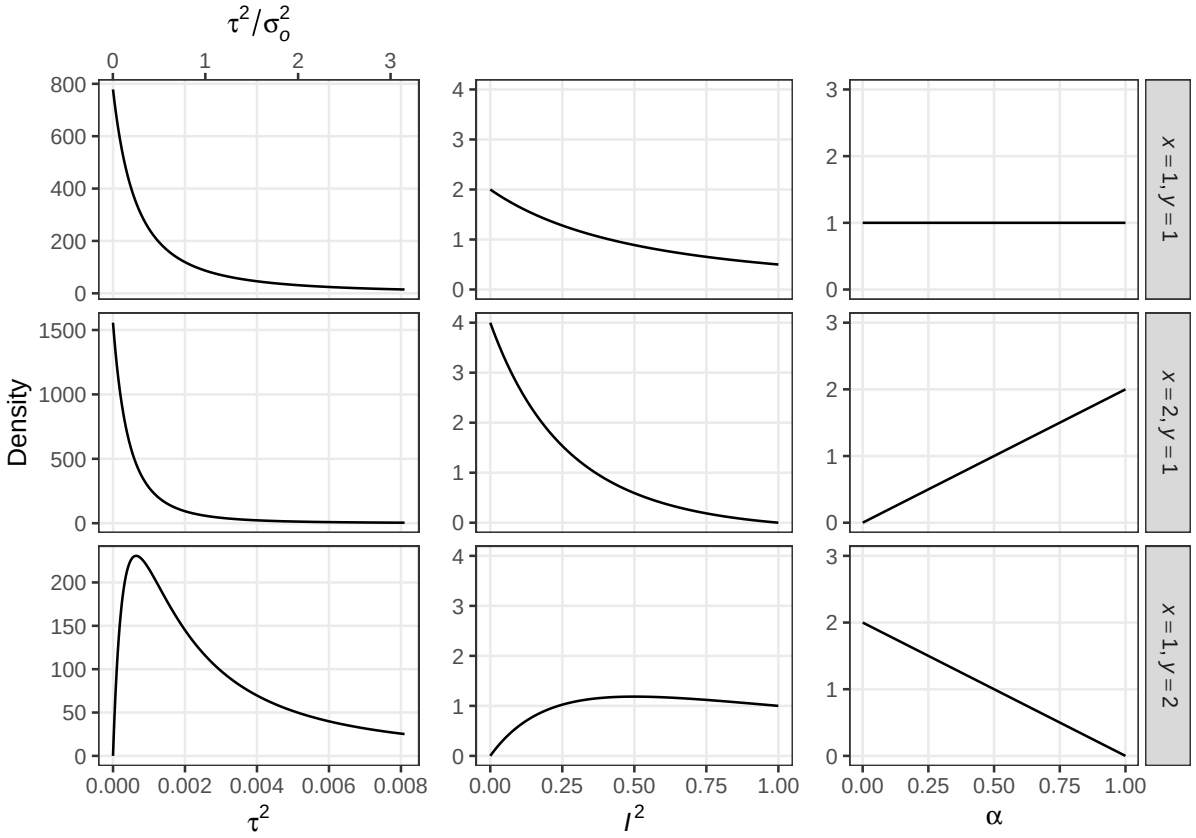

Figure 4: Priors on the heterogeneity $\tau^2 \sim \text{GF}(y, x, 2/\sigma_o^2)$ (left), the relative heterogeneity $I^2 = \tau^2/(\sigma_o^2 + \tau^2) \sim \text{GBe}(y, x, 2)$ (middle) and the power parameter $\alpha \sim \text{Be}(x, y)$ (right) that lead to matching marginal posteriors for effect 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 $I^2 = 0$.

3.4 Heterogeneity variance asymptotics

As for the power parameter α, we may also want to understand the asymptotic behavior of the marginal posterior of the heterogeneity τ^2. Assume that $\hat{\theta}_r$ and $\hat{\theta}_o$ are consistent estimators of their true underlying effect sizes θ_r and θ_o. When the standard errors from both studies go to zero (e. g., because the the sample size goes to infinity), the estimates will converge in probability to their true effect size. The marginal posterior (14) then becomes

$$f(\tau^2 \,|\, \theta_o, \theta_r) \propto f(\tau^2) \times \mathrm{N}(\theta_r \,|\, \theta_o, 2\tau^2). \tag{17}$$

The limiting marginal posterior (17) depends on the prior of τ^2, and many reasonable choices exist (for an overview see Röver et al., 2021). Since one wants to take into account that the heterogeneity may be zero, the prior should have support at $\tau^2 = 0$. This restriction excludes the conjugate inverse gamma prior, for instance. One distribution with positive density at $\tau^2 = 0$ for which further insight can be gained is the exponential distribution. That is, taking $\tau^2 \sim \mathrm{Exp}(\lambda)$ and assuming that both effect sizes are the same ($\theta_o = \theta_r$), the normalizing constant of the limiting posterior density (17) is available in closed form via the gamma function, leading to

$$f(\tau^2 \,|\, \theta_o = \theta_r, \lambda) = \sqrt{\frac{\lambda}{\tau^2\, \pi}} \exp\left(-\lambda \tau^2\right). \tag{18}$$

From (18) several important realizations can be made: Even when the standard errors from both studies go to zero ($\sigma_o \downarrow 0$ and $\sigma_r \downarrow 0$) and the underlying effect sizes are perfectly compatible ($\theta_o = \theta_r$) there is still uncertainty about the heterogeneity τ^2. Specifically, the limiting distribution is not a point mass at $\tau^2 = 0$, just as for the power parameter α. The reason for this is that the unit of information for τ^2 is the number of studies and not the number of samples within a study, so an infinite number of studies is needed to precisely estimate τ^2. At the same time, the density (18) goes to infinity for $\tau^2 \downarrow 0$, meaning that the value $\tau^2 = 0$ receives overwhelming support from the data, a result which will be useful in the next section on hypothesis testing. In contrast, the limiting density (7) of the power parameter α does not go to infinity at $\alpha = 1$ in the same situation but is bounded by the prior (except if the prior already has infinite density at $\alpha = 1$).

4 Hypothesis testing

Apart from the estimation of θ and α, one may also wish to test hypotheses regarding these parameters. The standard Bayesian approach is to quantify the strength of evidence that the data provide for two competing hypotheses by computing the Bayes factor (Jeffreys, 1961; Kass and Raftery, 1995), that is, the ratio of the two marginal likelihoods.

4.1 Hypotheses about the effect size θ

We may wish to quantify 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 from (1). The associated

Bayes factor is then given by

$$\text{BF}_{01}(\hat{\theta}_r \mid x, y) = \frac{f(\hat{\theta}_r \mid \mathcal{H}_0)}{f(\hat{\theta}_r \mid \mathcal{H}_1)} = \frac{\text{N}(\hat{\theta}_r \mid 0, \sigma_r^2)}{\int_0^1 \text{N}(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_r^2 + \sigma_o^2/\alpha)\ \text{Be}(\alpha \mid x, y)\, \text{d}\alpha}. \tag{19}$$

A reasonable choice for the prior of α under $\mathcal{H}_1$ is a uniform $\alpha \sim \text{Be}(1, 1)$ distribution. However, it is worth noting that fixing $\alpha = 1$ leads to

$$\text{BF}_{01}(\hat{\theta}_r \mid \alpha = 1) = \frac{\text{N}(\hat{\theta}_r \mid 0, \sigma_r^2)}{\text{N}(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_o^2 + \sigma_r^2)}, \tag{20}$$

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.

4.2 Hypotheses about the power parameter α

In order to quantify the compatibility between the original and the replication study we may wish to test hypotheses regarding the power parameter α. For example, we may wish to test $\mathcal{H}_\text{c}\colon \alpha = 1$ ("compatible") versus $\mathcal{H}_\text{d}\colon \alpha < 1$ ("different"). One approach is to assign a point prior $\mathcal{H}_\text{d}\colon \alpha = 0$. 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 choose an uninformative but proper initial prior such as the unit-information prior (Kass and Wasserman, 1995)

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

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

$$\text{BF}_\text{dc}(\hat{\theta}_r \mid \kappa^2) = \frac{f(\hat{\theta}_r \mid \mathcal{H}_\text{d})}{f(\hat{\theta}_r \mid \mathcal{H}_\text{c})} = \frac{\text{N}(\hat{\theta}_r \mid 0, \sigma_r^2 + \kappa^2)}{\text{N}(\hat{\theta}_r \mid s\hat{\theta}_o, \sigma_r^2 + s\sigma_o^2)} \tag{21}$$

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 assign priors to α under $\mathcal{H}_\text{d}$ and $\mathcal{H}_\text{c}$. A suitable class of priors are $\mathcal{H}_\text{d}\colon \alpha \sim \text{Be}(1, y)$ and $\mathcal{H}_\text{c}\colon \alpha \sim \text{Be}(x, 1)$ with $x, y > 1$. Two examples with $x = 2$ and $y = 2$ are shown in Figure 4. The $\text{Be}(1, y)$ prior has its highest density at $\alpha = 0$ and is monotonically decreasing; the $\text{Be}(x, 1)$ prior has its highest density at $\alpha = 1$ and is monotonically increasing. However, for small values of x and y there is relatively much overlap between the two priors, so one may want to specify large enough x and y that the two hypotheses can be separated by diagnostic data.

Finally, a compromise between the two mentioned approaches is to contrast the composite hypothesis $\mathcal{H}_\text{d}\colon \alpha \sim \text{Be}(1, y)$ to the simple hypothesis $\mathcal{H}_\text{c}\colon \alpha = 1$, 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

$$\text{BF}_{\text{dc}}(\hat{\theta}_r \mid y) = \frac{\int_0^1 \text{N}(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_r^2 + \sigma_o^2/\alpha)\ \text{Be}(\alpha \mid 1, y)\, \text{d}\alpha}{\text{N}(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_r^2 + \sigma_o^2)}. \tag{22}$$

The parameter y determines how much mass small values of α receive under $\mathcal{H}_\text{d}$. The simple hypothesis $\mathcal{H}_\text{d}\colon \alpha = 0$ 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". The Bayes factors contrasting $\mathcal{H}_0$ to $\mathcal{H}_1$ (column $\text{BF}_{01}(\hat{\theta}_r \mid x = 1, y = 1)$) 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 obtained from setting $\alpha = 1$ (column $\text{BF}_{01}(\hat{\theta}_r \mid \alpha = 1)$).

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 $H_0\colon \theta = 0$ to $H_1\colon \theta \neq 0$ for different priors for α under H_1, and Bayes factors contrasting $H_\text{d}\colon \alpha < 1$ to $H_\text{c}\colon \alpha = 1$.

	$\hat{\theta}_r$	σ_r	$\text{BF}_{01}(\hat{\theta}_r \mid x = 1, y = 1)$	$\text{BF}_{01}(\hat{\theta}_r \mid \alpha = 1)$	$\text{BF}_{\text{dc}}(\hat{\theta}_r \mid \kappa^2 = 2)$	$\text{BF}_{\text{dc}}(\hat{\theta}_r \mid y = 2)$
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

In order to compute the Bayes factor for testing simple $\mathcal{H}_\text{d}$ versus simple $\mathcal{H}_\text{c}$ 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 $\text{Var}(\hat{\theta}_i) = 4/n_i$ 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 $\text{BF}_{\text{dc}}(\hat{\theta}_r \mid \kappa^2 = 2)$ in Table 1 show that the data provide substantial and strong evidence for $\mathcal{H}_\text{c}$ in the first and second replication study, respectively, whereas the data indicate strong evidence for $\mathcal{H}_\text{d}$ in the third replication study. The Bayes factor $\text{BF}_{\text{dc}}(\hat{\theta}_r \mid y = 2)$ in the right-most column contrasts the simple $\mathcal{H}_\text{c}\colon \alpha = 1$ to the composite $\mathcal{H}_\text{d}\colon \alpha \sim \text{Be}(1, 2)$ and indicates absence of evidence for either hypothesis in the first and second replication, but strong evidence for $\mathcal{H}_\text{d}$ in the third replication. Compared to the simple versus simple approach, 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.

4.4 Bayes factor asymptotics

Similarly as for estimation, we may want to investigate the asymptotic behavior of the proposed Bayes factors. For instance, we may want to understand what happens when the standard error of the replication study σ_r becomes arbitrarily small (through an increase in sample size). Assume again that $\hat{\theta}_r$ is a consistent estimator of its true underlying effect size θ_r, so that as the standard error goes to zero, the estimate will converge in probability to the true effect size.

Assuming that the prior for α under $\mathcal{H}_1$ has mass for $\alpha > 0$, the Bayes factors for tests on the effect size θ from (19) and (20) are consistent, meaning that they will increasingly favor the correct hypothesis as the replication data accumulate (Bayarri et al., 2012). In contrast, the Bayes factors (21) and (22) do not grow unboundedly but converge to constants

$$\lim_{\sigma_r \downarrow 0} \text{BF}_{\text{dc}}(\theta_r \mid \kappa^2) = \sqrt{s\sigma_o^2/\kappa^2} \exp\left[-\frac{1}{2}\left\{\frac{\theta_r^2}{\kappa^2} - \frac{(\theta_r - s\hat{\theta}_o)^2}{s\sigma_o^2}\right\}\right] \tag{23}$$

and

$$\lim_{\sigma_r \downarrow 0} \text{BF}_{\text{dc}}(\theta_r \mid y) = \frac{M\{y, y + 3/2, (\theta_r - \hat{\theta}_o)^2/(2\sigma_o^2)\}\text{B}(3/2, y)}{\text{B}(1, y)} \tag{24}$$

with $M(a, b, z)$ the confluent hypergeometric function as used in Section 3.2. 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 (23) 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 $\lim_{\sigma_r \downarrow 0} \text{BF}_{\text{dc}}(\theta_r \mid \kappa^2 = 2) = 1/28$. Similarly, the bound (24) is minimized for $\theta_r = \hat{\theta}_o = 0.21$ since then the confluent hypergeometric function term becomes one, leading to $\lim_{\sigma_r \downarrow 0} \text{BF}_{\text{dc}}(\theta_r \mid y = 2) = \text{B}(3/2, y)/\text{B}(1, y) = 1/1.9$. Even in an infinitely precise replication study, we cannot find more evidence for $\mathcal{H}_{\text{c}}$. These results illustrate that also for tests related to α there are similar issues present as in estimation of α. That is, the maximum evidence for $\mathcal{H}_{\text{c}}$ is pre-determined by the prior, just as the limiting marginal posterior of α was.

4.5 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 (19) testing $\mathcal{H}_0 \colon \theta = 0$ versus $\mathcal{H}_1 \colon \theta \neq 0$ it is straightforward to show that it matches with the Bayes factor contrasting

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

for the replication data in in the hierarchical framework. The Bayes factor thus compares the likelihood of the replication data under the hypothesis $\mathcal{H}_0$ 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 $\mathcal{H}_1$ that θ_* follows the posterior based on the original data and an initial flat prior for θ_*. Setting the heterogeneity $\tau^2 = 0$ under $\mathcal{H}_1$ produces the standard replication Bayes factor

from (20).

The Bayes factor (21) 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

$$\mathcal{H}_\text{d}\colon \theta_* \sim \text{N}(0, \kappa^2) \qquad \text{versus} \qquad \mathcal{H}_\text{c}\colon \theta_* \sim \text{N}(s\,\hat{\theta}_o, s\,\sigma_o^2)$$

with $s = (\kappa^2/\sigma_o^2)/\{1 + (\kappa^2/\sigma_o^2)\}$ and assuming no heterogeneity $\tau^2 = 0$ under either hypothesis. 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.

The Bayes factor (22) testing $\mathcal{H}_\text{d}\colon \alpha \sim \text{Be}(1, y)$ versus $\mathcal{H}_\text{c}\colon \alpha = 1$ corresponds to a comparison between

$$\mathcal{H}_\text{d}\colon \tau^2 \sim \text{GF}(y, 1, \sigma_o^2/2) \qquad \text{versus} \qquad \mathcal{H}_\text{c}\colon \tau^2 = 0$$

and assuming $\theta_* \,|\, \tau^2 \sim \text{N}(\hat{\theta}_o, \sigma_o^2 + \tau^2)$ under both hypothesis in the hierarchical framework. The test for compatibility via the power parameter α is thus equivalent to a test for compatibility via the heterogeneity τ^2 after conditioning on the original data.

Like the original test on α, the equivalent test on τ^2 is inconsistent. However, choosing a different prior for τ^2 under $\mathcal{H}_\text{d}$ can produce a consistent test, for instance, the exponential prior $\mathcal{H}_\text{d}\colon \tau^2 \sim \text{Exp}(\lambda)$ as already considered in Section 3.4. This can be seen from the Savage-Dickey representation of the corresponding Bayes factor (Dickey, 1971). That is, when original and replication effect estimate are equal ($\hat{\theta}_o = \hat{\theta}_r$) , and their standard errors go to zero ($\sigma_o \downarrow 0$ and $\sigma_r \downarrow 0$), the ratio of the prior of τ^2 and its limiting posterior (18) evaluated at $\tau^2 = 0$ also goes to zero, rendering the test consistent.

To understand why the test with $\mathcal{H}_\text{d}\colon \tau^2 \sim \text{Exp}(\lambda)$ consistent, but the original test with $\mathcal{H}_\text{d}\colon \alpha \sim \text{Be}(1, y)$ is not, one can transform the consistent test on τ^2 to an equivalent test on α. The exponential prior for τ^2 implies a prior for α with density

$$f(\alpha \,|\, \lambda) = \frac{\lambda\,\sigma_o^2}{2\,\alpha^2}\,\exp\left\{-\frac{\lambda\,\sigma_o^2}{2}\left(\frac{1}{\alpha} - 1\right)\right\}. \tag{25}$$

Importantly, the prior (25) 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 λ stays 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.

4.6 Design

Now assume that the replication study has not yet been conducted and we wish to plan for a suitable sample size. In the case of the replication Bayes factor under normality (20), 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 (19), numerical integration or simulation is required to compute the probability of replication success as the marginal likelihood is not available in closed form under $\mathcal{H}_1$.

It is also possible to derive the probability of replication success at some level γ analytically for the simple-simple power parameter Bayes factor (21). With some algebra, one can show that $\text{BF}_{\text{dc}} \leq \gamma$ is equivalent to

$$\left(\hat{\theta}_r - \frac{s\hat{\theta}_o(\sigma_r^2 + \kappa^2)}{\kappa^2 - s\sigma_o^2}\right)^2 \leq X \tag{26}$$

for $\kappa^2 > s\sigma_o^2$ and with

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

Denote by m_i and v_i the mean and variance of $\hat{\theta}_r$ under hypothesis $i \in \{\text{d}, \text{c}\}$. The left hand side of (26) then follows a scaled non-central chi-squared distribution under both hypotheses. Hence the probability of replication success is given by

$$\Pr(\text{BF}_{\text{dc}} \leq \gamma \,|\, \mathcal{H}_i) = \Pr\left(\chi^2_{1,\lambda_i} \leq X/v_i\right) \tag{27}$$

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 (27) 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 $\mathcal{H}_\text{d}$ and $\mathcal{H}_\text{c}$. 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 $n_r \approx 4/\sigma_r^2$.

4.7 Example "Labels" (continued)

Figure 5 illustrates Bayesian design analysis based on the power parameter Bayes factor (21). 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 $\mathcal{H}_\text{c}\colon \alpha = 1$ (top) or $\mathcal{H}_\text{d}\colon \alpha = 0$ (bottom) is shown as a function of the relative sample size. In both cases, the probability is computed assuming that either $\mathcal{H}_\text{c}$ (green) or $\mathcal{H}_\text{d}$ (orange) 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 $\mathcal{H}_\text{d}$ is not much affected by the sample size of the replication study; it stays at almost zero under $\mathcal{H}_\text{c}$, while under $\mathcal{H}_\text{d}$ it increases from about 75% to about 90%. In contrast, the top panels show that the probability for finding strong evidence for $\mathcal{H}_\text{c}$ rapidly increases under $\mathcal{H}_\text{c}$ and seems to level off at an asymptote. Under $\mathcal{H}_\text{d}$ 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

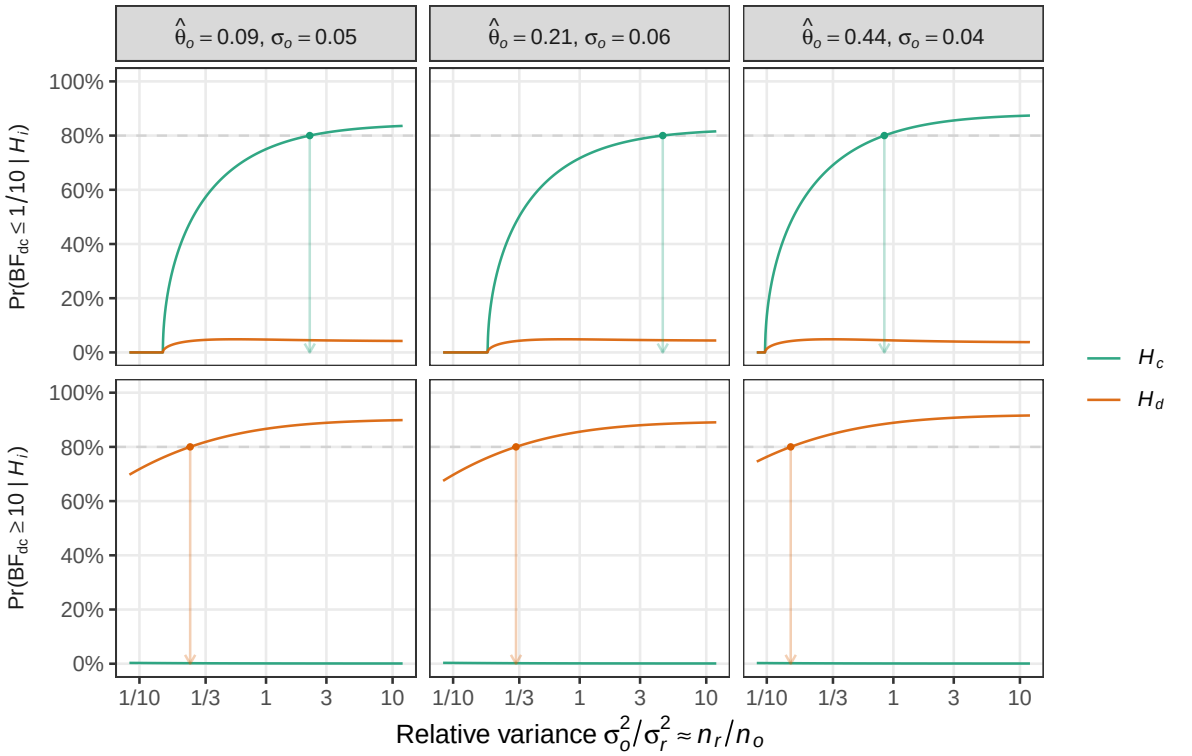

Figure 5: Probability of replication success as a function of relative variance for the three replications of experiment "Labels" regarded as original study. Relative sample size that correspond to a probability of 80% under the respective hypothesis are indicated by arrows.

errors require smaller relative sample sizes in the replication to achieve the same probability of replication success. Under $\mathcal{H}_c$ the required relative sample sizes are larger than under $\mathcal{H}_d$. However, while the probability of misleading evidence under $\mathcal{H}_c$ seems to be well controlled under the determined sample size, under $\mathcal{H}_d$ it stays roughly 5% for all three studies, and even for very large replication sample sizes. Choosing the sample size based on finding strong evidence for $\mathcal{H}_c$ assuming $\mathcal{H}_c$ is true thus guarantees appropriate error probabilities for finding strong evidence for $\mathcal{H}_d$ in all three studies. At the same time, it seems that the probability for finding misleading evidence for $\mathcal{H}_c$ cannot be reduced below around 5% which might undesirably high for certain applications.

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. An asymptotic analysis showed that the posterior of the power parameter α can hardly change from the prior when the outcomes from both studies are perfectly compatible. This means that uninformative priors (e. g., the uniform prior) on α strongly limit the possible degree of borrowing, whereas they do not limit the possible degree of discounting. Data analysts who have strong prior beliefs about the compatibility of both studies may therefore specify more informative priors that give more mass towards larger values of α. A pragmatic alternative is to specify α via an empirical Bayes approach, which permits complete pooling of both studies (Gravestock and Held, 2017). More research is needed to identify practical prior distributions

for α which alleviate the undesirable properties of the standard beta prior.

We also showed how the power prior 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 normal-normal hierarchical models. This connection provides an intuition for why even with two highly precise and compatible studies one cannot draw conclusive posterior inferences about the power parameter α; beta priors on α directly correspond to generalized beta priors on the relative heterogeneity I^2. When no heterogeneity is observed, both of these priors scale with the precision of the study, and their influence on the posterior does therefore not vanish, even when the precision becomes arbitrarily large.

Which of the two approaches should data analysts use in practice? We believe that the power parameter provides, at first sight, a more intuitive scale for assessing study compatibility compared to the heterogeneity variance. However, data analysts should be aware of the identified limitations such as Bayes factor inconsistency. Hierarchical modeling does not suffer from these limitations, and also generalizes better to non-normal likelihoods and prior distributions. In more complex scenarios hierarchical modeling might also be computationally easier to implement and handle. 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 (so-called *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, 2018). 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. Finally, 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 4.2.0 (R Core Team, 2020). The code to reproduce this manuscript as well as an R package for estimation and testing under the power prior framework are available at `https://github.com/SamCH93/ppRep`. A snapshot of the GitHub repository at the time of writing this article is archived at `https://doi.org/10.5281/zenodo.XXXXX`.

Acknowledgments

We thank Protzko et al. (2020) for publicly sharing their data. 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.

Appendix A Posterior distribution under the hierarchical model

Under the hierarchical model from (10), the joint posterior conditional on a heterogeneity τ^2 is given by

$$f(\theta_r, \theta_o, \theta_* \,|\, \hat{\theta}_o, \hat{\theta}_r, \tau^2) = \frac{\prod_{i\in\{o,r\}} \mathrm{N}(\hat{\theta}_i \,|\, \theta_i, \sigma_i^2)\ \mathrm{N}(\theta_i \,|\, \theta_*, \tau^2)\, k}{f(\hat{\theta}_o, \hat{\theta}_r \,|\, \tau^2)} \tag{28}$$

with normalizing constant

$$\begin{aligned} f(\hat{\theta}_o, \hat{\theta}_r \,|\, \tau^2) &= \int \prod_{i\in\{o,r\}} \mathrm{N}(\hat{\theta}_i \,|\, \theta_*, \sigma_i^2)\ \mathrm{N}(\theta_i \,|\, \theta_*, \tau^2)\, k \,\mathrm{d}\theta_o \mathrm{d}\theta_r \mathrm{d}\theta_* \\ &= \int \prod_{i\in\{o,r\}} \mathrm{N}(\hat{\theta}_i \,|\, \theta_*, \sigma_i^2 + \tau^2) k \,\mathrm{d}\theta_* \\ &= k\ \mathrm{N}(\hat{\theta}_r \,|\, \hat{\theta}_o, \sigma_o^2 + \sigma_r^2 + 2\tau^2). \end{aligned} \tag{29}$$

To obtain the marginal posterior distribution of the replication effect size θ_r we need to integrate out θ_o and θ_* from (28). 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 (11).

When the heterogeneity τ^2 is also assigned a prior distribution, the posterior distribution can be factorized in the posterior (28) 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 (29), the

marginal posterior of τ^2 can be derived to be

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

Appendix B The generalized beta and F distributions

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

$$f(x \mid a, b, \lambda) = \frac{\lambda^a\, x^{a-1}\, (1-x)^{b-1}}{\mathrm{B}(a,b)\, \{1 - (1-\lambda)x\}^{a+b}}\, \mathbf{1}_{[0,1]}(x) \tag{30}$$

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

$$f(z \mid a, b, \lambda) = \frac{\lambda^a\, z^{a-1}}{\mathrm{B}(a,b)\, (1 + \lambda z)^{a+b}}\, \mathbf{1}_{[0,\infty)}(z) \tag{31}$$

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. 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. (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.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, 42(1):204–223. doi:10.1214/aoms/1177693507.

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.

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. (2018). 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. 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.

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. 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.

```
## print R sessionInfo to see system information and package versions
## used to compile the manuscript
sessionInfo()

## R version 4.2.0 (2022-04-22)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.4 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] ReplicationSuccess_1.2 hypergeo_1.2-13        dplyr_1.0.9
## [4] xtable_1.8-4           colorspace_2.0-3       ggplot2_3.3.5
## [7] ppRep_0.42             knitr_1.39
##
## loaded via a namespace (and not attached):
##  [1] RColorBrewer_1.1-3 ggpubr_0.4.0       highr_0.9          pillar_1.7.0
##  [5] compiler_4.2.0     tools_4.2.0        digest_0.6.29      viridisLite_0.4.0
##  [9] evaluate_0.15      lifecycle_1.0.1    tibble_3.1.6       gtable_0.3.0
## [13] pkgconfig_2.0.3    rlang_1.0.2        cli_3.3.0          DBI_1.1.2
## [17] xfun_0.30          withr_2.5.0        stringr_1.4.0      generics_0.1.2
## [21] vctrs_0.4.1        contfrac_1.1-12    elliptic_1.4-0     cowplot_1.1.1
## [25] grid_4.2.0         tidyselect_1.1.2   glue_1.6.2         deSolve_1.32
## [29] R6_2.5.1           rstatix_0.7.0      fansi_1.0.3        carData_3.0-5
## [33] car_3.0-12         tidyr_1.2.0        farver_2.1.0       purrr_0.3.4
## [37] magrittr_2.0.3     backports_1.4.1    scales_1.2.0       ellipsis_0.3.2
## [41] MASS_7.3-57        abind_1.4-5        assertthat_0.2.1   ggsignif_0.6.3
## [45] labeling_0.4.2     utf8_1.2.2         stringi_1.7.6      munsell_0.5.0
## [49] broom_0.8.0        crayon_1.5.1
```