Mixture priors for replication studies

Roberto Macrì Demartino[*a], **Leonardo Egidi**[b], **Leonhard Held**[c], and **Samuel Pawel**[c]

[a] Department of Statistical Sciences, University of Padova, Via C. Battisti 241, Padova, 35121, Italy.

[b] Department of Economics, Business, Mathematics, and Statistics "Bruno de Finetti", University of Trieste, Via A. Valerio 4/1, Trieste, 34127, Italy

[c] Epidemiology, Biostatistics and Prevention Institute, Center for Reproducible Science, University of Zurich, Hirschengraben 84, Zurich, 8001, Switzerland

Abstract

Replication of scientific studies is important for assessing the credibility of their results. However, there is no consensus on how to quantify the extent to which a replication study replicates an original result. We propose a novel Bayesian approach based on mixture priors. The idea is to use a mixture of the posterior distribution based on the original study and a non-informative distribution as the prior for the analysis of the replication study. The mixture weight then determines the extent to which the original and replication data are pooled.

Two distinct strategies are presented: one with fixed mixture weights, and one that introduces uncertainty by assigning a prior distribution to the mixture weight itself. Furthermore, it is shown how within this framework Bayes factors can be used for formal testing of scientific hypotheses, such as tests regarding the presence or absence of an effect. To showcase the practical application of the methodology, we analyze data from three replication studies. Our findings suggest that mixture priors are a valuable and intuitive alternative to other Bayesian methods for analyzing replication studies, such as hierarchical models and power priors. We provide the free and open source R package `repmix` that implements the proposed methodology.

Keywords: Bayesian inference, Borrowing, Effect size, Evidence synthesis, Historical data

1 Introduction

The integrity and credibility of scientific research heavily relies on the replicability of its results (National Academies of Sciences, Engineering, and Medicine, 2019). However, in recent years, an increasing number of published findings failed to replicate, leading to growing concerns about a "replication crisis" in several scientific fields (Open Science Collaboration, 2015; Camerer et al., 2016, 2018; Errington et al., 2021). As a consequence, there is an increasing emphasis within the scientific community on the importance of replication studies (NWO, 2016; Nature Communications, 2022). Establishing the success of a replication remains a challenging task. Multiple statistical methodologies, ranging from frequentist to Bayesian paradigms and even hybrid models of both, have been suggested to quantify the degree of success a replication study achieved in replicating the original result (Bayarri and Mayoral, 2002a,b; Verhagen and Wagenmakers, 2014; Patil et al., 2016; Johnson

*Corresponding author e-mail: `roberto.macridemartino@phd.unipd.it`

et al., 2017; Harms, 2019; Hedges and Schauer, 2019; Held, 2020; Mathur and VanderWeele, 2020; Pawel and Held, 2020, 2022; Held et al., 2022b; Micheloud et al., 2023; Pawel et al., 2024, among others).

Analyzing replication studies involves per definition the use of historical data – the data from the original study. Given the inherent nature of sequential information updating, Bayesian methods are natural for this purpose. Consequently, an intuitive way to incorporate historical information is to use a prior distribution based on the data from the original study for the analysis of the replication data. In its simplest form, one could use the posterior distribution of the model parameters based on the original data as the prior for the replication analysis. However, this may be problematic if there is heterogeneity between the two studies, as the resulting posterior may then conflict with both studies. This is of particular concern in the replication setting, where replication studies often show less impressive effects than their original counterparts (e.g., the effect estimates in Open Science Collaboration, 2015, were on average only half as large as the original ones), often argued to happen because of stricter control of biases and researcher degrees of freedom, for example, via preregistration of the replication study.

A variety of more sophisticated Bayesian methods have been proposed to mitigate potential conflict between historical and current data, and "borrow information" from the historical data in an adaptive way (for an overview see e.g., Lesaffre et al., 2024). Notably, power priors (Chen and Ibrahim, 2000; Duan et al., 2006), hierarchical models (Thall et al., 2003; Berry et al., 2013), and mixture priors (Schmidli et al., 2014; Yang et al., 2023) are three prominent approaches in this domain. The power prior, in its basic version, is derived by updating an initial prior distribution with the likelihood of the historical data raised to the power parameter δ, ranging between zero and one, which determines the degree to which historical data influences the prior distribution. Power priors evaluate two primary concepts of successful replication (Pawel et al., 2024). Firstly, they ensure the replication study confirms the presence of a tangible effect, often by assessing the effect size θ, and checking if it differs significantly from zero. Secondly, they assess how well the original data matches with the replication data, as a δ value close to one means both studies align seamlessly, while a value close to zero implies a disagreement between the original and the replication study.

Hierarchical modeling offers an alternative way to incorporate historical data into Bayesian analyses. The idea is to assume a hierarchical model where the true original θ_o and replication effect sizes θ_r are sampled themselves from a distribution around an overall effect size θ. The variance τ^2 of this distribution then determines the similarity between the studies, a value of zero corresponding to identical true effects while a large value corresponds to heterogeneity. Works by Bayarri and Mayoral (2002a,b) and Pawel and Held (2020) have effectively applied this approach in replication scenarios.

Mixture priors represent yet another way to adaptively borrow information from historical data (O'Hagan and Pericchi, 2012; Egidi et al., 2022). Essentially, a mixture prior combines a prior based on the historical data with a non-informative one, allocating distinct mixing weights to each component. The informative prior encourages information borrowing, while the non-informative prior indicates limited or no use of historical information. The robust meta-analytic predictive (MAP) prior presented by Schmidli et al. (2014), which mixes a MAP prior derived from multiple historical studies with a non-informative prior, is an example of a mixture prior used for historical data borrowing. In the replication setting, Consonni and Egidi (2023) have proposed a mixture prior modification of the reverse-Bayes method from Pawel and Held (2022) to limit prior-data conflict between the original study and a "sceptical prior" that is used to challenge it. The set of 21 replication studies from the Social Sciences Replication Project have also been jointly analyzed with a Bayesian mixture model to estimate an overall true positive rate and an effect size deflation factor (Camerer et al., 2018, see also `https://osf.io/nsxgj`).

However, apart from these two works, mixture prior modeling has not been applied to replication studies in any way, particularly not in its most basic form of using a mixture prior based on the original study for the analysis of the replication study.

The aim of this paper is therefore to present a novel and conceptually intuitive Bayesian approach for quantifying replication success based on mixture priors. The idea is to use a mixture of the posterior distribution based on the original study and a non-informative distribution as the prior for the analysis of the replication study. The mixture weight then determines the extent to which the original and replication data are pooled. This methodology is illustrated using data from three replication studies, which were part of the replication project from Protzko et al. (2023), detailed in the following Section 2. Section 3 then describes the process of deriving mixture priors from data of an original study within a meta-analytic framework, presenting a general approach for integrating original data into the mixture prior. In this exploration, two distinct approaches are examined: the first fixes the mixture weights, while the second introduces uncertainty by assigning a prior distribution on the mixture weight parameter. In Section 4 different hypotheses regarding the underlying parameters of interests are examined. Bayes factors are derived offering a quantitative measure of evidence for one hypothesis over another. Finally, Section 5 provides concluding remarks about similarities and differences between the discussed method and established approaches, particularly hierarchical models and power priors. Additionally, the strengths and limitations of the mixture prior approach are emphasized, along with insights into potential extensions.

2 Running example

We examine a particular experiment from communication science titled "Labels" which was part of the large-scale replication project by Protzko et al. (2023). The original authors hypothesized that the type of label used by a person to describe another person can indicate something about the preferences of the person themselves. For example, when someone uses the term "denier" to describe someone else who does not believe in global warming, the authors hypothesized that this is an indication that the speaker believes in global warming.

The original study found evidence for this hypothesis. Its main finding was drawn from a sample of 1577 participants which led to a standardized mean difference effect estimate of $\hat{\theta}_o = 0.21$ and standard error $\sigma_o = 0.05$, suggesting a positive effect of "labelling". Subsequently, this experiment was replicated by three other labs. The first replication yielded a smaller effect estimate, with $\hat{\theta}_{r_1} = 0.09$ and $\sigma_{r_1} = 0.05$. In contrast, the other two replications reported either the same effect estimate, $\hat{\theta}_{r_2} = 0.21$ and $\sigma_{r_2} = 0.04$, or a larger one, $\hat{\theta}_{r_3} = 0.44$ and $\sigma_{r_3} = 0.06$, compared to the original study. Figure 1 shows the effect size estimates along with their 95% confidence intervals for the original study. its three independent replications, and the pooled replication.

3 Mixture prior modeling of replication studies

In the following, we use a meta-analytic framework which can be applied to a broad range of data types and effect sizes (see e.g., chapter 2.4 in Spiegelhalter et al., 2004). Define by θ the unknown effect size, with $\hat{\theta}_o$ and $\hat{\theta}_{r_i}$ being the estimated effect size from the original study and replication $i = 1, \ldots, m$, respectively. As assumed by Held (2020) and Pawel and Held (2022), it is common to specify that the likelihood of the effect size

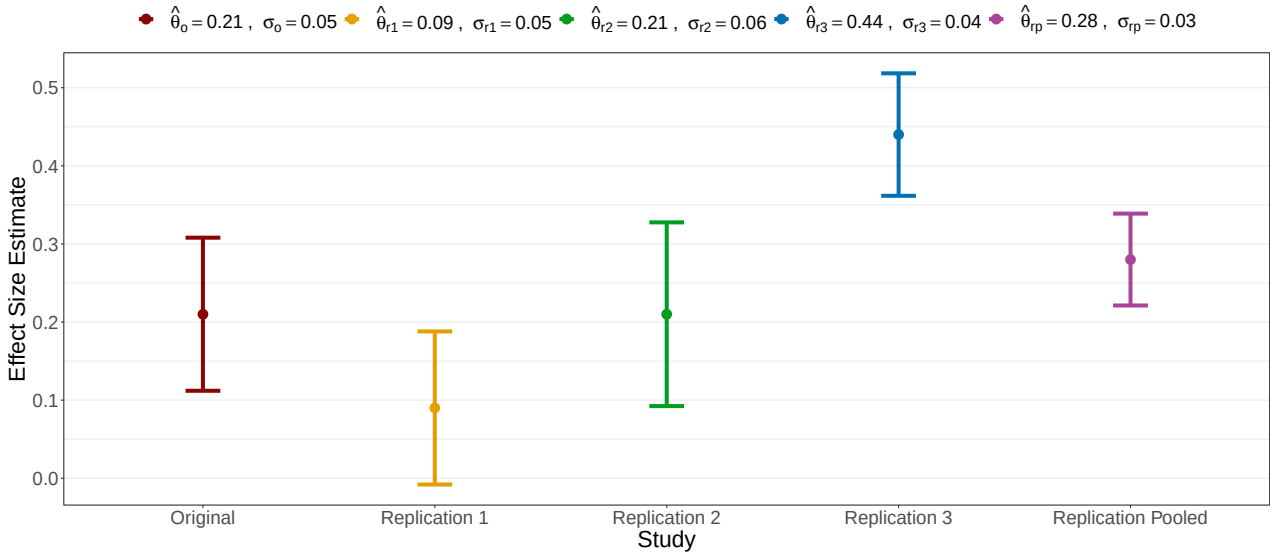

Figure 1: Effect size estimates (standardized mean difference) and 95% CI for the "Labels" original study, the three independent replications, and the pooled replication.

estimates is approximately normal

$$\hat{\theta}_o \mid \theta \sim \mathrm{N}\left(\theta, \sigma_o^2\right) \qquad \hat{\theta}_{r_i} \mid \theta \sim \mathrm{N}\left(\theta, \sigma_{r_i}^2\right),$$

where σ_i represents the standard error of an estimate, which is assumed to be known. There are circumstances under which the effect size might need a particular transformation, such as a logit function or a log function transformation, to refine the normal distribution approximation. Additionally, adjusting the effect size for confounders via regression might also be necessary. Finally, define the pooled replication effect size estimate and its standard error by

$$\hat{\theta}_{r_p} = \frac{\sum_{i=1}^{m} \hat{\theta}_{r_i}/\sigma_{r_i}^2}{\sum_{i=1}^{m} 1/\sigma_{r_i}^2} \qquad \sigma_{r_p} = \sqrt{\frac{1}{\sum_{i=1}^{m} 1/\sigma_{r_i}^2}},$$

which are sufficient statistics for inference regarding the effect size parameter θ, that is, we have that the likelihood of a sample of independent replication studies is

$$\prod_{i=1}^{m} \mathrm{N}(\hat{\theta}_{r_i} \mid \theta, \sigma_{r_i}^2) = K \times \mathrm{N}(\hat{\theta}_{r_p} \mid \theta, \sigma_{r_p}^2),$$

with $\mathrm{N}(\cdot \mid m, v)$ the normal density function with mean m and variance v and K a constant that does not depend on the effect size θ. In the following, we will investigate posterior distribution and Bayes factor analyses related to the effect size θ and based on the likelihood of the pooled replication effect size estimate and standard error $\hat{\theta}_{r_p} \mid \theta \sim \mathrm{N}(\theta, \sigma_{r_p}^2)$. For both analyses, the constant K cancels out and the approach thus encompasses both the analysis of a single replication study ($m = 1$ so that $\hat{\theta}_{r_p} = \hat{\theta}_{r_1}$ and $\sigma_{r_p} = \sigma_{r_1}$) or multiple replication studies ($m > 1$).

The aim is now to develop a mixture prior for the effect size θ that combines two distinct components. The first component is derived from the original study, akin to the meta-analytic-predictive (MAP) prior described by

Spiegelhalter et al. (2004) and Neuenschwander et al. (2010); and the second component is a normal prior that provides an alternative in case there is conflict between the replication and original data

$$\pi(\theta \mid \hat{\theta}_o, \omega) = \omega \mathrm{N}(\theta \mid \hat{\theta}_o, \sigma_o^2) + (1-\omega)\mathrm{N}(\theta \mid \mu, \tau^2). \tag{1}$$

The mean μ and variance τ^2 of the alternative are typically specified such that the prior is proper but non-informative (e.g., $\mu = 0$ and τ^2 large). Clearly, by setting $\omega = 1$, we obtain a prior that leads to a complete pooling of the data from both studies, while setting $\omega = 0$ completely discounts the original data. For a $0 < \omega < 1$, there is a gradual compromise between these two extremes.

In a mixture prior as in (1), setting an appropriate mixing weight ω, is a complex but crucial task. It is essential that the chosen ω accurately reflects the level of agreement between the original and replication studies. A prior that places too much weight towards the non-informative component can undermine the effectiveness of borrowing from the original study, leading to an underestimate of the real agreement between the original and replication studies. On the contrary, a mixing weight skewed heavily towards the informative prior may result in overestimating the confidence on the similarity between the two studies, introducing a potential bias. In the following we will discuss two strategies for determining the value of ω. The first strategy involves fixing ω on a predetermined value that is considered reasonable, while the second employs an additional prior specification by taking ω as a random quantity.

3.1 Fixed weight parameter

After observing the replication data, the mixture prior (1) is updated yielding the posterior distribution

$$\pi(\theta \mid \hat{\theta}_o, \hat{\theta}_r, \omega) = \frac{\mathrm{N}(\hat{\theta}_r \mid \theta, \sigma_r^2)\left\{\omega \mathrm{N}(\theta \mid \hat{\theta}_o, \sigma_o^2) + (1-\omega)\mathrm{N}(\theta \mid \mu, \tau^2)\right\}}{f(\hat{\theta}_r \mid \hat{\theta}_o, \omega)}, \tag{2}$$

where the marginal likelihood is

$$\begin{aligned} f(\hat{\theta}_r \mid \hat{\theta}_o, \omega) &= \int_\Theta \mathrm{N}(\hat{\theta}_r \mid \theta, \sigma_r^2)\left\{\omega \mathrm{N}(\theta \mid \hat{\theta}_o, \sigma_o^2) + (1-\omega)\mathrm{N}(\theta \mid \mu, \tau^2)\right\} \mathrm{d}\theta \\ &= \omega \mathrm{N}(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_r^2 + \sigma_o^2) + (1-\omega)\mathrm{N}(\hat{\theta}_r \mid \mu, \sigma_r^2 + \tau^2). \end{aligned} \tag{3}$$

For a normal mixture model, there exists a closed-form solution for the marginal likelihood, as denoted in equation (3), based on which it can be shown that the posterior is again a mixture of two normals

$$\pi(\theta \mid \hat{\theta}_o, \omega) = \omega' \mathrm{N}(\theta \mid m_1, v_1) + (1-\omega')\mathrm{N}(\theta \mid m_2, v_2),$$

with updated means and variances

$$\begin{aligned} m_1 &= (\hat{\theta}_o/\sigma_o^2 + \hat{\theta}_r/\sigma_r^2) \times v_1, & v_1 &= (1/\sigma_o^2 + 1/\sigma_r^2)^{-1}, \\ m_2 &= (\mu/\tau^2 + \hat{\theta}_r/\sigma_r^2) \times v_2, & v_2 &= (1/\tau^2 + 1/\sigma_r^2)^{-1}, \end{aligned}$$

and updated weight

$$\omega' = \left\{1 + \frac{1-\omega}{\omega} \times \frac{\mathrm{N}(\hat{\theta}_r \mid \mu, \tau^2 + \sigma_r^2)}{\mathrm{N}(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_o^2 + \sigma_r^2)}\right\}^{-1}.$$

The two posterior components thus represent two ordinarily updated normal posteriors, while the initial weight along with the relative predictive accuracy of the replication data under either component determines the updated weight. The fact that for a fixed mixture weight, the posterior distribution is again a mixture distribution is known from general Bayesian theory (Bernardo and Smith, 1994; Neuenschwander et al., 2023). The mixture representation of the posterior also shows that the non-informative component has to be proper ($\tau^2 < \infty$) to enable borrowing, as otherwise the updated weight will be $\omega' = 1$, leading always to a complete pooling with the historical data regardless of conflict.

There are different approaches for specifying the mixture weight ω. A straightforward approach involves assigning to ω a value that is reasonable, based on domain-expert knowledge, regarding the agreement between the two studies. Alternatively, the empirical Bayes estimate of ω may be used, which represents the value that maximizes the marginal likelihood function (3). Finally, in order to assess prior sensitivity, a reverse-Bayes approach (Good, 1950; Best et al., 2021; Held et al., 2022a) may be used to find the mixture weight such that a certain posterior distribution is obtained.

Returning to the "Labels" experiment from Protzko et al. (2023) introduced in Section 2, Figure 2 shows the shifts in the posterior distribution for the effect size (2) under different fixed weights assigned to the mixture prior. Here, the non-informative prior component in (1) is constructed to be a unit-information normal distribution centred at a mean $\mu = 0$ and with variance $\tau^2 = 2$. A unit-information prior (Kass and Wasserman, 1995) is structured to provide only a minimal amount of information. Essentially, its variance is set so that the prior's information has a content equivalent to a unit sample size. The use of unit-information prior is illustrated in several studies such as Ntzoufras et al. (2003) for binary response models, Overstall and Forster (2010) apply this principle to generalized linear mixed models, and Sabanés Bové and Held (2011) demonstrate its application in the context of generalized linear models. For further details see also Consonni et al. (2018). We see that, varying ω, within the range from 0 to 1, induces a progressive transformation of the posterior distribution. At $\omega = 0$, the posterior distribution virtually aligns with the likelihood of the replication study as the influence of the non-informative component is minimal compared to the replication data. Conversely, as ω increases towards 1, the posterior distribution gradually becomes more influenced by the prior associated with the original study, leading to a posterior that lies somewhere in between the replication and original likelihood, as the replication borrows information from the original study.

Based on the reverse-Bayes approach (Best et al., 2021; Held et al., 2022a), a tipping point analysis was conducted to assess the influence of the mixing weight ω on the resulting posterior distribution. This analysis focuses on the question: "How much does the mixing weight has to change for the conclusion of the analysis to change?" Figure 3 shows the posterior median and the 95% highest posterior density interval (HPDI) of the effect size for each weight value associated to the original study component in (1). We see that the second, third, and the pooled replication scenarios are robust with respect to the choice of weights, as the effect size posterior median and its corresponding 95% HPDI remain substantially above zero across all prior weights, thereby suggesting robust evidence for a genuine effect. In contrast, the first replication is less stable, as the 95% HPDI includes zero up to about a weight of 0.1. Thus, this first replication can only be considered as providing

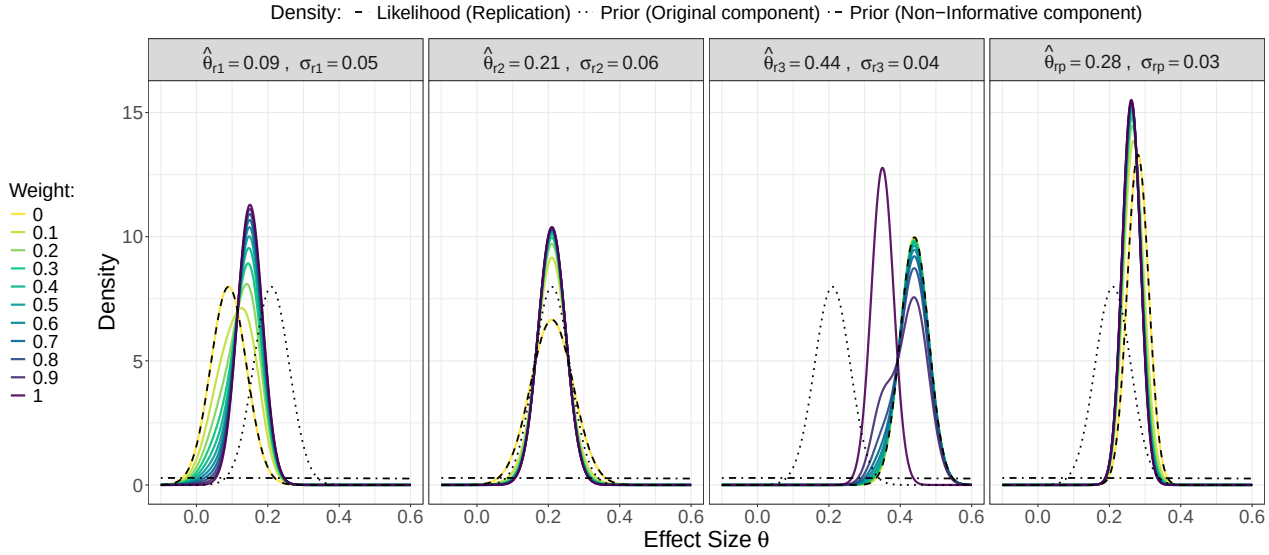

Figure 2: "Labels" experiment. Each colored line represents the posterior distribution of effect size θ under different fixed weights. The black dotted line indicates the prior distribution based on the original study. The dash-dotted line represents a non-informative prior. The dashed line corresponds to the likelihood based on replication data.

evidence for a genuine effect if a mixture weight of at least 0.1 seems plausible, as the replication study alone (i.e., a mixture weight of zero) fails to do so. It is important to note that the posterior median can be a misleading point estimate in the case of bimodality. This does not seem to be a problem in our analysis, as only the posterior of the third replication shows a slight "hump" in the posterior distribution for certain weight values, while the posteriors of the remaining replications appear unimodal. However, if assessing bimodality by looking at the posterior density is not possible, it may be advisable to at least compute numerical summaries that quantify potential bimodality (see e.g. sections 2.2-2.10 in O'Hagan and Forster, 2004).

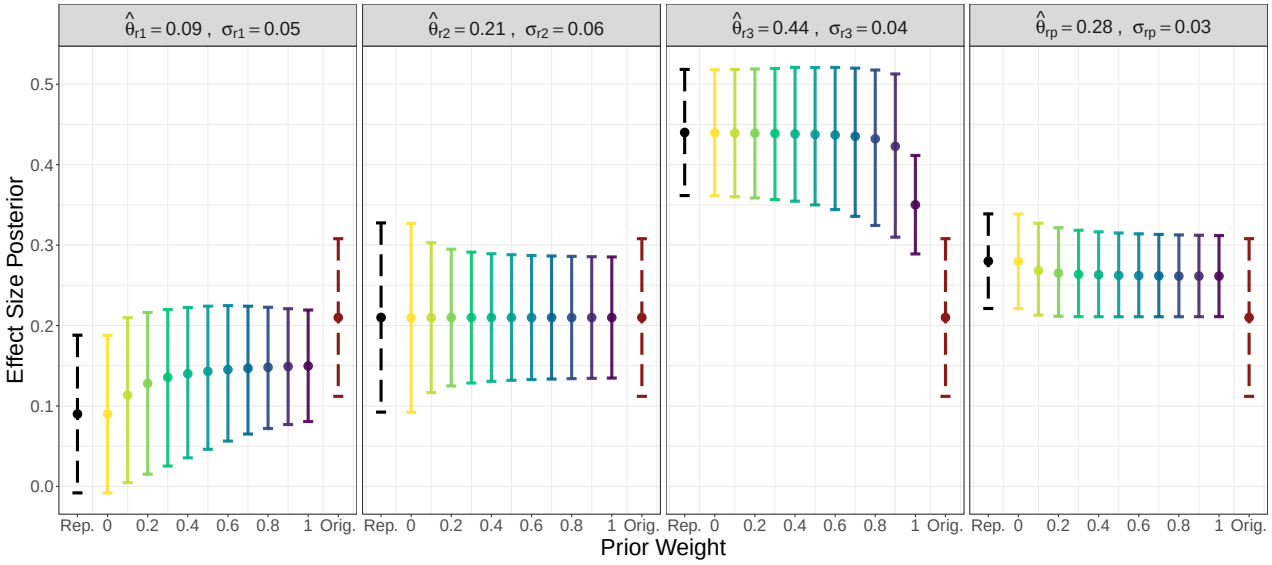

Figure 3: "Labels" experiment. Posterior median (points) and 95% highest posterior density interval (HPDI) of the effect size posterior against mixture prior weight assigned to the original study component. On the left and right side of each panel, the corresponding replication study effect estimate and the the original study effect estimate with 95% confidence interval.

3.2 Prior on the weight parameter

We now introduce an extension of the mixture prior in (1) assuming uncertainty on the weight ω. This approach considers ω as a random quantity, requiring the specification of a prior distribution $\pi(\omega)$. A natural choice is a Beta distribution

$$\omega \mid \eta, \nu \sim \text{Beta}(\eta, \nu),$$

since ω is a proportion. Consequently, this formulation leads to the joint prior distribution for the effect size θ and the weight ω

$$\begin{aligned} \pi(\theta, \omega \mid \hat{\theta}_o, \eta, \nu) &= \pi(\omega \mid \eta, \nu)\pi(\theta \mid \omega, \hat{\theta}_o) \\ &= \text{Beta}(\omega \mid \eta, \nu) \times \left\{\omega \text{N}(\theta \mid \hat{\theta}_o, \sigma_o^2) + (1-\omega)\text{N}(\theta \mid \mu, \tau^2)\right\}, \end{aligned} \tag{4}$$

where $\text{Beta}(\cdot \mid \eta, \nu)$ is the Beta density function with the strictly positive shape parameters $\eta, \nu > 0$. Given the joint prior distribution (4) and in light of the replication data, the joint posterior distribution is then

$$\pi(\theta, \omega \mid \hat{\theta}_r, \hat{\theta}_o, \eta, \nu) = \frac{\text{N}(\hat{\theta}_r \mid \theta, \sigma_r^2) \times \text{Beta}(\omega \mid \eta, \nu) \times \left\{\omega \text{N}(\theta \mid \hat{\theta}_o, \sigma_o^2) + (1-\omega)\text{N}(\theta \mid \mu, \tau^2)\right\}}{f(\hat{\theta}_r \mid \hat{\theta}_o, \eta, \nu)}. \tag{5}$$

The marginal likelihood in the normal mixture model with random weights can be determined through a closed-form solution, similar to Equation (3). In this scenario, it depends on the expected value of the weight parameter ω and on the updated normal prior components of the mixture

$$\begin{aligned} f(\hat{\theta}_r \mid \hat{\theta}_o, \eta, \nu) &= \int\int \text{N}(\hat{\theta}_r \mid \theta, \sigma_r^2) \times \text{Beta}(\omega \mid \eta, \nu) \times \left\{\omega \text{N}(\theta \mid \hat{\theta}_o, \sigma_o^2) + (1-\omega)\text{N}(\theta \mid \mu, \tau^2)\right\} \text{d}\theta\text{d}\omega \\ &= \int \text{Beta}(\omega \mid \eta, \nu) \times \left\{\omega \text{N}(\hat{\theta}_r \mid \sigma_r^2 + \sigma_o^2) + (1-\omega)\text{N}(\hat{\theta}_r \mid \mu, \sigma_r^2 + \tau^2)\right\} \text{d}\omega \\ &= \left(\frac{\eta}{\eta+\nu}\right) \times \left\{\text{N}(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_r^2 + \sigma_o^2) - \text{N}(\hat{\theta}_r \mid \mu, \sigma_r^2 + \tau^2)\right\} + \text{N}(\hat{\theta}_r \mid \mu, \sigma_r^2 + \tau^2). \end{aligned}$$

Consequently, in the case of a random weight, the marginal likelihood is similar to that in Equation (3), with the difference of replacing the fixed weight with the expected weight over the prior. By integrating θ out in (5), the marginal posterior distribution of ω can be expressed as

$$\pi(\omega \mid \hat{\theta}_r, \hat{\theta}_o, \eta, \nu) = \frac{\text{Beta}(\omega \mid \eta, \nu) \times \left\{\omega \text{N}(\hat{\theta}_r, \sigma_r^2 + \sigma_o^2) + (1-\omega)\text{N}(\hat{\theta}_r \mid \mu, \sigma_r^2 + \tau^2)\right\}}{f(\hat{\theta}_r \mid \hat{\theta}_o, \eta, \nu)}. \tag{6}$$

The marginal posterior of θ is given by

$$\pi(\theta \mid \hat{\theta}_r, \hat{\theta}_o, \eta, \nu) = \frac{\text{N}(\hat{\theta}_r \mid \theta, \sigma_r^2) \times \left(\frac{\eta}{\eta+\nu}\right) \times \left\{\text{N}(\theta \mid \hat{\theta}_o, \sigma_o^2) - \text{N}(\theta \mid \mu, \tau^2)\right\} + \text{N}(\theta \mid \mu, \tau^2)}{f(\hat{\theta}_r \mid \hat{\theta}_o, \eta, \nu)}.$$

In summary, when introducing uncertainty in the mixture weight ω via a Beta prior, the marginal likelihood of the data, the joint and marginal posteriors of the effect size θ, and the mixture weight ω are still available in closed-form. Moreover, the marginal likelihood and marginal posterior of θ are of the same form as with a fixed mixture weight ω as shown in the previous section, but with ω replaced by its expected value under its prior distribution.

Figure 4 shows the contour plot of the joint posterior distribution for the effect size θ and the weight parameter ω considering the data from the "Labels" experiment, its three replications, and the pooled replication. In our analysis, we employ a mixture prior, as in (4), in which the informative prior component is derived from the original study, while the non-informative prior is a unit-informative prior as in Section 3.1. Additionally, we adopt a flat prior distribution for the weight parameter choosing a Beta$(1, 1)$. We see that for the first, second,

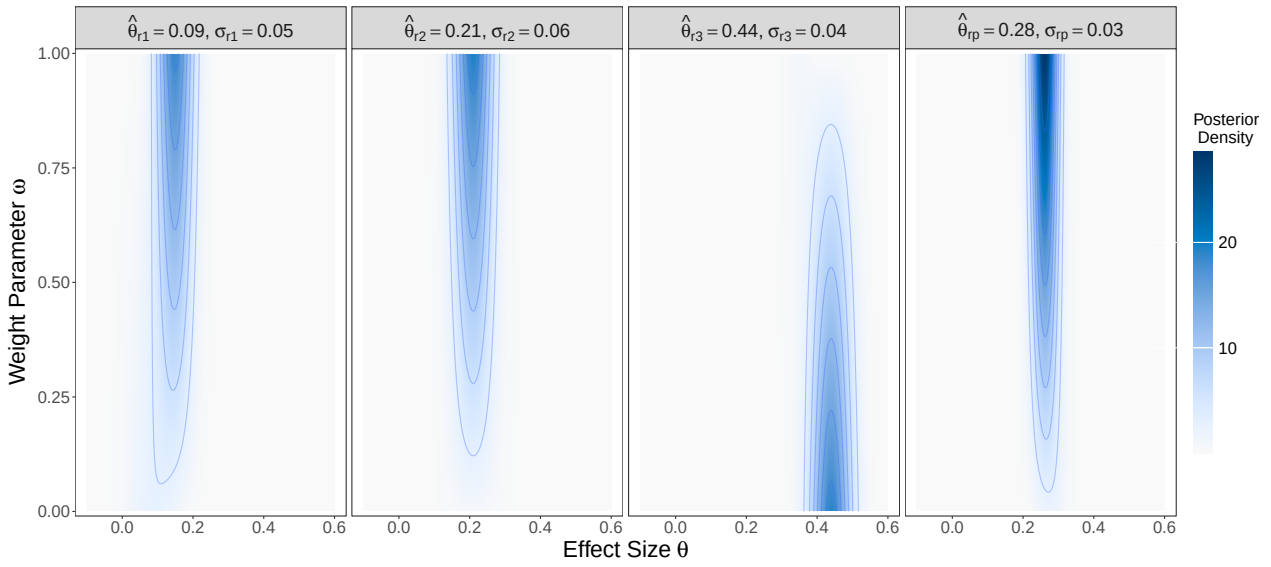

Figure 4: Joint posterior distribution of the effect size θ and the weight parameter ω considering the data from the "Labels" experiment, its three external replications, and the pooled replication.

and pooled replications, the posterior distribution is concentrated around weight parameter values close to one, reflecting the similarity between the original and replication results. In contrast, for the third replication the posterior distribution is concentrated around zero, indicating a conflict between the original study and the results of this replication. In addition, because it is based on three replications instead of just one, the posterior based on the pooled replications is much more peaked than the others.

Figure 5 displays the marginal posterior distributions of the effect size θ (left) and the weight parameter ω (right). The plot related to θ is enriched by contrasting it with the posterior distribution of θ based solely on the replication data, represented as a dashed line. This effectively illustrates the added value from integrating the original data through a mixture prior. The blue marginal posterior, corresponding to the most divergent estimate $\hat{\theta}_{r_3} = 0.44$, shows a tendency to incorporate less information, leading to a more heavy-tailed posterior distribution despite the smallest standard error associated, $\sigma_{r_3} = 0.04$, among the three external replications. The discrepancy with the original study increases the variance of the posterior distribution, as evident when comparing with the replication-only posterior shown by the dashed blue line. This is further highlighted in the 95% HPDI, which ranges from 0.35 to 0.52, slightly exceeding the 95% HPDI range of 0.36 to 0.52 observed when the replication data is analyzed without considering the original study, represented by the dashed horizontal blue bar. Conversely, the green marginal posterior, associated with the most coherent replication $\hat{\theta}_{r_2} = 0.21$, results in a noticeably narrower 95% HPDI compared to the one derived solely from the replication data. Additionally, the magenta marginal posterior based on the pooled replication $\hat{\theta}_{r_p} = 0.21$ results to be the most peaked density. It is worth noting that these marginal posteriors are equivalent to those obtained when the weight parameter is fixed at $\omega = 0.5$, as shown in Figure 2, because the expected value of a Beta$(1, 1)$ distribution is 0.5.

The right panel in Figure 5 shows the marginal posterior distribution for the weight parameter ω, under the assumption of a flat prior distribution for ω. Following the formula as detailed in (6) and under a flat prior, this yields a linearly increasing/decreasing posterior density. However, for non-flat priors (i.e., Beta(η, ν) with $\eta \neq 1$ and $\nu \neq 1$), the posterior density of the weight ω is not linear anymore. The first, second, and pooled replications, highlighted in yellow, green, and magenta respectively, display linear marginal posterior distributions that increase monotonically, indicating a peak at $\omega = 1$. This suggests compatibility between the two replications and the pooled replication with respect to the original study. Conversely, the linear marginal distribution of the third replication, illustrated in blue, exhibits a monotonically decreasing trend with the most probable value at $\omega = 0$. This trend suggests a notable disagreement between this replication and the original study. Nevertheless, it is worth noting that the HPDIs remain considerably wide across all the replication scenarios, despite their large sample sizes.

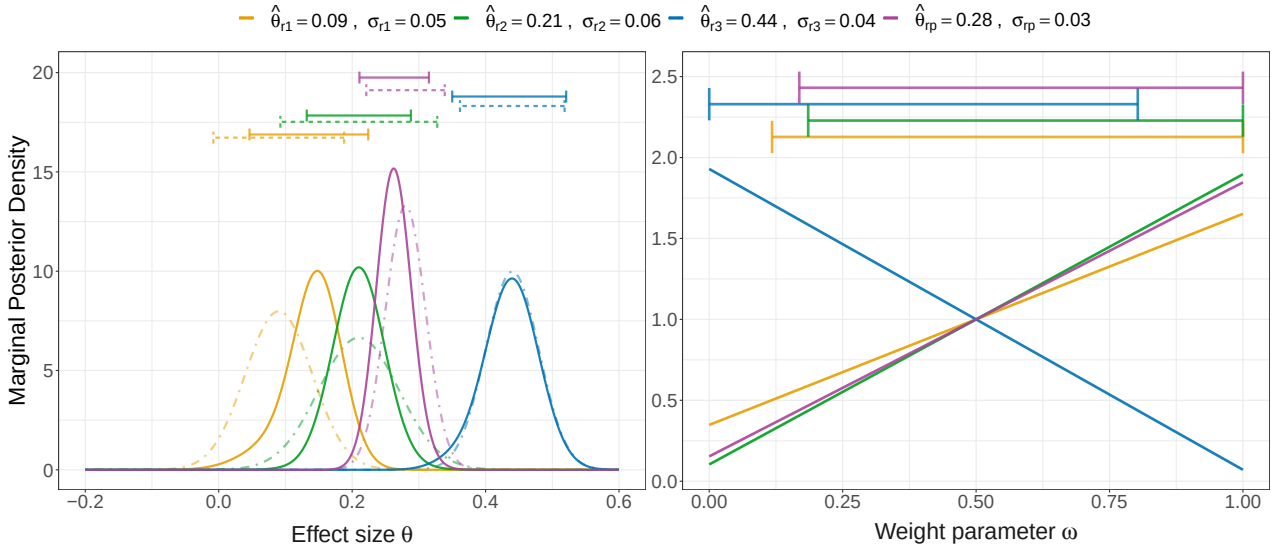

Figure 5: Marginal posterior distributions of the effect size θ (left) and the weight parameter ω (right) considering the data from the "Labels" experiment, its three external replications, and the pooled replication. The dashed lines represent the posterior density of the effect size θ, derived exclusively from the replication data, without considering the original data, and assuming a uniform prior for the effect size $\pi(\theta) \propto 1$. The horizontal error bars indicate the 95% highest posterior density credible intervals (HPDI).

4 Hypothesis testing

Estimating the parameters of a model is one aspect, but in statistical analysis one may also want to test the plausibility of different scientific hypotheses. Within the Bayesian framework, the Bayes factor is a key tool for assessing and comparing hypotheses about the parameters (Good, 1958; Jeffreys, 1961; Kass and Raftery, 1995; Gronau et al., 2017; Schönbrodt and Wagenmakers, 2018; Schad et al., 2023, among others). Let us consider the replication data $\hat{\theta}_r$ and let $\mathcal{H}_0$ and $\mathcal{H}_1$ be two competing hypothesis. The Bayes Factor is then given by the updating factor of the prior odds of the hypotheses to their posterior odds

$$\text{BF}_{01} = \frac{\Pr(\mathcal{H}_0 \mid \hat{\theta}_r)}{\Pr(\mathcal{H}_1 \mid \hat{\theta}_r)} \Big/ \frac{\Pr(\mathcal{H}_0)}{\Pr(\mathcal{H}_1)} = \frac{f(\hat{\theta}_r \mid \mathcal{H}_0)}{f(\hat{\theta}_r \mid \mathcal{H}_1)},$$

which simplifies to the ratio of marginal likelihoods (or evidences) as shown by the second equality. As such, the Bayes Factor is a quantitative tool to measure the relative evidence that we have for $\mathcal{H}_0$ over $\mathcal{H}_1$. For example, when the a priori probabilities of both hypotheses are assumed to be equal, a Bayes factor greater than one indicates that the data are more likely under $\mathcal{H}_0$ than $\mathcal{H}_1$. Conversely, a Bayes factor less than one suggests that $\mathcal{H}_1$ is more in agreement with the observed data. A value approximately equal to one implies that the data do not distinctly favor any model, indicating similar levels of empirical support for $\mathcal{H}_0$ and $\mathcal{H}_1$. To interpret the Bayes Factor effectively, various categorizations have been proposed. One of the most notable is outlined by Jeffreys (1961), as detailed in Table 1.

Table 1: Scale of evidence proposed by Jeffreys (1961).

BF_{01}	$\log_{10}(\text{BF}_{01})$	Evidence for $\mathcal{H}_0$
1 to 3.2	0 to 0.5	Barely worth mentioning
3.2 to 10	0.5 to 1	Substantial evidence
10 to 31.6	1 to 1.5	Strong evidence
31.6 to 100	1.5 to 2	Very strong evidence
> 100	> 2	Decisive evidence

4.1 Hypothesis testing for the mixture weight ω

To determine how closely the replication aligns with the original study, we may perform hypothesis testing on the mixture weight parameter ω. A key goal is testing whether the original and replication studies are consistent with each other, formulated as the hypothesis $\mathcal{H}_c : \omega = 1$. This hypothesis may be tested against the alternative hypothesis that suggests the data from the studies should be entirely disregarded, indicated as $\mathcal{H}_d : \omega = 0$. Contrary to the power prior approach (Pawel et al., 2024), the point hypothesis $\mathcal{H}_d : \omega = 0$ avoids leading to an improper mixture prior. As a result, the Bayes factor derived in this context does not encounter problematic issues related to the dependence on the ratio of the two arbitrary constants, since it is based on the ratio of two well-defined marginal likelihoods

$$\begin{aligned}\text{BF}_{dc}(\hat{\theta}_r \mid \mathcal{H}_d : \omega = 0) &= \frac{f\{\hat{\theta}_r \mid \mathcal{H}_d : \theta \mid \omega \sim \omega\text{N}(\hat{\theta}_o, \sigma_o^2) + (1-\omega)\text{N}(\mu, \tau^2), \omega = 0\}}{f\{\hat{\theta}_r \mid \mathcal{H}_c : \theta \mid \omega \sim \omega\text{N}(\hat{\theta}_o, \sigma_o^2) + (1-\omega)\text{N}(\mu, \tau^2), \omega = 1\}} \\ &= \frac{\text{N}(\hat{\theta}_r \mid \mu, \sigma_r^2 + \tau^2)}{\text{N}(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_r^2 + \sigma_o^2)}.\end{aligned} \tag{7}$$

A more flexible hypothesis to consider is that the data exhibit a certain level of compatibility or disagreement. A suitable hypothesis is defined by the prior class $\mathcal{H}_d : \omega \sim \text{Beta}(1, \nu)$, where $\nu > 1$. In this class of distributions, the density is maximized at $\omega = 0$ and decreases consistently from there. This encodes a hypothesis where the importance of the original data is systematically reduced. The degree of this reduction is dictated by the parameter ν. In the asymptotic case where $\nu \to \infty$, the hypothesis simplifies to $\mathcal{H}_d : \omega = 0$, implying a complete

discounting of the original data. Consequently, the Bayes factor is

$$\begin{aligned}\text{BF}_{dc}\{\hat{\theta}_r \mid \mathcal{H}_d : \omega \sim \text{Beta}(1,\nu)\} &= \frac{f\{\hat{\theta}_r \mid \mathcal{H}_d : \theta \mid \omega \sim \omega\text{N}(\hat{\theta}_o, \sigma_o^2) + (1-\omega)\text{N}(\mu, \tau^2), \omega \sim \text{Beta}(1,\nu)\}}{f\{\hat{\theta}_r \mid \mathcal{H}_c : \theta \mid \omega \sim \omega\text{N}(\hat{\theta}_o, \sigma_o^2) + (1-\omega)\text{N}(\mu, \tau^2), \omega = 1\}} \\ &= \frac{\left(\frac{\eta}{\eta+\nu}\right) \times \left\{\text{N}(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_r^2 + \sigma_o^2) - \text{N}(\hat{\theta}_r \mid \mu, \sigma_r^2 + \tau^2)\right\} + \text{N}(\hat{\theta}_r \mid \mu, \sigma_r^2 + \tau^2)}{\text{N}(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_r^2 + \sigma_o^2)}.\end{aligned}$$

4.2 Hypothesis testing for the effect size θ

In the assessment of hypotheses regarding the magnitude of the effect size θ, the analysis typically involves a comparative evaluation between the null hypothesis, $\mathcal{H}_0 : \theta = 0$, which posits absence of the effect, and the alternative hypothesis, $\mathcal{H}_1 : \theta \neq 0$, suggesting the presence of an effect. The null hypothesis $\mathcal{H}_0$ represents a singular value within the possible range of θ values, while the alternative hypothesis $\mathcal{H}_1$ requires a prior specification for both θ and ω.

To address this, the use of a mixture prior as in equation (1) is proposed. Specifically, the first mixture prior component is based on the empirical data from the original study $\hat{\theta}_o$ while the second component is designed to have the same amount of information content equivalent to a single observation. This approach is complemented by the specification of a suitable Beta prior for the weight parameter ω. Consequently, the Bayes factor is

$$\begin{aligned}\text{BF}_{01}\{\hat{\theta}_r \mid \mathcal{H}_1 : \omega \sim \text{Beta}(\eta,\nu)\} &= \frac{f(\hat{\theta}_r \mid \mathcal{H}_0 : \theta = 0)}{f\{\hat{\theta}_r \mid \mathcal{H}_1 : \theta \mid \omega \sim \omega\text{N}(\hat{\theta}_o, \sigma_o^2) + (1-\omega)\text{N}(\mu, \tau^2), \omega \sim \text{Beta}(\eta,\nu)\}} \\ &= \frac{\text{N}(\hat{\theta}_r \mid 0, \sigma_r^2)}{\left(\frac{\eta}{\eta+\nu}\right) \times \left\{\text{N}(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_r^2 + \sigma_o^2) - \text{N}(\hat{\theta}_r \mid \mu, \sigma_r^2 + \tau^2)\right\} + \text{N}(\hat{\theta}_r \mid \mu, \sigma_r^2 + \tau^2)}.\end{aligned}$$

It is important to emphasize that, as parallel discussed in Pawel et al. (2024) for the power parameter in the power prior approach, assigning a point mass to the weight parameter $\omega = 1$ leads to the Bayes factor contrasting a point null hypothesis to the posterior distribution of the effect size based on the original data that is the *replication Bayes factor* under normality (Verhagen and Wagenmakers, 2014; Pawel and Held, 2022). In detail, it is

$$\begin{aligned}\text{BF}_{01}\{\hat{\theta}_r \mid \mathcal{H}_1 : \omega = 1\} &= \frac{f(\hat{\theta}_r \mid \mathcal{H}_0 : \theta = 0)}{f\{\hat{\theta}_r \mid \mathcal{H}_1 : \theta \mid \omega \sim \omega\text{N}(\hat{\theta}_o, \sigma_o^2) + (1-\omega)\text{N}(\mu, \tau^2), \omega = 1\}} \\ &= \frac{\text{N}(\hat{\theta}_r \mid 0, \sigma_r^2)}{\text{N}(\hat{\theta}_r \mid \hat{\theta}, \sigma_r^2 + \sigma_o^2)}.\end{aligned}$$

Similar to the power prior formulation, the mixture prior version of the replication Bayes factor represents a generalization of the standard replication Bayes factor that provides a flexible and controlled approach for combining original and replication data.

4.3 Posterior distribution and Bayes factor asymptotics

In delving deeper into the proposed mixture model, a key focus is on examining the asymptotic characteristics of the marginal posterior distribution and the Bayes factor for the weight parameter. Specifically, let us consider the Bayes factor contrasting $\mathcal{H}_d\colon \theta \sim \text{N}(\mu, \tau^2)$ to $\mathcal{H}_c\colon \theta \sim \text{N}(\hat{\theta}_o, \sigma^2)$ for the replication data $\hat{\theta}_r \mid \theta \sim \text{N}(\theta, \sigma_r^2)$ as

in (7). Subsequently, the marginal posterior distribution in (6) can be expressed in terms of the Bayes factor

$$\begin{aligned}\pi(\omega \mid \hat{\theta}_r, \hat{\theta}_o) &= \frac{\pi(\omega)\left\{\omega \mathrm{N}(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_r^2 + \sigma_o^2) + (1-\omega)\mathrm{N}(\hat{\theta}_r \mid \mu, \sigma_r^2 + \tau^2)\right\}}{\left(\frac{\eta}{\eta+\nu}\right) \times \left\{\mathrm{N}(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_r^2 + \sigma_o^2) - \mathrm{N}(\hat{\theta}_r \mid \mu, \sigma_r^2 + \tau^2)\right\} + \mathrm{N}(\hat{\theta}_r \mid \mu, \sigma_r^2 + \tau^2)} \\ &= \frac{\pi(\omega)\left\{\omega + (1-\omega)\mathrm{BF}_{dc}(\hat{\theta}_r)\right\}}{\left(\frac{\eta}{\eta+\nu}\right) \times \left\{1 - \mathrm{BF}_{dc}(\hat{\theta}_r)\right\} + \mathrm{BF}_{dc}(\hat{\theta}_r)}.\end{aligned} \tag{8}$$

We investigate the behavior of the limiting marginal posterior distribution in (8) when the Bayes factor tends to zero and when it tends towards infinity, respectively. In these cases, we have that

$$\begin{aligned}\lim_{\mathrm{BF}_{dc}(\hat{\theta}_r)\downarrow 0} \pi(\omega \mid \hat{\theta}_r, \hat{\theta}_o) &= \frac{\pi(\omega)}{\mathbb{E}_{\pi(\omega)}(\omega)}\omega = \mathrm{Beta}(\omega \mid \eta + 1, \nu), \\ \lim_{\mathrm{BF}_{dc}(\hat{\theta}_r)\uparrow +\infty} \pi(\omega \mid \hat{\theta}_r, \hat{\theta}_o) &= \frac{\pi(\omega)}{1 - \mathbb{E}_{\pi(\omega)}(\omega)}(1-\omega) = \mathrm{Beta}(\omega \mid \eta, \nu + 1).\end{aligned}$$

This means that even when we find overwhelming evidence in favor of $\mathcal{H}_d$ or $\mathcal{H}_c$, the posterior distribution is only slightly changed from the prior (i.e., "updated by one observation" from the prior). For example, for a flat prior with $\nu = \eta = 1$, the limiting posteriors are given by the Beta(2, 1) and Beta(1, 2) distributions, respectively, which correspond to densities that are linearly increasing (decreasing) from 0 (2) to 2 (0). We can see from Figure 5 that the marginal posteriors for the second, third, and pooled "Labels" replications are not too different from these two asymptotic distributions.

While the previous calculations assumed that the Bayes factor can go to infinity or zero, thereby overwhelmingly favoring one of the contrasted models, it is unclear whether this is even possible. We therefore now aim to explore the effects on the Bayes factor (7) when the standard error of the replication study σ_r becomes arbitrarily small. This could occur due to an increase in the sample size which is typically inversely related to the squared standard error. The limiting Bayes factor as the replication standard error σ_r approaches zero is

$$\lim_{\sigma_r^2 \downarrow 0} \mathrm{BF}_{dc}(\hat{\theta}_r) = \frac{\mathrm{N}(\hat{\theta}_r \mid \mu, \tau^2)}{\mathrm{N}(\hat{\theta}_r \mid \hat{\theta}_o, \sigma_o^2)}.$$

Consequently, for finite τ^2 and σ_o^2, the Bayes factor is bounded and cannot converge to either zero or $+\infty$. However, if the original standard error σ_o also approaches zero, the Bayes factor in (7) behaves differently. In this case, the Bayes factor approaches

$$\lim_{\sigma_r^2, \sigma_o^2 \downarrow 0} \mathrm{BF}_{dc}(\hat{\theta}_r) = \frac{\mathrm{N}(\hat{\theta}_r \mid \mu, \tau^2)}{\delta_{\hat{\theta}_o}(\hat{\theta}_r)},$$

where $\delta_{\hat{\theta}_o}(\cdot)$ represents the Dirac delta function. When the standard errors from both original and replication go to zero, the Bayes factor thus shows the correct asymptotic behavior, converging to zero when their effect sizes are the same while converging to infinity when they are not.

4.4 Hypothesis testing for the "Labels" experiment

We will now illustrate the results of the proposed hypothesis tests in Sections 4.1 and 4.2 using the data obtained from the "Labels" experiment, as described in Section 2. Specifically, a comprehensive analysis is performed to understand the behaviour of these parameters in the replication study.

To evaluate the agreement between the initial study and subsequent replications, Table 2 shows the results of the hypothesis tests concerning the mixture weight parameter ω. The fourth column shows the Bayes factor contrasting two point hypotheses: $\mathcal{H}_d : \omega = 0$ and $\mathcal{H}_c : \omega = 1$. This analysis reveals substantial and strong evidence for $\mathcal{H}_c$ in the first, second, and pooled replication scenarios, respectively. Conversely, the third replication study shows strong evidence favoring $\mathcal{H}_d$. While the Bayes factors based on the Beta$(1, 2)$ prior under $\mathcal{H}_d$ (fifth column) still point in the same direction, the extent of evidence is lower than for the point hypothesis.

Table 2: "Labels" experiment. Hypothesis tests for the mixture weight ω.

Replication	$\hat{\theta}_r$	σ_r	$\text{BF}_{\text{dc}}(\hat{\theta}_r \mid \mathcal{H}_d : \omega = 0)$	$\text{BF}_{\text{dc}}\{\hat{\theta}_r \mid \mathcal{H}_d : \omega \sim \text{Beta}(1, 2)\}$
1	0.09	0.05	1/4.8	1/2.1
2	0.21	0.06	1/18	1/2.7
3	0.44	0.04	27	19
Pooled	0.28	0.03	1/12	1/2.6

Table 3 presents the outcomes of our hypothesis tests regarding the effect size parameter θ. Specifically, the fourth column shows the Bayes factors contrasting the null hypothesis ($\mathcal{H}_0 : \theta = 0$) to the alternative hypothesis ($\mathcal{H}_1 : \theta \neq 0$), with a Beta$(1, 1)$ prior for the weight parameter ω under $\mathcal{H}_1$. The results suggest that there is absence of evidence for either hypothesis in the first replication. Conversely, the Bayes factor $\text{BF}_{01}\{\hat{\theta}_r \mid \mathcal{H}_1 : \omega \sim \text{Beta}(1, 1)\}$ indicates strong evidence in favor of $\mathcal{H}_1$ for the second, third, and pooled replications. In addition, the evidence from the replication Bayes factor under normality, $\text{BF}_{01}(\hat{\theta}_r \mid \mathcal{H}_1 : \omega = 1)$, shown in the last column, leads to the same qualitative conclusions thus corroborating the previous findings.

Table 3: "Labels" experiment. Hypothesis tests for the effect size θ.

Replication	$\hat{\theta}_r$	σ_r	$\text{BF}_{01}\{\hat{\theta}_r \mid \mathcal{H}_1 : \omega \sim \text{Beta}(1, 1)\}$	$\text{BF}_{01}(\hat{\theta}_r \mid \mathcal{H}_1 : \omega = 1)$
1	0.09	0.05	1.2	2
2	0.21	0.06	1/351	1/185
3	0.44	0.04	< 1/1000	< 1/1000
Pooled	0.28	0.03	< 1/1000	< 1/1000

In summary, the findings from our analysis indicate that among the three replications, only the second one aligns with the original study's results and also offers evidence for a non-zero effect. While the first replication slightly aligns with the initial findings, it fails to offer substantial evidence supporting an effect different from zero. Conversely, the third replication presents significant evidence for an effect that is non-zero but does not align with the findings of the original study. However, when pooled, the replications align with the original study's findings and provide evidence for a non-zero effect, indicating that the replication effort was successful overall.

5 Discussion

In this paper, we introduced a novel Bayesian method for analyzing data from replication studies. By using a mixture prior that mixes the posterior based on the original study with a non-informative prior, our method addresses the issue of potential conflict between original and replication study, as in such cases the information from the original study can be discounted. A crucial element is the mixture weight parameter ω. We explored two distinct strategies for setting this weight parameter. The first strategy involves fixing the weight to a specific value, for example, on the basis of expert knowledge or an empirical Bayes estimate. The sensitivity of this choice may then be assessed with a reverse-Bayes tipping point analysis (Best et al., 2021; Held et al., 2022a). The second strategy introduces a level of uncertainty by assigning a prior distribution to the mixture weight parameter. We then showed that the prior on the weight strategy is equivalent to the fixed weight strategy using the expected value of the weight's prior as fixed weight. However, the uncertain weight strategy also provides data analysts with a posterior distribution of the weight, which can be used for quantitatively assessing the degree of study compatibility, yet the extent to which this posterior can be updated from the prior was also shown to be limited. Importantly, both strategies yield the same results for the effect size when the fixed weight equals the expectation of the prior distribution. The only difference lies in the additional posterior distribution for the weight parameter.

Scientists should choose between these two strategies based on the characteristics of their study. Fixed weights can be more straightforward and are based on prior knowledge. They are suitable in situations where there is a reasonable confidence about the degree of agreement between the original and replication studies. A tipping point analysis can additionally help to assess how robust the analysis is to the choice of the weight. On the other hand, the random weight approach provides an additional posterior distribution for the weight parameter, showing the uncertainty related to this parameter.

We also presented Bayesian hypothesis tests for assessing the magnitude of the effect size θ and to determine how closely the replications align with the original study. We analyzed the asymptotic behavior of the marginal posterior distribution for the weight parameter when the Bayes factor tends to zero or towards infinity. Moreover, we examined how the Bayes factor related to the effect size behaves as the replication study's standard error σ_r tends to zero. Our findings reveal that the Bayes factor contrasting $\mathcal{H}_d : \theta \sim \mathrm{N}(\mu, \tau^2)$ to $\mathcal{H}_c : \theta \sim \mathrm{N}(\hat{\theta}_o, \sigma^2)$, for finite τ^2 and σ_o^2, is inconsistent. However, when the original study's standard error σ_o also approaches zero, the behavior of the Bayes factor changes, leading to correct asymptotic behavior and consistency.

The mixture prior approach we developed presents some similarities with two well-established methods in the replication setting – power priors (Pawel et al., 2024) and hierarchical models (Bayarri and Mayoral, 2002a,b; Pawel and Held, 2020). All three approaches exhibit similar strengths in assessing differences between original and replication studies providing valuable inferences that complement each other. Analogously to the heterogeneity variance and the power parameter, the mixture weight ω controls the degree of compatibility between the original and replication studies. Nevertheless, we think that our approach has some practical advantages. First, the mixture weight parameter ω seems to be a more straightforward and intuitive discounting measure, making this approach more accessible for the analysts. Second, the inherent structure of the mixture prior provides computational advantages. Notably, the calculation of the marginal likelihood in the random weight scenario is similar to that in the fixed weight scenario, with the only difference being the replacement of the fixed weight with the expected weight over the prior, which is computationally advantageous. This is particularly evident when compared to the computationally-prohibitive normalizing constant of the normalized

power prior (Lesaffre et al., 2024). Finally, when multiple original studies are involved, our mixture prior approach may facilitates their inclusion into the analysis. Specifically, this can be achieved by using two or more informative components derived from the original studies, along with a non-informative component.

Our method relies on the widely-used meta-analytic assumption that the distribution of effect estimates can be accurately approximated by a normal distribution with known variance, making it adaptable to a broad range of effect sizes from various data models across different research fields. However, this assumption becomes too strong in presence of small sample sizes and/or extreme effect size values at the boundary of the parameter space (e.g., very small or large probabilities). Future research could thus adapt our approach to specific data models (e.g., binomial or t-student distribution), especially in the presence of small sample sizes.

In this paper, we analyze replications both individually by directly comparing each one with the original study, and simultaneously by pooling them into a unique replication without assuming heterogeneity among the replications. An alternative pooling approach would be to assume a hierarchical model for the replication effect sizes that incorporates potential between-replication heterogeneity with a heterogeneity variance parameter. However, it remains unclear how to specify this parameter or a prior distribution for it. Consequently, an opportunity for future research could be to explore methods to specify a fixed value for this additional parameter or to elicit a prior distribution for it.

Software and Data Availability

All analyses were conducted in the R programming language version 4.2.3 (R Core Team, 2023). The code and data to reproduce this manuscript are openly available at `https://github.com/RoMaD-96/MixRep`. We provide an R package `repmix` for analysis of replication studies using the mixture prior framework. The package is currently available on GitHub and can be installed by running `remotes::install_github(repo = "SamCH93/repmix")` (requiring the `remotes` package available on CRAN). We plan to release the package on CRAN in the future.

References

Bayarri, M. and Mayoral, A. (2002a). Bayesian analysis and design for comparison of effect-sizes. *Journal of Statistical Planning and Inference*, 103(1):225–243.

Bayarri, M. J. and Mayoral, A. M. (2002b). Bayesian design of "successful" replications. *The American Statistician*, 56(3):207–214.

Bernardo, J. and Smith, A. (1994). *Bayesian Theory*. Wiley.

Berry, S. M., Broglio, K. R., Groshen, S., and Berry, D. A. (2013). Bayesian hierarchical modeling of patient subpopulations: Efficient designs of phase II oncology clinical trials. *Clinical Trials*, 10(5):720–734. PMID: 23983156.

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.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J., and Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644.

Chen, M.-H. and Ibrahim, J. G. (2000). Power prior distributions for regression models. *Statistical Science*, 15(1):46 – 60.

Consonni, G. and Egidi, L. (2023). Assessing replication success via skeptical mixture priors. *arXiv preprint arXiv:2401.00257*.

Consonni, G., Fouskakis, D., Liseo, B., and Ntzoufras, I. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, 13(2):627 – 679.

Duan, Y., Ye, K., and Smith, E. P. (2006). Evaluating water quality using power priors to incorporate historical information. *Environmetrics*, 17(1):95–106.

Egidi, L., Pauli, F., and Torelli, N. (2022). Avoiding prior–data conflict in regression models via mixture priors. *Canadian Journal of Statistics*, 50(2):491–510.

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., and Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, 10:e71601.

Good, I. (1950). Probability and the weighing of evidence. *Journal of the Institute of Actuaries*, 76(3):293–296.

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

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., and Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, 81:80–97.

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

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.

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

Held, L., Matthews, R., Ott, M., and Pawel, S. (2022a). Reverse-Bayes methods for evidence assessment and research synthesis. *Research Synthesis Methods*, 13(3):295–314.

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

Jeffreys, H. (1961). *The Theory of Probability*. Oxford University Press, third edition.

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., and Mandal, S. (2017). On the reproducibility of psychological science. *Journal of the American Statistical Association*, 112(517):1–10. PMID: 29861517.

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

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Statist. Assoc.*, 90(431):928–934.

Lesaffre, E., Qi, H., Banbeta, A., and van Rosmalen, J. (2024). A review of dynamic borrowing methods with applications in pharmaceutical research. *Brazilian Journal of Probability and Statistics*, 38(1).

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.

Micheloud, C., Balabdaoui, F., and Held, L. (2023). Assessing replicability with the sceptical p-value: Type-I error control and sample size planning. *Statistica Neerlandica*, 77(4):573–591.

National Academies of Sciences, Engineering, and Medicine (2019). *Reproducibility and Replicability in Science*. National Academies Press.

Nature Communications (2022). Replication studies hold the key to generalization [editorial]. *Nature Communications*, 13(1).

Neuenschwander, B., Capkun-Niggli, G., Branson, M., and Spiegelhalter, D. J. (2010). Summarizing historical information on controls in clinical trials. *Clinical Trials*, 7(1):5–18. PMID: 20156954.

Neuenschwander, B., Wandel, S., Roychoudhury, S., and Schmidli, H. (2023). On fixed and uncertain mixture prior weights. *arXiv preprint arXiv:2306.15197*.

Ntzoufras, I., Dellaportas, P., and Forster, J. J. (2003). Bayesian variable and link determination for generalised linear models. *Journal of Statistical Planning and Inference*, 111(1):165–180.

NWO (2016). Make replication studies a normal part of science.

O'Hagan, A. and Forster, J. (2004). *Kendall's Advanced Theory of Statistic 2B*. Wiley & Sons, Chichester, second edition.

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

Overstall, A. M. and Forster, J. J. (2010). Default Bayesian model determination methods for generalised linear mixed models. *Computational Statistics & Data Analysis*, 54(12):3269–3288.

O'Hagan, A. and Pericchi, L. (2012). Bayesian heavy-tailed models and conflict resolution: A review. *Brazilian Journal of Probability and Statistics*, 26(4):372 – 401.

Patil, P., Peng, R. D., and Leek, J. T. (2016). What should researchers expect when they replicate studies? a statistical view of replicability in psychological science. *Perspectives on Psychological Science*, 11(4):539–544. PMID: 27474140.

Pawel, S., Aust, F., Held, L., and Wagenmakers, E.-J. (2024). Power priors for replication studies. *TEST*, 33:127–154.

Pawel, S. and Held, L. (2020). Probabilistic forecasting of replication studies. *PLOS ONE*, 15(4):1–23.

Pawel, S. and Held, L. (2022). The sceptical Bayes factor for the assessment of replication success. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):879–911.

Protzko, J., Krosnick, J., Nelson, L. D., Nosek, B. A., Axt, J., Berent, M., Buttrick, N., DeBell, M., Ebersole, C. R., Lundmark, S., and et al. (2023). High replicability of newly discovered social-behavioural findings is achievable. *Nature Human Behaviour*.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sabanés Bové, D. and Held, L. (2011). Hyper-g priors for generalized linear models. *Bayesian Analysis*, 6(3):387 – 410.

Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., and Vasishth, S. (2023). Workflow techniques for the robust use of Bayes factors. *Psychological Methods*, 28(6):1404–1426.

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.

Schönbrodt, F. D. and Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, 25(1):128–142.

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

Thall, P. F., Wathen, J. K., Bekele, B. N., Champlin, R. E., Baker, L. H., and Benjamin, R. S. (2003). Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes. *Statistics in Medicine*, 22(5):763–780.

Verhagen, J. and Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, 143(4):1457–1475.

Yang, P., Zhao, Y., Nie, L., Vallejo, J., and Yuan, Y. (2023). Sam: Self-adapting mixture prior to dynamically borrow information from historical data in clinical trials. *Biometrics*, 79(4):2857–2868.