TEST
https://doi.org/10.1007/s11749-025-00985-7

ORIGINAL PAPER

Assessing replication success via skeptical mixture priors

Guido Consonni[1] · **Leonardo Egidi**[2]

Received: 9 July 2024 / Accepted: 23 August 2025

Abstract

There is growing interest in the analysis of replication studies aimed at reassessing original findings across a wide range of scientific disciplines. In the context of hypothesis testing for effect sizes, two Bayesian approaches stand out for their principled use of the Bayes factor (BF): the replication BF and the skeptical BF. The latter, built around the skeptical prior, represents the perspective of an investigator who remains unconvinced by the original results and seeks to critically reassess them. In this paper, we adopt the skeptical viewpoint and introduce a novel mixture prior that incorporates skepticism while offering control over prior-data conflict. We study the consistency properties of the resulting skeptical mixture Bayes factor and examine its relationship to the standard skeptical BF. Through a focused simulation study, we conduct a sensitivity analysis of the skeptical mixture BF with respect to prior-data conflict, covering a range of plausible experimental scenarios. Our results show broad agreement with the standard skeptical BF under typical conditions. However, in situations where the standard skeptical BF suffers from severe prior-data conflict, our approach can yield a meaningful adjustment in the reported strength of replication success. Finally,we illustrate the practical application of our method using case studies from the Social Sciences Replication Project.

Keywords Bayes factor · Bayesian hypothesis testing · Consistency · Prior-data conflict · Replication studies · Reverse-Bayes

Mathematics Subject Classification 62A01 · 62F15

✉ Leonardo Egidi
legidi@units.it

Guido Consonni
guido.consonni@unicatt.it

[1] Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Largo A. Gemelli 1, 20123 Milano, Italy

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

Published online: 10 September 2025

1 Introduction and background

The so-called "replication crisis" has raised serious concerns about the reliability of scientific findings across a wide range of disciplines (Open Science Collaboration 2015; Camerer et al. 2018). This has sparked growing interest in the analysis of replication studies. Several attempts have been made to precisely define the notion of "replication success"; see, for example, Hutton et al. (2020), Anderson and Maxwell (2016), Johnson et al. (2017), Ly et al. (2019), Hedges and Schauer (2019), Harms (2019), and Held (2020).

Within the Bayesian framework, the Bayes factor (BF) (Kass and Raftery 1995) has become an important tool for evaluating the strength of evidence in replication studies. Two notable examples are the *replication* BF introduced by Verhagen and Wagenmakers (2014) and the *skeptical* BF developed by Pawel and Held (2022).

This paper focuses on the latter approach, which combines reverse-Bayes analysis (Good 1950) with Bayesian hypothesis testing. Specifically, the skeptical BF approach constructs a skeptical prior for the effect size such that the original study's findings are no longer convincing from a Bayesian perspective. This skeptical prior is then contrasted with an advocate prior, that is the reference posterior for the effect size obtained from the original study. Replication success is declared if the replication data favor the advocate prior over the skeptical prior more strongly than the original data favored the skeptical prior over the null hypothesis. The skeptical Bayes factor thus determines the highest level at which replication success can be claimed.

A key strength of the skeptical BF approach is its ability to integrate multiple aspects of replicability. In particular, it ensures that both the original and replication studies provide substantial evidence against the null hypothesis and penalizes discrepancies between their effect estimates. For further details, we refer the reader to (Pawel and Held 2022).

Building on this framework, we propose a novel extension based on the *skeptical mixture prior*. Our method introduces a flexible mixture prior governed by two hyperparameters: one controlling the prior mass assigned to the null effect in the discrete component, and another regulating the variance of the continuous component. This added flexibility helps to mitigate the consistency issues inherent in the standard skeptical prior. Moreover, it facilitates the assessment of prior-data conflict (Evans and Moshonov 2006) and enables sensitivity analysis to support practitioners in their evaluation.

The remainder of this paper is organized as follows. The rest of this section provides additional background and motivation for our methodology. Section 1.1 reviews the replication BF, while Section 1.2 outlines the main features of the skeptical BF. Sections 1.3 and 1.4 focus on the consistency properties of both the replication BF and the skeptical BF, examining scenarios where replication sample sizes increase, while the original data remain fixed. In this context, we show that the skeptical BF fails to achieve consistency, irrespective of the true effect size. Section 1.4 further discusses information consistency and argues that this property lies outside the scope of the skeptical BF due to the non-nested structure of the hypotheses.

In Section 1.5, we address the issue of prior-data conflict and discuss why this concern is particularly relevant for the skeptical approach in replication studies.

Section 2 constitutes the core of the paper. It presents our proposed skeptical mixture prior and its associated Bayes factor, examines its consistency properties, and highlights connections to the standard skeptical BF. Additionally, we illustrate the method through examples and report results from a focused simulation study designed to assess the sensitivity of the Bayes factor to prior-data conflict across a variety of plausible scenarios.

In Section 3, we apply our methodology to selected case studies from the Social Sciences Replication Project (Camerer et al. 2018) and assess its performance in terms of robustness and comparison with alternative Bayesian approaches.

Finally, Section 4 summarizes the main advantages of our proposed method and outlines possible extensions. To maintain the flow of the main discussion, technical details and proofs have been collected in the supplementary material.

1.1 The replication Bayes factor

Consider two Bayesian models for the same observable y

$$H_j : \{f(y \mid H_j, \theta_j);\ f(\theta_j \mid H_j)\},\ j = 1, 2, \tag{1}$$

where $f(y \mid H_j, \theta_j)$ is the sampling distribution under H_j indexed by parameter θ_j, and $f(\theta_j \mid H_j)$ is the corresponding parameter prior. We evaluate the plausibility of H_1 relative to H_2 based on data y through the Bayes factor (BF)

$$BF_{1:2}(y) = \frac{f(y \mid H_1)}{f(y \mid H_2)}, \tag{2}$$

where $f(y \mid H_j) = \int f(y \mid H_j, \theta_j) f(\theta_j \mid H_j) d\theta_j$ is the marginal likelihood of y under H_j, also named the marginal likelihood of H_j.

In expression (1) both the data distribution and the prior may depend on H_j. In our setting, however, the family of data distributions is the same under H_1 and H_2 with the same parameter θ say, so that $f(y \mid H_j, \theta_j) = f(y \mid \theta)$; as a consequence, H_j characterizes only the prior for θ, and model comparison reduces to a Bayesian hypothesis testing problem.

Let θ be the effect of a treatment on an outcome of interest, and $\hat{\theta}_o$ and $\hat{\theta}_r$ denote estimators (typically MLE) of θ obtained under the *original* and the *replication* study, respectively, with corresponding standard errors σ_o and σ_r. Following common practice in meta-analytic studies, we further assume that the sample sizes n_k are sufficiently large to justify a normal distribution for the estimators, so that $\hat{\theta}_k \mid \theta \sim N(\theta, \sigma_k^2)$ with σ_k known, $k \in \{o, r\}$. This represents a reasonable approximation for various types of effect sizes, including means and mean differences, odds ratios, hazard ratios, risk ratios or correlation coefficients, usually after a suitable transformation; see, for instance, (Spiegelhalter et al. 2003, Section 2.4).

The following notation will also be useful in the sequel. Denote the z-values associated to the estimates of the two studies with $z_o = \hat{\theta}_o/\sigma_o$, $z_r = \hat{\theta}_r/\sigma_r$, respectively; the relative effect estimate with $d = \hat{\theta}_r/\hat{\theta}_o$; the variance ratio with $c = \sigma_o^2/\sigma_r^2$. Since

for many types of effect sizes the variances are inversely proportional to the sample size, often one can safely assume that $\sigma_k^2 = \sigma^2/n_k, k \in \{o, r\}$, where σ^2 is the unitary variance in each study. In this case $c = n_r/n_o$, the ratio between the replication and the original sample size.

Of particular interest in our setting is the situation wherein z_o is sufficiently large in absolute value, so that the original experiment is believed to provide substantial evidence that there is truly an effect, i.e., $\theta \neq 0$. To evaluate to what extent a replication study resulted in a success, thus confirming the original finding, Verhagen and Wagenmakers (2014) compared two hypotheses using replication data $\hat{\theta}_r \mid \theta \sim N(\theta, \sigma_r^2)$. The first one is the standard null hypothesis $H_0 : \theta = 0$ of no effect. The second one reflects the opinion of an *advocate* who believes the effect to be consistent with that found in the original study. This is quantified through a posterior distribution on θ, conditionally on the original data $\hat{\theta}_o$, and based on a flat prior for θ. The resulting *advocate prior* becomes

$$H_A : \theta \sim N(\hat{\theta}_o, \sigma_o^2).$$

The BF of H_0 against H_A,

$$BF_{0:A}(\hat{\theta}_r) = \frac{f(\hat{\theta}_r \mid H_0)}{f(\hat{\theta}_r \mid H_A)} \equiv BF_R,$$

is named the *Replication Bayes factor*. It can be verified that

$$BF_R = \sqrt{1+c}\, \exp\left\{-\frac{z_o^2}{2}\left(d^2 c - \frac{(1-d)^2}{1/c+1}\right)\right\}. \tag{3}$$

Replication success is declared whenever BF_R is sufficiently low to provide convincing evidence against H_0, based on conventional evidence thresholds essentially dating back to Jeffreys (1961); see, for instance, (Schönbrodt and Wagenmakers 2018, Table 1). It may be observed that the replication BF is a *partial* BF (O'Hagan and Forster 2004) for checking H_0 against its complement $\theta \neq 0$ when the prior under H_A is flat. In this context, $\hat{\theta}_r$ is used as comparison data and $\hat{\theta}_o$ as training data; see also Ly et al. (2019). Notice that BF_R provides an answer to the following question: "In the replication experiment, is the effect absent or is it similar to what was found in the original one?", where the latter supposition is represented through H_A. This should be contrasted with more traditional default Bayesian testing methods, where the alternative is usually a relatively uninformative prior centered on the null value $\theta = 0$; see, for instance, Wetzels and Wagenmakers (2012). We highlight that BF_R establishes a useful connection between the replication and the original experiment, through the advocate prior. However, replication success is declared on the basis of the evidential strength against H_0 when compared to H_A *solely* under the replication data. In other words, there is no explicit consideration of the evidence against H_0 provided by the original data. This issue in taken up in the next section.

Table 1 Thirteen studies from the *Social Sciences Replication Project* (Camerer et al. 2018)

Study	z_o	z_r	n_o	n_r	c	d	g_S	P_S	$\psi_{SM,\alpha}$	$h_{SM,\alpha}$	BF_S	BF_R	BF_{SM}
$\alpha = 0.05$													
Aviezer et al	6.80	3.93	15	14	0.92	0.60	0.24	<0.001			0.01	<0.001	
Balafoutas and Sutter	2.37	2.28	72	243	3.48	0.52	0.25	0.03	0.56	0.97	0.64	0.26	0.55
Derex et al	4.04	2.97	51	65	1.29	0.65	0.40	<0.001	0.93	115.40	0.12	0.03	0.04
Duncan et al	2.83	4.41	15	92	7.42	0.57	0.50	0.02	0.66	2.67	0.32	<0.001	0.26
Gneezy et al	3.00	3.71	178	407	2.31	0.81	1.03	0.04	0.46	2.13	0.14	<0.001	0.14
Hauser et al	6.96	5.21	40	22	0.51	1.04	0.72	<0.001			<0.001	<0.001	
Janssen et al	5.76	2.24	63	42	0.65	0.48	0.03	<0.001			0.63	0.61	
Karpicke and Blunt	4.24	2.75	40	49	1.24	0.58	0.26	<0.001	0.94	1512.98	0.18	0.08	0.08
Kovacs et al	2.22	6.44	24	95	4.38	1.38	3.95	0.32	<0.001	0.29	0.31	<0.001	0.65
Morewedge et al	2.63	3.44	32	89	2.97	0.76	0.97	0.06	<0.001	0.81	0.26	0.01	0.28
Nishi et al	2.85	2.55	200	480	2.42	0.57	0.35	0.01	0.74	3.59	0.4	0.12	0.28
Pyc and Rawson	2.27	2.63	36	306	9.18	0.38	0.09	0.03	0.71	1.10	0.85	0.25	0.67
Wilson et al	4.25	4.10	30	39	1.33	0.83	0.83	<0.001	0.92	69.79	0.02	<0.001	0.01

Table 1 continued

Study	z_o	z_r	n_o	n_r	c	d	g_S	P_S	$\psi_{SM,\alpha}$	$h_{SM,\alpha}$	BF_S	BF_R	BF_{SM}
$\alpha = 0.1$													
Aviezer et al	6.80	3.93	15	14	0.92	0.60	0.24	<0.001			0.01	<0.001	
Balafoutas and Sutter	2.37	2.28	72	243	3.48	0.52	0.25	0.03	0.63	3.11	0.64	0.26	0.47
Derex et al	4.04	2.97	51	65	1.29	0.65	0.40	<0.001	0.88	343.51	0.12	0.03	0.04
Duncan et al	2.83	4.41	15	92	7.42	0.57	0.50	0.02	0.67	6.20	0.32	<0.001	0.22
Gneezy et al	3.00	3.71	178	407	2.31	0.81	1.03	0.04	0.61	5.82	0.14	<0.001	0.13
Hauser et al	6.96	5.21	40	22	0.51	1.04	0.72	<0.001			<0.001	<0.001	
Janssen et al	5.76	2.24	63	42	0.65	0.48	0.03	<0.001			0.63	0.61	
Karpicke and Blunt	4.24	2.75	40	49	1.24	0.58	0.26	<0.001			0.18	0.08	
Kovacs et al	2.22	6.44	24	95	4.38	1.38	3.95	0.32	<0.001	0.83	0.31	<0.001	0.44
Morewedge et al	2.63	3.44	32	89	2.97	0.76	0.97	0.06	0.41	2.55	0.26	0.01	0.24
Nishi et al	2.85	2.55	200	480	2.42	0.57	0.35	0.01	0.73	8.65	0.4	0.12	0.24
Pyc and Rawson	2.27	2.63	36	306	9.18	0.38	0.09	0.03	0.69	3.27	0.85	0.25	0.56
Wilson et al	4.25	4.10	30	39	1.33	0.83	0.83	<0.001	0.87	208.72	0.02	<0.001	0.01

Effect values, originally expressed as sample correlation coefficients, were subsequently turned into effect estimates $\hat{\theta}$ using Fisher z-transformation. Reported are the z-values for the original (z_o) and replication studies (z_r); $c = \sigma_o^2/\sigma_r^2$, and relative effect estimates $d = \hat{\theta}_r/\hat{\theta}_o$. Based on the choice $\gamma_S = BF_S$ and $\gamma_{SM} = BF_{SM}$ for skepticism, respectively, prior hyperparameters are shown, namely the relative variance g_S for the skeptical prior and the pair $(\psi_{SM,\alpha}, h_{SM,\alpha})$ for the skeptical mixture prior, the latter based on three possible prior-data conflict scenarios for $\alpha = \{0.05, 0.1\}$, when these thresholds are achievable. P_S indicates the p-value for prior-data conflict under the skeptical prior: note that the p-value for prior-data conflict under the skeptical mixture prior, if it exists, coincides with α by construction. The skeptical Bayes factors BF_S, the replication Bayes factor BF_R and the skeptical mixture Bayes factor $BF_{SM}(\alpha)$ are reported in the last three columns

1.2 The skeptical Bayes factor

Pawel and Held (2022) propose a different route to establish replication success. Their key idea is to compare two particular BF's: one based on the original data and the other one on the replication data. For the former they compare the standard null hypothesis of no effect H_0 against that of a *skeptic* who is unconvinced by the result. This view is operationalized through a *skeptical* Normal prior, centered on zero and with a variance $\sigma_S^2 = g\sigma_o^2$, where g is chosen so that the resulting BF provides unconvincing evidence against the null hypothesis. In other words, the skeptic wishes to "challenge" the original finding and requires the Bayes factor to attain a value so that s/he cannot take a definitive commitment against the null, and thus, further investigation (namely the replication experiment) is called for. More formally, let $0 < \gamma < 1$ be a level of skepticism and g_γ be the value of the relative sufficiently skeptical prior variance corresponding to this level γ such that the comparison of

$$H_0 : \theta = 0 \quad \text{vs} \quad H_S : \theta \sim N(0, g_\gamma \sigma_o^2) \tag{4}$$

leads to a Bayes factor of H_0 against H_S equal to γ, that is

$$BF_{0:S}(\hat{\theta}_o; g_\gamma) = \gamma,$$

where

$$BF_{0:S}(\hat{\theta}_o; g_\gamma) = \sqrt{(1 + g_\gamma)} \cdot \exp\left\{-\frac{1}{2} \cdot \frac{g_\gamma}{1 + g_\gamma} \cdot z_o^2\right\}. \tag{5}$$

In principle, γ would be set to a level such that values of $BF_{0:S}$ equal to, or lower than γ, would be considered adequate evidence against H_0.

For instance, $\gamma = 1/10$ could be a suitable choice, because values of $BF_{0:S}$ in the interval (1/10,1/3) provide only moderate evidence against H_0, while those in the interval (1/30, 1/10) imply strong evidence against H_0; see again Table 1 of Schönbrodt and Wagenmakers (2018) (notice, however, that the BF in their table is the reciprocal of ours). A slightly different classification scheme for Bayes factor evidence is available in Kass and Raftery (1995).

We note that the prior $N(0, g_\gamma \sigma_o^2)$ represented by H_S in (4) is named the skeptical prior (at level γ) and is constructed through a *reverse-Bayes* methodology, a technique dating back to Good (1950). The term "reverse" is employed because the prior is specified in such a way to induce a specific value for the BF *after* the data $\hat{\theta}_o$ will be collected; see Held et al. (2022) for an insightful discussion of reverse-Bayes ideas.

The value of g_γ, whose dependence on the original data is omitted for simplicity, can be explicitly computed as in Pawel and Held (2022, formula (3)). It must be pointed out, however, that g_γ will not exist when $BF_{0:S}(\hat{\theta}_o, g)$ is always above γ for any $g > 0$: this happens, for instance, when $|z_o| \leq 1$, and $\gamma \leq 1$; but it may also happen for $1 < |z_o| \leq 2$ if γ is smaller than 1/3. However, these situations are of little interest, as values of z_o smaller than two are generally not considered sufficient to claim the presence of an effect or justify replication. On the other hand, when $BF_{0:S} = \gamma$ is

attainable, there will typically be two values of g_γ leading to this result. The higher value, which is usually much greater than the smaller value, is merely an instance of the Jeffreys–Lindley paradox (Shafer 1982) and is accordingly discarded because it represents vagueness rather than skepticism.

Clearly, the skeptical prior is data-dependent; however, its use is confined to obtain a BF whose value, on the original data, is set based on external considerations. The skeptical distribution will then be used as a regular prior to construct a BF based on the replication data $\hat{\theta}_r$, and in that context is *not* data-dependent.

Turning to replication data, the next step involves comparing the skeptical prior $H_S : \theta \sim N(0, g_\gamma \sigma_o^2)$ against the advocate prior $H_A : \theta \sim N(\hat{\theta}_o, \sigma_o^2)$, leading to

$$BF_{S:A}(\hat{\theta}_r; g_\gamma) = \sqrt{\frac{1/c+1}{1/c+g_\gamma}} \exp\left\{-\frac{z_o^2}{2}\left(\frac{d^2}{1/c+g_\gamma} - \frac{(d-1)^2}{1/c+1}\right)\right\}, \tag{6}$$

where $z_o = \hat{\theta}_o/\sigma_o$, $d = \hat{\theta}_r/\hat{\theta}_o$, and $c = \sigma_o^2/\sigma_r^2$. Replication success at level γ is declared if

$$BF_{S:A}(\hat{\theta}_r; g_\gamma) \leq BF_{0:S}(\hat{\theta}_o; g_\gamma) = \gamma. \tag{7}$$

In the words of Pawel and Held (2022) "It is natural to consider a replication successful if the replication data favor the advocate over the skeptic to a higher degree than the skeptic's initial objection to the original study".

Rather than fixing a value γ and then checking whether Equation (7) holds, one might instead look for the smallest γ satisfying (7), namely

$$BF_S \equiv \inf\{\gamma : BF_{S:A}(\hat{\theta}_r; g_\gamma) \leq \gamma\}. \tag{8}$$

The value BF_S is called the *skeptical BF*, and represents the *smallest* γ level for which replication success can be established. Clearly, the smaller the BF_S, the stronger the degree of replication success. For instance, if $BF_S = 1/6$, this means that the replication experiment can at most successfully support the original findings with a moderate level of evidence. However, the level improves to strong if $BF_S = 1/20$, say.

It may happen that BF_S does not exist, because there is no γ for which replication success can be established, but this usually occurs when $|z_o|$, or $|d| = |\hat{\theta}_r|/|\hat{\theta}_o|$, or both are too small. More details are provided in Pawel and Held (2022).

1.3 Consistency

Model selection consistency (Liang et al. 2008), or simply consistency, is the property of a statistical procedure to recover the true model (or hypothesis) as the sample size grows. Below we analyze separately the behavior of the replication and the skeptical Bayes factor. In both cases consistency is evaluated relative to a sequence of replication datasets whose sample size is assumed to grow indefinitely.

Proposition 1 *Consider a sequence of replication datasets with increasing sample size* $n_r = 1, 2, \ldots$. *Assume there exists a corresponding sequence of estimators* $\{\hat{\theta}_r^{(n_r)}\}_{n_r=1}^{\infty}$ *of a common parameter* θ *whose distribution for sufficiently large* n_r *can be approximated as* $\hat{\theta}_r^{(n_r)} \mid \theta \sim N(\theta, (\sigma_r^{(n_r)})^2)$ *with* $\sigma_r^{(n_r)}$ *known. Denote with* $BF_R^{(n_r)}$ *the replication BF based on* $\hat{\theta}_r^{(n_r)}$. *Let* θ^* *denote the true value of* θ. *Then the following limits in probability hold*

$$\begin{aligned} &\text{if } \theta^* = 0, \quad BF_R^{(n_r)} \underset{n_r \to \infty}{\to} \infty \text{ at rate } O(\sqrt{n_r}); \\ &\text{if } \theta^* \neq 0, \quad BF_R^{(n_r)} \underset{n_r \to \infty}{\to} 0 \text{ at rate } \exp\{-K n_r\}, \end{aligned} \tag{9}$$

where $K > 0$ *is a positive constant. As a consequence,* BF_R *is* consistent.

Proof See supplementary material.

Proposition 2 *Under the assumptions of Proposition 1, let* $p_S(\cdot)$ *and* $p_A(\cdot)$ *denote the density of the skeptical and the advocate prior leading to* (6). *Let* θ^* *denote the true value of* θ. *Then the following limit in probability holds*

$$BF_{S:A}(\hat{\theta}_r^{(n_r)}; g) \underset{n_r \to \infty}{\to} \frac{p_S(\theta^*)}{p_A(\theta^*)}. \tag{10}$$

Proof See supplementary material.

It follows from Proposition 2 that consistency does not hold for $BF_{S:A}$ because it converges to a constant irrespective of the true value θ^*. Ly and Wagenmakers (2022), discussing Bayes factors for "peri-null" hypotheses, also mention, as a particular case, the inconsistency of $BF_{S:A}$. We note that both the consistency of BF_R and the inconsistency of $BF_{S:A}$ reported in Proposition 1 and Proposition 2, respectively, are in accord with theoretical results on the asymptotic behavior of Bayes factors under rather general conditions on model and priors presented in Dawid (2011).

Proposition 2 highlights the fact that the pair $\{H_S; H_A\}$ leading to $BF_{S:A}$ is a comparison between two *opinions* (priors) on the parameter for the *same* model. The bottom line is that even an infinite replication sample size cannot favor one over the other overwhelmingly.

1.4 Information consistency

Besides consistency, another useful criterion to evaluate a Bayes factor is *information consistency*. Bayarri et al. (2012) present this criterion with regard to two *nested* models, M_0 and M, with M_0 (the null model) nested in M. Let $\Lambda_{M_0:M}(y)$ be the usual likelihood ratio between M_0 and M given the data y, and consider a sequence of data vectors $\{y_m\}$ of *fixed* sample size, such that

$$\lim_{m \to \infty} \Lambda_{M:M_0}(y_m) = \infty, \tag{11}$$

so that, in the limit, the data provide overwhelming evidence in favor of M. It is then required that the BF in favor of M follows suit and diverges accordingly. We show in the supplementary material (Proposition S.1) that BF_R is information consistent. On the other hand, the notion of information consistency becomes vacuous when it comes to the skeptical $BF_{S:A}$, because the two models under comparison are not nested, as already mentioned at the end of Section 1.2. We note that some concerns about information consistency are addressed in Pawel and Held (2022 Section 3.4), to whom we refer for further details.

1.5 Prior-data conflict

The skeptical prior is constructed in such a way that $BF_{0:S}(\hat{\theta}_o; g_\gamma) = \gamma$ so that the original finding is made unconvincing at level γ. How reasonable is the skeptical prior $\theta \sim N(0, g_\gamma \sigma_o^2)$ relative to the information on θ provided by the data $\hat{\theta}_o$ in the original experiment?

The same concern applies to $N(0, g_S \sigma_o^2)$, where g_S is the value corresponding to BF_S defined in Equation (8) (assuming it exits). Surely a skeptical prior which is at odds with the original data would appear suspicious to an external agent (e.g., a regulatory agency, such as the Food and Drug Administration (FDA) or the European Medicines Agency (EMA)). Indeed being skeptical does not mean being unrealistic.

We address this issue using the notion of *prior-data conflict* (Evans and Moshonov 2006; Egidi et al. 2021); see also Held (2020) in the context of replication studies. Notice that both Evans and Moshonov (2006) and Held (2020) use this concept to define features of the prior for a given statistical model whose structure is not questioned. This is also the approach we follow in this paper. In this section we simply sketch the idea. Consider a statistic T having distribution $f_T(t|\theta)$ and a prior $\theta \sim \pi(\theta)$. The marginal density of T is given by

$$m_T(t) = \int f_T(t|\theta)\pi(\theta)d\theta, \tag{12}$$

where t ranges over the set of values of T. Let t_{obs} be the observed value of T. The p-value for prior-data conflict (Evans and Moshonov 2006) is defined as:

$$P(t_{obs}) = \Pr^{m_T}\{t : m_T(t) \leq m_T(t_{obs})\}, \tag{13}$$

where $Pr^{m_T}(\cdot)$ is the probability computed under the marginal $m_T(\cdot)$ in (12). The index $P(t_{obs})$ can be interpreted as a measure of surprise of the value t_{obs} relative to our uncertainty on T described in (13). Intuitively, if $P(t_{obs})$ is small, a surprising value has occurred, suggesting prior-data conflict. In particular, if $m_T(\cdot)$ is unimodal, $P(t_{obs})$ provides the tail probabilities under $m_T(\cdot)$, where the tails are the t-values whose density is below the cutoff $m_T(t_{obs})$.

2 The skeptical mixture prior and its Bayes factor

Our novel contribution is to generalize the skeptical prior $H_S : \theta \sim N(0, g_\gamma \sigma_o^2)$ employed in (4) to a *mixture* prior composed of a point mass and a continuous component. These type of priors have been already implemented as variants of the classic spike-and-slab prior (Mitchell and Beauchamp 1988), and they have also been used as data distribution in genomic studies (Taylor and Pollard 2009).

Specifically, we define the *family of skeptical mixture priors* at level $\gamma \in (0, 1)$ as

$$\tilde{H}_{SM} : \theta \sim \psi_\gamma \delta_0 + (1 - \psi_\gamma) N(0, h_\gamma \sigma_o^2), \quad (\psi_\gamma, h_\gamma) \in U_\gamma, \tag{14}$$

where $0 \leq \psi_\gamma \leq 1$ is a weight, δ_0 is the Dirac measure at $\theta = 0$, $h_\gamma > 0$ is the relative variance, and U_γ is the set of pairs (ψ_γ, h_γ) such that the BF for the comparison of $H_0 : \theta = 0$ against the hypothesis H_{SM} described in (14) is equal to γ, that is

$$BF_{0:SM}(\hat{\theta}_o; \psi_\gamma, h_\gamma) = \gamma, \tag{15}$$

then in symbols $U_\gamma = \{(\psi_\gamma, h_\gamma)\ s.t.\ BF_{0:SM}(\hat{\theta}_o; \psi_\gamma, h_\gamma) = \gamma\}$. It can be checked that

$$BF_{0:SM}(\hat{\theta}_o; \psi_\gamma, h_\gamma) = \{\psi_\gamma + (1 - \psi_\gamma) BF_{S:0}(\hat{\theta}_o; h_\gamma)\}^{-1},$$

where $BF_{S:0}(\hat{\theta}_o; h_\gamma)$ is the reciprocal of $BF_{0:S}(\hat{\theta}_o; h_\gamma)$ defined in (5) with $g_\gamma = h_\gamma$.

Family (14) is empty if condition (15) cannot be fulfilled.

It is worth emphasizing that, differently from the skeptical prior, our skeptical mixture comprises a *family* of distributions, which includes the skeptical prior (4) as a special case by setting $(\psi_\gamma = 0, h_\gamma = g_\gamma)$.

2.1 Prior-data conflict under the skeptical mixture prior

Consider $\hat{\theta}_o | \theta \sim N(\theta, \sigma_o^2)$, with σ_o^2 known, and assume that θ is distributed according to the skeptical mixture prior at level γ, (14). The marginal density of the estimator $\hat{\theta}_o$ is

$$m(\hat{\theta}_o) = \int N(\hat{\theta}_o | \theta, \sigma_o^2) dF_{SM}(\theta),$$

where $N(\hat{\theta}_o | \theta, \sigma_o^2)$ is a shorthand notation for the sampling density of $\hat{\theta}_o$ and $F_{SM}(\theta)$ the cdf of the mixture prior (14). We obtain

$$m(\hat{\theta}_o) = \psi_\gamma N(\hat{\theta}_o | 0, \sigma_o^2) + (1 - \psi_\gamma) N(\hat{\theta}_o | 0, \sigma_o^2 (1 + h_\gamma)). \tag{16}$$

Simplifying the notation, the structure of Equation (16) can be formally written as

$$m_T(t) = \psi N(t | 0, \sigma^2) + (1 - \psi) N(t | 0, \sigma^2 (1 + h)). \tag{17}$$

To evaluate the p-value for prior-data conflict $P(t_{obs})$ defined in (13) with regard to (17), it is expedient to introduce an auxiliary random variable V having a Bern(ψ) distribution and define the joint density of (T, V) as $h(t, v|\theta) = f(t|v, \theta)g(v)$ where $g(0) = \psi$, $g(1) = (1 - \psi)$ and

$$f(t|v, \theta) = \begin{cases} N(t|0, \sigma^2) & \text{if} \quad v = 0 \\ N(t|\theta, \sigma^2) & \text{if} \quad v = 1. \end{cases}$$

Let $\theta \sim N(\theta|0, \sigma^2 \cdot h)$. Then, marginally

$$\begin{aligned} h(t) &= \sum_v \left\{ \int h(t, v|\theta) N(\theta|0, \sigma^2 \cdot h) d\theta \right\} g(v) \\ &= \psi \int N(t|0, \sigma^2) p(\theta) d\theta + (1 - \psi) \int N(t|\theta, \sigma^2) p(\theta) d\theta \\ &= \psi N(t; 0, \sigma^2) + (1 - \psi) N(t; 0, \sigma^2 \cdot (1 + h)), \end{aligned}$$

which coincides with (17).

Since V is ancillary, one can condition on it to compute prior-data conflict; see Evans and Moshonov (2006). Hence,

$$\begin{aligned} \mathrm{P}(t_{obs}|v = 0) &= \Pr\left\{ N(T|0, \sigma^2) \le N(t_{obs}|0, \sigma^2) \right\} \\ \mathrm{P}(t_{obs}|v = 1) &= \Pr\left\{ N(T|0, \sigma^2(1 + h)) \le N(t_{obs}|0, \sigma^2)(1 + h) \right\}, \end{aligned}$$

whence

$$\mathrm{P}(t_{obs}) = \psi \mathrm{P}(t_{obs}|v = 0) + (1 - \psi)\mathrm{P}(t_{obs}|v = 1). \tag{18}$$

Lemma *Let* $T \sim f(t) = N(t|0, \tau^2)$. *Then*

$$Pr\{f(T) \le f(t_{obs})\} = Pr\left\{ U \ge \left(\frac{t_{obs}}{\tau} \right)^2 \right\},$$

where $U \sim \chi^2(1)$, *a chi-squared distribution with one df.*

Proof $f(T) \le f(t_{obs})$ iff $\left(\frac{T}{\tau}\right)^2 \ge \left(\frac{t_{obs}}{\tau}\right)^2$, and $(T/\tau)^2 \equiv U \sim \chi^2(1)$. □

Using the lemma together with (18) and reverting to the notation used in (16), the p-value for prior-data conflict based on the skeptical mixture prior (14) is

$$P(\hat{\theta}_o; \psi_\gamma, h_\gamma) = \psi_\gamma(1 - G_1(z_o^2)) + (1 - \psi_\gamma)(1 - G_1(z_o^2/(1 + h_\gamma))), \tag{19}$$

where $G_1(\cdot)$ is the cdf of a chi-squared distribution with one df.

Since any element in the *set* U_γ of hyperparameters $\{(\psi_\gamma, h_\gamma)\}$ describing the family (14) leads to a BF equal to γ, a skeptic is offered the opportunity to identify

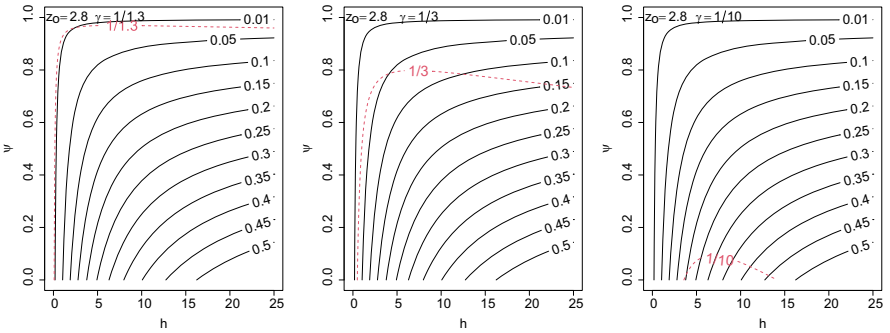

Fig. 1 α-contours of p-values for prior-data conflict (solid black line). Contour for $BF_{0:SM}(\hat{\theta}_o; \psi, h) = \gamma$ (dashed red line), for $z_o = 2.8$ and three selected values of γ. The intersection between the p-value at level α and the $BF_{0:SM}$ at level γ yields the pair of solutions $(\psi_{\gamma,\alpha}, h_{\gamma,\alpha})$ (color figure online)

a specific pair $(\psi_{\gamma,\alpha}, h_{\gamma,\alpha})$ by adding the additional constraint that the prior-data conflict is equal to α. As for ordinary p-values, very small values of α suggest that the observed $\hat{\theta}_o$ is highly unlikely to occur under the skeptical mixture prior. Conventional thresholds for declaring prior-data conflict are $\alpha = \{0.10, 0.05, 0.01\}$ where a lower value represents a stronger conflict.

The task of identifying the pair $(\psi_{\gamma,\alpha}, h_{\gamma,\alpha})$ can be visually represented by plotting in the (h, ψ) space: i) the contour lines realizing $P(\hat{\theta}_o; \psi, h) = \alpha$ for a grid of α-values, where $P(\hat{\theta}_o; \psi, h)$ is identical to the expression in (19) except that (h, ψ) are *unconstrained*; ii) the contour line U_γ for a given γ, and finally looking for possible points of intersection.

Figure 1 illustrates our procedure with $z_o = 2.8$, three levels of γ and a grid of values for α. Notice that both the γ and α-contours are concave. Additionally the α-contours are increasing over the h-range considered in the plot. To see why this occurs, simply observe that if we raise the mass ψ and keep h fixed (or set it to a lower value), this will reduce the area in the tails of $m_T(\cdot)$; see (12). Accordingly, to keep α constant, both h and ψ must jointly increase. Finally, as α increases, the α-contour allows only values $h > h_\alpha$ with h_α increasing in α.

We now consider the three γ-contours in the three panels of Figure 1. To understand the difference, we argue as follows. First of all the value $z_o = 2.8$ represents a substantive effect size: in a frequentist setting it would be regarded as highly significant (two-sided p-value ≈ 0.005). Consider the left panel, characterized by a low level of skepticism with $\gamma = 1/1.3$ sufficiently high to represent merely "anecdotal" evidence against H_0. To achieve this relatively high level of γ, the prior must be essentially concentrated on $\theta = 0$, implying an extremely strong prior belief in favor of the null hypothesis. This situation can be obtained by letting h be negligibly different from zero, so that the two components of the skeptical mixture, namely the Dirac measure and the Normal distribution, are essentially indistinguishable and produce a unitary mass on zero, regardless of the value of ψ. This explains the almost vertical part of the $\gamma = 1/1.3$-contour. Non-negligible h values start being allowed only for extremely high values of ψ, and hereafter h can increase further as ψ monotonically declines,

because the Jeffreys–Lindley paradox starts kicking-in. Such priors will determine an exceptionally strong level of prior-data conflict ($\alpha < 0.01$) for most (h, ψ) values on the vertical part, reaching the level $\alpha = 0.01$ only when h starts getting bigger. The level $\alpha = 0.05$ is not even reached over the range of h values considered in the plot, although it will be eventually intersected as h is allowed to increase.

The central panel presents a less extreme scenario. Here the level of skepticism against H_0 is increased because γ is lowered to $1/3$, bordering between anecdotal and moderate evidence. Accordingly, the vertical part of the $\gamma = 1/3$-contour is less steep; in particular, ψ is never above the 0.8 threshold, and correspondingly the level of prior-data conflict is always above 0.01 with intersections available at $\alpha = 0.05$ and 0.10 in the range of h presented in the plot.

The right plot further increases the level of skepticism to the even smaller value $\gamma = 1/10$, now bordering between moderate and strong evidence against H_0. In this scenario, ψ can never exceed 0.1 while h ranges between 4 and 13 approximately. In this scenario prior-data conflict is very mild with α always above 0.2.

We conclude by remarking that the ordinary skeptical prior is represented by the $(h_\gamma, \psi_\gamma = 0)$ point on the corresponding $\gamma = BF_S$-contour. For $BF_S = 1/1.3 \approx 0.77$, h is approximately 0.08

with a very high level of prior-data conflict ($\alpha < 0.01$). For $BF_S = 1/3$ $h \approx 0.5$ with moderately high prior-data conflict ($0.01 < \alpha < 0.05$). Finally, in the right scenario, h is close to 4 with $\alpha = 0.2$, suggesting no prior-data conflict.

2.2 The skeptical mixture Bayes factor and its relationship with the skeptical Bayes factor

Consider now the comparison

$$H_{SM} : \theta \sim \psi_{\gamma,\alpha}\delta_0 + (1 - \psi_{\gamma,\alpha})N(0, h_{\gamma,\alpha}\sigma_o^2) \quad \text{versus} \quad H_A : \theta \sim N(\hat{\theta}_o, \sigma_o^2), \tag{20}$$

where H_{SM} represents the skeptical mixture prior with γ level of skepticism and α level of conflict, while H_A is the advocate prior. Let $p_A(\cdot)$ be the density function of the advocate prior and let $f(\hat{\theta}_r \mid H_A) = \int f(\hat{\theta}_r \mid \theta) p_A(\theta) d\theta$ denote the marginal density of $\hat{\theta}_r$ conditionally on H_A. Similarly let $p_S(\theta; h_{\gamma,\alpha}) = N(\theta; 0, h_{\gamma,\alpha}\sigma_o^2)$ be the density function of the continuous component of the skeptical mixture prior and let $f(\hat{\theta}_r \mid H_S, h_{\gamma,\alpha}) = \int f(\hat{\theta}_r \mid \theta) p_S(\theta; h_{\gamma,\alpha}) d\theta$ denote the marginal density of $\hat{\theta}_r$ conditionally on H_S with given $h_{\gamma,\alpha}$, as in (4). Finally, let $P_{SM}(\cdot)$ be the cdf of the skeptical mixture prior (20), which is everywhere continuous, except in $\theta = 0$, where it makes a jump equal to $\psi_{\gamma,\alpha}$.

Then

$$\begin{aligned} BF_{SM:A}(\hat{\theta}_r; \psi_{\gamma,\alpha}, h_{\gamma,\alpha}) &= \frac{\int f(\hat{\theta}_r \mid \theta) d P_{SM}(\theta)}{f(\hat{\theta}_r \mid H_A)} \\ &= \frac{1}{f(\hat{\theta}_r \mid H_A)} \times \Big(\psi_{\gamma,\alpha} f(\hat{\theta}_r \mid \theta = 0) + (1 - \psi_{\gamma,\alpha}) f(\hat{\theta}_r \mid H_S) \Big) \\ &= \psi_{\gamma,\alpha} BF_R + (1 - \psi_{\gamma,\alpha}) BF_{S:A}(\hat{\theta}_r; h_{\gamma,\alpha}), \end{aligned} \tag{21}$$

where BF_R is the replication BF defined in (3), and $BF_{S:A}$ is the BF comparing the skeptical and the advocate prior defined in (6) having relative variance $h_{\gamma,\alpha}$, which, differently from g_γ in (6), incorporates the prior-data conflict constraint. We then declare *replication success* at level γ iff

$$BF_{SM:A}(\hat{\theta}_r; \psi_{\gamma,\alpha}, h_{\gamma,\alpha}) \leq \gamma, \tag{22}$$

that is, the data favor the advocate over the skeptical mixture prior at a higher level than the skeptic's initial objection. The lower this value, the stronger the claim of replication success.

We remark that definition (22) depends on the threshold γ, as well as α. In analogy with equation (8), the *skeptical mixture Bayes factor* is defined as

$$BF_{SM}(\alpha) = \inf\{\gamma : BF_{SM:A}(\hat{\theta}_r; \psi_{\gamma,\alpha}, h_{\gamma,\alpha}) \leq \gamma\}. \tag{23}$$

As for the skeptical Bayes factor, it may also happen that $BF_{SM:A}(\hat{\theta}_r; \psi_{\gamma,\alpha}, h_{\gamma,\alpha})$ remains below $BF_{0:SM}(\hat{\theta}_o; \psi_{\gamma,\alpha}, h_{\gamma,\alpha})$ for all values of the hyperparameters. In such situations we set $BF_{SM}(\alpha)$ to be equal to the minimum value taken on by $BF_{0:SM}(\hat{\theta}_o; \psi_{\gamma,\alpha}, h_{\gamma,\alpha})$. Finally, there could be situations wherein replication success cannot be established for any level of γ and α, so that $BF_{SM}(\alpha)$ does not exist. This means that the replication study is unsuccessful since it is impossible for the advocate to convince the skeptic at any level of evidence.

The following represents an important feature of our proposal.

Result 1 *Under the skeptical mixture prior introduced in (20), if $\psi_{\gamma,\alpha} > 0$ and the true value is $\theta^* = 0$, then $BF_{SM:A}(\hat{\theta}_r; \psi_{\gamma,\alpha}, h_{\gamma,\alpha})$ is consistent.*

Proof For $n_r \to \infty$, the result follows immediately from (21) and the fact that $BF_R \to \infty$, if $\theta^* = 0$ because of (9), while $BF_{S:A}(\hat{\theta}_r; h_{\gamma,\alpha})$ converges to a constant because of (10). □

Thus, if the effect is truly absent, this will be flagged by $BF_{SM:A}$ with unlimited evidence if the sample size grows indefinitely. On the other hand if $\theta^* \neq 0$, then the continuous skeptical component of the mixture will take the lead, and $BF_{SM:A}$ will converge to the constant $(1 - \psi_{\gamma,\alpha})\frac{p_S(\theta^*; h_{\gamma,\alpha})}{p_A(\theta^*)}$; see Proposition 2. While this result is only partial, it is particularly useful in a replication setting wherein correctly ascertaining the lack of an effect may prove very valuable to contrast an original finding possibly pointing in a different direction.

Our next result explores some relationships between the skeptical and the skeptical mixture Bayes factor. Essentially, it states that if the α-level of $BF_{SM}(\alpha)$ is larger than the corresponding value for BF_S, then its relative variance must also be larger than the corresponding value for BF_S; and *vice versa*.

Proposition 3 *Let $\mathcal{A}_{SM} = \{\alpha : BF_{SM}(\alpha) \text{ exists}\}$. For $\alpha \in \mathcal{A}_{SM}$ denote with $h_{SM,\alpha}$ the corresponding relative variance in the skeptical mixture prior realizing $BF_{SM}(\alpha)$.*

Assume the skeptical Bayes factor BF_S in (8) exists and let $g_S > 0$ be its corresponding relative variance and α_S its corresponding level of prior-data conflict.

If $c = \sigma_o^2/\sigma_r^2 = 1$ and $\alpha \in \mathcal{A}_{SM}$ the following two conditions are equivalent

C1 $\alpha \geq \alpha_S$;

C2 $h_{SM,\alpha} \geq g_S$.

Proof See supplementary material.

Proposition 4 *Assume the skeptical Bayes factor BF_S exists. If $c = \sigma_o^2/\sigma_r^2 = 1$ and $d = 1$, then $BF_S = 2^{1/4} \cdot \exp\left\{-\frac{z_o^2}{4}\right\}$.*

Proof See supplementary material.

2.3 Example

By way of illustration, consider the same setting discussed in Pawel and Held (2022 Sect. 2.2) with $z_o = 3$, $z_r = 2.5$ and $c = \sigma_o^2/\sigma_r^2 = 1$, so that $d = \hat{\theta}_r/\hat{\theta}_o = 0.83$. This setup is meant to represent a situation often encountered in practice with the replication study providing a somewhat weaker evidence against the null than the original study. Additionally, we fix the p-value for prior-data conflict at level $\alpha = 0.1$.

In Figure 2 we see four curves. Two, namely $BF_{0:S}$ (solid dark brown) and $BF_{0:SM}$ (dashed light brown), are based on the original data, while $BF_{S:A}$ (solid dark blue) and $BF_{SM:A}$ (dashed light blue) refer to replication data.

All curves are plotted as a function of their relative variance. Additionally all skeptical mixture priors realize a p-value for prior-data conflict equal to $\alpha = 0.1$. The replication Bayes factor BF_R is also included and appears as a constant green line because its corresponding prior has no hyperparameters. The black cross represents BF_S, i.e., the skeptical BF, while the green one represents $BF_{SM}(\alpha)$, the skeptical mixture BF.

The solid brown curve $BF_{0:S}(\hat{\theta}_o)$ exhibits the usual pattern, decreasing up to a certain point and then increasing. On the other hand, the pattern of the dashed brown curve $BF_{0:SM}(\hat{\theta}_o)$ is quite distinct. First of all, it exists only for h larger than a threshold, $h(z_o, \alpha)$ say; next it is monotonically increasing over the range of values considered in the plot. To understand the first point, set $\psi = 0$ in the expression of the p-value for prior-data conflict (19) and derive that when $\alpha = 0.1$: the curve must start at $h = h(z_o, \alpha) = \frac{z_o^2}{G_1^{-1}(1-\alpha)} - 1 = \frac{3^2}{G_1^{-1}(0.9)} - 1 \approx 2.33$. To further understand the behavior of the curves, recall that $z_o = 3$ is a result which exhibits appreciable evidence against the null. Consider $BF_{0:S}(\hat{\theta}_o)$ first. The curve is first monotonically decreasing and then increasing. This happens because, as the relative variance increases, it will push mass toward areas in the θ-space better supported by the data, and thus against the null, and this will reduce the Bayes factor in favor of H_0. However, the curve will start increasing when the relative variance becomes too high pushing mass to the tails of the θ-space, thus causing H_0 to gain evidence in comparison with H_{SM} (Jeffreys–Lindley paradox).

Now turn to $BF_{0:SM}(\hat{\theta}_o)$. The main difference with the skeptical $BF_{0:S}(\hat{\theta}_o)$ is that the curve is always monotonically increasing. The reason why this occurs is because

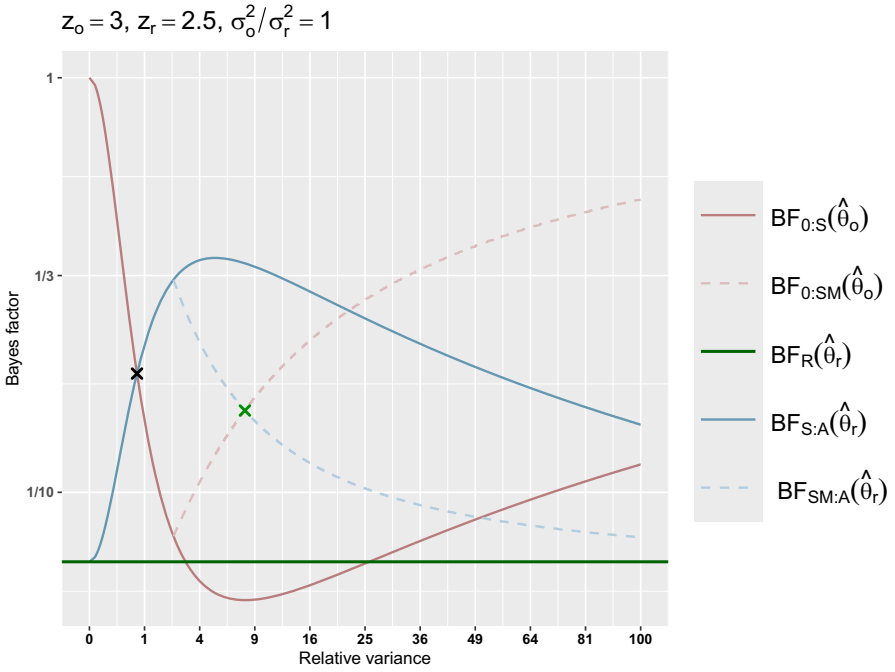

Fig. 2 Bayes factors $BF_{0:S}(\hat{\theta}_o; g)$, $BF_{S:A}(\hat{\theta}_r; g)$, $BF_{0:SM}(\hat{\theta}_o; \psi, h)$, $BF_{SM:A}(\hat{\theta}_r; \psi, h)$ and $BF_R(\hat{\theta}_r)$ as a function of the relative variance. The black cross represents the skeptical Bayes factor BF_S, while the green one represents the skeptical mixture Bayes factor $BF_{SM}(\alpha)$, with $\alpha = 0.1$ (color figure online)

the skeptical mixture incorporates a constraint on prior-data conflict which is absent in the skeptical prior. Specifically, pairs (h, ψ) on the same α-contour level are positively related; see Figure 1. This implies that, as the relative variance increases, so does ψ. Alternatively said, the effect of increasing the relative variance in the skeptical mixture is now counterbalanced by ψ.

Now let us turn to the curve $BF_{S:A}(\hat{\theta}_r)$. In this case $z_r = 2.5$, so that evidence is still against the null, although to a lesser extent than in the original experiment. As the relative variance increases, we know that mass is pulled away from values around zero which are not supported by the replication data, and this will initially benefit the skeptical hypothesis although this phenomenon will gradually reverse as the variance gets too large. This explains why the plot of $BF_{S:A}(\hat{\theta}_r)$ almost mirrors that for $BF_{0:S}(\hat{\theta}_o)$. Similar considerations apply to the behavior of $BF_{SM:A}(\hat{\theta}_r)$ in comparison to $BF_{0:SM}(\hat{\theta}_o)$.

We now investigate the role played by prior-data conflict on the skeptical mixture Bayes factor $BF_{SM}(\alpha)$ defined in (23). Because of the presence of two hyperparameters and the prior-data constraint embedded in our prior, our methodology is more complex than in the standard skeptical approach with only a single hyperparameter. As a consequence, fewer analytical results are available. Nevertheless we performed extensive numerical investigations to highlight some important features of our methodology. A summary of the results is reported in Figure 3 with special emphasis on the

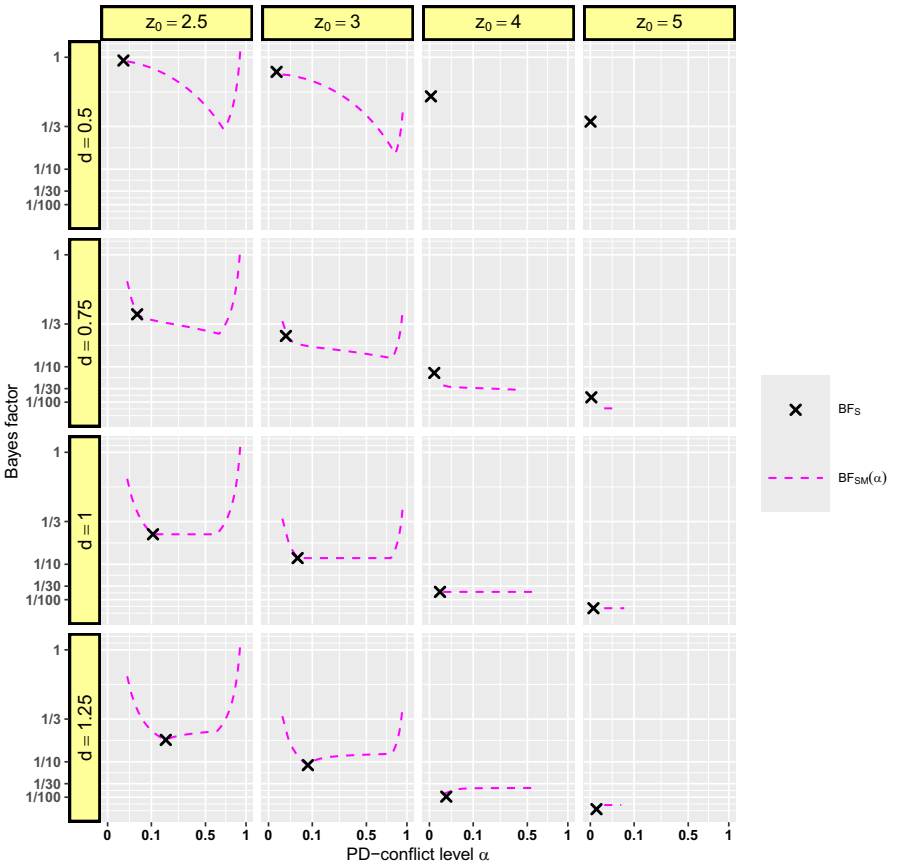

Fig. 3 Skeptical and skeptical mixture Bayes factors BF_S and $BF_{SM}(\alpha)$ for varying z_o and d as functions of the prior-data conflict threshold α. In all examples $c = \sigma_o^2/\sigma_r^2 = 1$. $z_o = \hat{\theta}_o/\sigma_o$ represents the z-value associated to the estimate $\hat{\theta}_o$ of the effect under the original study, whereas $d = \hat{\theta}_r/\hat{\theta}_o$ denotes the relative effect estimate

behavior of $BF_{SM}(\alpha)$ as a function of α. We have identified sixteen scenarios stemming from the combination of four values for z_o and four values of d. To better isolate their role we assumed throughout $c = \sigma_o^2/\sigma_r^2 = 1$ so that the original and replication study present a comparable level of precision in their estimation of the effect. Specifically, we let $z_o \in \{2.5, 3, 4, 5\}$, representing levels of increasing evidence against the null hypothesis in the original study, and $d = \hat{\theta}_r/\hat{\theta}_o \in \{0.5, 0.75, 1, 1.25\}$, identifying different replicability ratios with $d = 1$ representing a benchmark in which the original and replication effect estimates are identical.

Each panel reports the corresponding plot of $BF_{SM}(\alpha)$, whenever it exists, against α. The skeptical Bayes factor BF_S is marked by a black cross in correspondence of the realized level of prior-data conflict attained by the skeptical prior.

The behavior of BF_S appears quite natural: for a given d it declines as z_o increases, that is the level of skepticism increases as the original data provide stronger evidence

against H_0. A similar shape occurs for each given z_o as d grows, in this case because the replication data provide stronger evidence against H_0. As can be expected, the prior-data conflict realized under the skeptical prior increases (α decreases) as z_o becomes larger and becomes appreciable when $z_o \geq 3$, with α never exceeding the level 0.1, suggesting some incompatibility of the skeptical prior with respect to the original data.

The behavior of the $BF_{SM}(\alpha)$ curves can be separated into two parts depending on the value of z_o.

Specifically, for $z_o \in \{4, 5\}$, either $BF_{SM}(\alpha)$ does not exist ($d = 0.5$), or it is slightly declining ($d = 0.75$), essentially constant ($d = 1$) or slightly increasing ($d = 1.25$). Thus, when the evidence against the null hypothesis is highly substantial—the two-sided p-value in these situations is of the order of 10^{-5} or lower—the skeptical mixture BF is in the neighborhood of 1/30 and can approach 1/100 representing strong, respectively very strong, evidence against the null according to conventional classification schemes for the Bayes factor; see, e.g., (Schönbrodt and Wagenmakers 2018, Table 1). We note that these values are also similar to the standard skeptical BF_S. In conclusion, when $c = 1$ and $z_o \in \{4, 5\}$ replication success can be declared at a strong/very strong level, and this conclusion is robust to the choice of the prior-data conflict level α.

When $z_o \in \{2.5, 3\}$, pointing to a relatively weaker effect, the curve takes on a bathtub shape for d exceeding 0.75, while for $d = 0.5$ its form is more akin to a wedge. Either way, the left arm of the curve is downward sloping, followed by a gently declining or constant floor and then by a rising right arm (when $d = 0.5$ there is no floor in the curve). Interestingly, both arms in the bathtub curves correspond to situations wherein the $BF_{SM:A}(\hat{\theta}_r; \psi_{\gamma,\alpha}, h_{\gamma,\alpha})$ is always less than $BF_{0:SM}(\hat{\theta}_o; \psi_{\gamma,\alpha}, h_{\gamma,\alpha})$ so that $BF_{SM}(\alpha)$ is chosen to be the minimum value of $BF_{0:SM}(\hat{\theta}_o; \psi_{\gamma,\alpha}, h_{\gamma,\alpha})$; see Equation (23). Additionally the upward sloping right arm is due to the Jeffreys–Lindley paradox because increasing levels of α require increasing levels of the relative variances as already described in Figure 1. More specifically, for $z_o = 2.5$, replication success can be established only with a level of evidence which is only anecdotal because $1/3 < BF_{SM}(\alpha) < 1$. On the other hand for higher values of d, the evidence level is either anecdotal (usually in the two arms) or moderate because $BF_{SM}(\alpha)$ goes below the level $1/3$ along the floor of the curve. It is worth pointing out that the anecdotal level is reached only for values of α in the neighborhood of either endpoint of the interval $(0, 1)$. One could possibly argue that for more reasonable values of prior-data conflict (e.g., $\alpha = 0.1$) replication success can be established at a moderate level. Somewhat reassuringly this conclusion agrees with that based on the standard BF_S; interestingly, however, this is not the case for $d = 0, 75$, where the analysis based on the BF_S would declare success only at the anecdotal level, while our analysis would upgrade this conclusion to moderate and with a more reasonable value for α. For $z_o = 3$ results are highly robust for $d \geq 0.75$ because replication success can be declared at the moderate level and in fact close to the strong level when the curve approaches the value $1/10$ for the intermediate values of α.

In conclusion, Figure 3 shows across several concrete scenarios that the replication assessment is *robust* with respect to the choice of α. From this perspective it represents

a valuable tool for the practitioner, who can run the procedure using their available data.

3 Case studies

In this section we consider real data sets from the *Social Sciences Replication Project* (SSRP) (Camerer et al. 2018). In 2016 SSRP planned to replicate a collection of experimental studies in the social sciences published in two high-profile journals, *Nature* and *Science*, in the period 2010-2015. Specifically, 21 studies were selected because they satisfied three criteria: (1) they tested for an experimental treatment effect between or within subjects, (2) they tested at least one clear hypothesis with a statistically significant finding, and (3) they were performed on students or other accessible subject pools. For each study, further actions were implemented to determine which treatment effect to consider for replication. Additionally, to deal with the possibility of inflated effect sizes in the original studies, the authors adopted a high-powered design and a two-stage procedure to implement the replication experiment; details are spelled out in their report (Camerer et al. 2018). Eventually, the authors discovered that only for 13 studies, out of 21, there was a significant effect in the same direction as in the original experiment (Camerer et al. 2018, Figure 1b). Interestingly, it is only for them that the skeptical BF_S is well defined, as reported in Pawel and Held (2022 Section 5 and Table 2). In this section we analyze these 13 studies using our skeptical mixture approach. We used the observations reported in the `SSRP` dataset contained in the `R` package `ReplicationSuccess` (Held 2020).

Effect estimates for each study were originally reported on the correlation scale r. The Fisher z-transformation was then applied to obtain an approximate Normal distribution for the estimator $\hat{\theta} = tanh^{-1}(r)$, having an approximate variance $Var(\hat{\theta}) = 1/(n-3)$, so that $c \approx n_r/n_o$.

For each study, we report in Table 1 the basic summary statistics $(z_o, z_r, n_o, n_r, c, d)$ in columns 1 through 6. Next we report the hyperparameter of the skeptical - respectively skeptical mixture- prior, namely g_S and $(\psi_{SM,\alpha}, h_{SM,\alpha})$ each computed in correspondence of the degrees of skepticism $\gamma_S = BF_S$ and $\gamma_{SM} = BF_{SM}(\alpha)$; see Equations (8) and (23). Additionally, P_S denotes the realized p-value for prior-data conflict: note that P_{SM}, if it exists, is equal by definition to the value $\alpha \in \{0.05, 0.10\}$ in each block, see Equation (19). The last three columns report the three Bayes factors, namely the skeptical BF_S, the replication BF_R and the skeptical mixture $BF_{SM}(\alpha)$ (when it exits; otherwise the corresponding entry is void). For completeness, the BF curves for each of the 13 studies, and separately for $\alpha = \{0.05, 0.1\}$, are plotted in Figures II and III of the supplementary material, whose format is the same as Figure 2 in Section 2.3. The choice of the two values for α is done for illustrative purposes only: $\alpha = 0.05$ corresponds to a significant level of prior-data conflict, while $\alpha = 0.1$ indicates a weakly significant level, similarly to what happens for the interpretation of p-values. In general, a more complete analysis of the case studies could be done across several values of α to assess sensitivity; see Figure 3.

We now summarize the main features which emerge.

Overall the value of $BF_{SM}(\alpha)$, when it exists, is generally less than or equal to BF_S. The only exceptions are represented by the studies Kovacs et al. (Kovács et al. 2010) and Morewedge et al. (Morewedge et al. 2010) (only for $\alpha = 0.05$). Notice, however, that for these two studies the curve of $BF_{S:A}(\hat{\theta}_r)$ always lies below the curve of $BF_{0:SM}(\hat{\theta}_o)$ as reported in Figures II and III of the supplementary material. In this case the value assigned to $BF_{SM}(\alpha)$ is by default the minimum value taken on by $BF_{0:SM}(\hat{\theta}_o)$; see (23). While for Morewedge et al. the difference between the two skeptical BF's is numerically trivial, this is not the case for the study by Kovacs et al., although the interpretation is essentially unchanged because both BF's belong to the "anecdotal" evidence range. As a consequence, under our mixture methodology, replication success is established with a stronger level of evidence (higher skepticism) than under the standard skeptical approach; see also Figure 3 for similar situations.

From a more practical perspective, using a conventional classification for the Bayes factor (Schönbrodt and Wagenmakers 2018), we note that replication success is robust to the choice of α for the following cases (in bracket the level of evidence which is also shared by BF_S): Balafoutas and Sutter (Balafoutas and Sutter 2012) (anecdotal); Duncan et al. (Duncan et al. 2012) (moderate); Gneezy et al. (Gneezy et al. 2014) (moderate); Kovacs et al. (anecdotal), Pyc and Rawson (Pyc and Rawson 2010) (anecdotal), Morewedge et al (anecdotal); Wilson et al. (Wilson et al. 2014) (very strong).

For some cases, namely Aviezer et al. (Aviezer et al. 2012), Hauser et al. (Hauser et al. 2014) and Janssen et al. (Janssen et al. 2010), $BF_{SM}(\alpha)$ does not exist for each of the considered α-values. On the other hand, if we are willing to tolerate an extremely high level of prior-data conflict, we will get a value for $BF_{SM}(\alpha)$, and this is typically similar to BF_S which exists for these studies. However, this happens at the expense of an extremely high prior-data conflict, so that these results should be considered with great caution because they are obtained under a very unreasonable prior. Finally, for three cases our skeptical mixture method reports replication success at a stronger level of evidence relative to the standard skeptical approach:

(reported in bracket): Derex et al. (Derex et al. 2013) (from moderate to strong); Karpicke and Blunt (Karpicke and Blunt 2011) (from moderate to strong); Nishi et al. (Nishi et al. 2015) (from anecdotal to moderate).

4 Discussion

In the context of replication studies, we introduced a method for quantifying the success of a replication experiment in reproducing the results of the original study. We used a meta-analytic framework with effect size estimators approximately normally distributed with known variances, which is often a reasonable assumption for large sample sizes, possibly after a suitable transformation.

Throughout we systematically used the Bayes factor (BF) as a measure of evidence, coupled with reverse-Bayes techniques to elicit a skeptical prior, along the lines originally presented in Pawel and Held (2022). We proposed a novel skeptical mixture prior which adds a component to regulate prior-data conflict in the prior.

This feature enhances the flexibility of our method and can be useful when the standard skeptical prior exhibits an extreme conflict with the original data. Instead of

imposing a fixed constraint on prior-data conflict, we perform a sensitivity analysis across a range of levels to gain deeper understanding and insight into the problem. Through a focused simulation study, we assess the sensitivity of our methodology to the level of prior-data conflict, and demonstrate its robustness across a variety of plausible scenarios. Reassuringly, our results show a broad agreement with those obtained using the standard skeptical approach although, in a few cases, we observe a meaningful shift in the degree of replication success.

Our skeptical mixture prior is characterized by two hyperparameters, ψ and h, which control distinct aspects of the prior: the probability mass assigned to the null value of the effect (zero in our setup) and the relative variance, respectively. By imposing a constraint α on prior-data conflict, the analysis effectively proceeds along α-contours in the (h, ψ)-plane, followed by a sensitivity analysis. An alternative and more direct approach would be to work with the unconstrained pair (h, ψ), thereby extending the framework proposed by Pawel and Held (2022), which relies solely on the relative variance. In this setting, the Bayes factor would naturally extend to a surface in three-dimensional space. We have briefly explored this extension separately, as it may represent an independent avenue for future research.

We conclude by noting that prior-data conflict was measured using the p-value proposed by Evans and Moshonov (2006). However, we emphasize that our framework can be employed with alternative measures of conflict, such as those presented in Reimherr et al. (2021) or in Young and Pettit (1996) and Veen et al. (2018).

Supplementary information

The supplementary material contains technical results and figures from the case studies. Specifically: the proofs of Propositions 1, 2, 3, and 4 and the proof of information consistency for the replication BF.

Part of the data appearing in Table 1 was downloaded from Samuel Pawel's GitHub page: https://github.com/SamCH93/ReplicationSuccess.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11749-025-00985-7.

Acknowledgements We express our thanks to Leonhard Held and Samuel Pawel (University of Zurich) for useful discussions on the issue of replication studies.

Author Contributions The authors contributed equally to this work.

Funding Open access funding provided by Università degli Studi di Trieste within the CRUI-CARE Agreement. Partial financial support for GC was provided by Università Cattolica del Sacro Cuore, Milan, projects D1 years 2022-2024.

Data Availability The authors used publicly available data from https://github.com/SamCH93/ReplicationSuccess

Code Availability R code required to reproduce the results of the paper can be found at https://github.com/LeoEgidi/Consonni_Egidi_Replication_Code.

Declarations

Conflict of interest: The authors report that there are no Conflict of interest to declare.

References

Anderson SF, Maxwell SE (2016) There?s more than one way to conduct a replication study: beyond statistical significance. Psych Meth 21(1):1–12

Aviezer H, Trope Y, Todorov A (2012) Body cues, not facial expressions, discriminate between intense positive and negative emotions. Science 338(6111):1225–1229

Balafoutas L, Sutter M (2012) Affirmative action policies promote women and do not harm efficiency in the laboratory. Science 335(6068):579–582

Bayarri MJ, Berger JO, Forte A, García-Donato G (2012) Criteria for Bayesian model choice with application to variable selection. Ann Stat 40(3):1550–1577

Camerer CF, Dreber A, Holzmeister F, Ho T-H, Huber J, Johannesson M, Kirchler M, Nave G, Nosek BA, Pfeiffer T et al (2018) Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. Nat Hum Behav 2(9):637–644

Dawid PA (2011) Posterior model probabilities. In: Bandyopadhyay PS, Forster M (eds) Philosophy of Statistics. Elsevier, Amsterdam, pp 607–630

Derex M, Beugin M-P, Godelle B, Raymond M (2013) Experimental evidence for the influence of group size on cultural complexity. Nature 503(7476):389–391

Duncan K, Sadanand A, Davachi L (2012) Memory?s penumbra: episodic memory decisions induce lingering mnemonic biases. Science 337(6093):485–487

Egidi L, Pauli F, Torelli N (2021) Avoiding prior-data conflict in regression models via mixture priors. Can J Stat 50(2):491–510

Evans M, Moshonov H (2006) Checking for prior-data conflict. Bayes Anal 1(4):893–914

Gneezy U, Keenan EA, Gneezy A (2014) Avoiding overhead aversion in charity. Science 346(6209):632–635

Good IJ (1950) Probability and the weighing of evidence. Philosophy 26(97):163–164

Harms C (2019) A Bayes factor for replications of ANOVA results. Am Stat 73(4):327–339

Hauser OP, Rand DG, Peysakhovich A, Nowak MA (2014) Cooperating with the future. Nature 511(7508):220–223

Hedges LV, Schauer JM (2019) Statistical analyses for studying replication: meta-analytic perspectives. Psych Meth 24(5):557–570

Held L (2020) A new standard for the analysis and design of replication studies. J R Stat Soc Ser A Stat Soc 183(2):431–448

Held L, Matthews R, Ott M, Pawel S (2022) Reverse-Bayes methods for evidence assessment and research synthesis. Res Synth Meth 13(3):295–314

Hutton JL, Diggle PJ, Bird SM, Hennig C, Longford N, Mathur MB, Vander Weele TJ, Ioannidis JPA, Chai CP, Dowe DL et al (2020) Discussion on the meeting on signs and sizes understanding and replicating statistical findings? J Roy 183(2):449–469

Janssen MA, Holahan R, Lee A, Ostrom E (2010) Lab experiments for the study of social-ecological systems. Science 328(5978):613–617

Jeffreys H (1961) Theory of probability, 3rd edn. University Press, Oxford

Johnson VE, Payne RD, Wang T, Asher A, Mandal S (2017) On the reproducibility of psychological science. J Am Stat Ass 112(517):1–10

Karpicke JD, Blunt JR (2011) Retrieval practice produces more learning than elaborative studying with concept mapping. Science 331(6018):772–775

Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Ass 90(430):773–795

Kovács ÁM, Téglás E, Endress AD (2010) The social sense: Susceptibility to others? beliefs in human infants and adults. Science 330(6012):1830–1834
Liang F, Paulo R, Molina G, Clyde MA, Berger JO (2008) Mixtures of g-priors for Bayesian variable selection. J Am Stat Ass 103(481):410–423
Ly A, Wagenmakers E-J (2022) Bayes factors for peri-null hypotheses. TEST 31(4):1121–1142
Ly A, Etz A, Marsman M, Wagenmakers E-J (2019) Replication Bayes factors from evidence updating. Behav Res Meth 51:2498–2508
Mitchell TJ, Beauchamp JJ (1988) Bayesian variable selection in linear regression. J Am Stat Ass 83(404):1023–1032
Morewedge CK, Huh YE, Vosgerau J (2010) Thought for food: imagined consumption reduces actual consumption. Science 330(6010):1530–1533
Nishi A, Shirado H, Rand DG, Christakis NA (2015) Inequality and visibility of wealth in experimental social networks. Nature 526(7573):426–429
O'Hagan A and Forster JJ (2004) Kendall's Advanced Theory of Statistics, Volume 2B: Bayesian Inference vol 2. Arnold, London
Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science 349(6251):4716
Pawel S, Held L (2022) The sceptical Bayes factor for the assessment of replication success. J R Stat Soc Ser B Stat Meth 84(3):879–911
Pyc MA, Rawson KA (2010) Why testing improves memory: mediator effectiveness hypothesis. Science 330(6002):335–335
Reimherr M, Meng X-L, Nicolae DL (2021) Prior sample size extensions for assessing prior impact and prior-likelihood discordance. J R Stat Soc Ser B Stat Meth 83(3):413
Schönbrodt FD, Wagenmakers E-J (2018) Bayes factor design analysis: planning for compelling evidence. Psych Bull Rev 25(1):128–142
Shafer G (1982) Lindley's paradox. J Am Stat Ass 77(378):325–334
Spiegelhalter DJ, Abrams KR, Myles JP (2003) Bayesian approaches to clinical trials and health-care evaluation. Wiley, Hoboken
Taylor S, Pollard K (2009) Hypothesis tests for point-mass mixture data with application to Omics data with many zero values. Stat Appl Gen Molec Biol 8(1):1–43
Veen D, Stoel D, Schalken N, Mulder K, Schoot R (2018) Using the data agreement criterion to rank experts? Beliefs. Entropy 20(8):592
Verhagen J, Wagenmakers E-J (2014) Bayesian tests to quantify the result of a replication attempt. J Exper Psych General 143(4):1457–1475
Wetzels R, Wagenmakers E-J (2012) A default Bayesian hypothesis test for correlations and partial correlations. Psych Bull Rev 19(6):1057–1064
Wilson TD, Reinhard DA, Westgate EC, Gilbert DT, Ellerbeck N, Hahn C, Brown CL, Shaked A (2014) Just think: the challenges of the disengaged mind. Science 345(6192):75–77
Young KDS, Pettit LI (1996) Measuring discordancy between prior and data. J R Stat Soc Ser B Stat Meth 58(4):679–689