Behavior Research Methods (2024) 56:4085–4102
https://doi.org/10.3758/s13428-024-02350-2

ORIGINAL MANUSCRIPT

Bayesian evidence synthesis as a flexible alternative to meta-analysis: A simulation study and empirical demonstration

Elise van Wonderen[1,2] · **Mariëlle Zondervan-Zwijnenburg**[2] · **Irene Klugkist**[2]

Accepted: 26 January 2024 / Published online: 26 March 2024

Abstract

Synthesizing results across multiple studies is a popular way to increase the robustness of scientific findings. The most well-known method for doing this is meta-analysis. However, because meta-analysis requires conceptually comparable effect sizes with the same statistical form, meta-analysis may not be possible when studies are highly diverse in terms of their research design, participant characteristics, or operationalization of key variables. In these situations, Bayesian evidence synthesis may constitute a flexible and feasible alternative, as this method combines studies at the hypothesis level rather than at the level of the effect size. This method therefore poses less constraints on the studies to be combined. In this study, we introduce Bayesian evidence synthesis and show through simulations when this method diverges from what would be expected in a meta-analysis to help researchers correctly interpret the synthesis results. As an empirical demonstration, we also apply Bayesian evidence synthesis to a published meta-analysis on statistical learning in people with and without developmental language disorder. We highlight the strengths and weaknesses of the proposed method and offer suggestions for future research.

Keywords Research synthesis · Bayes factor · Robustness · Conceptual replications · Experimental psychology · Informative hypotheses

Introduction

Science is by nature a cumulative endeavor, in which we build upon results from previous studies to inform theory and generate new hypotheses. However, a great challenge in building theories is the high prevalence of conflicting results in psychological research, caused in part by small sample sizes and reliance on null hypothesis significance testing (e.g., Button et al., 2013; Open Science Collaboration, 2015; Van Calster et al., 2018). A commonly used method to help remedy this issue is to increase the robustness of scientific findings by means of meta-analysis (e.g., Cooper et al., 2019; Lipsey & Wilson, 2001). In a meta-analysis, the researcher quantitatively summarizes results across studies by computing a weighted mean effect size and corresponding confidence intervals to investigate whether there is evidence for an effect when all studies are taken together. Meta-analysis therefore helps mitigate the issue of underpowered studies, as a small effect may be statistically non-significant in each individual study, but significant when all these studies are combined. Furthermore, it can be investigated whether there is significant heterogeneity in effect sizes across studies, and if so, which study-level variables can explain part of this variation (e.g., Berkey et al., 1995; van Houwelingen et al., 2002). This way, meta-analysis can help explain conflicting results in the literature and contribute to more coherent theories.

Although meta-analysis is a powerful and versatile method for aggregating multiple studies, conducting a meta-analysis is sometimes challenging or even impossible for the set of studies a researcher wishes to combine. As studies are combined at the level of the effect size, meta-analysis requires the effect sizes across studies to be conceptually comparable and have the same statistical form (Lipsey & Wilson, 2001). These requirements are unlikely to be met if studies differ considerably regarding research design, operationalization of key variables, and statistical models used. For example, when studies have measured variables

✉ Elise van Wonderen
e.vanwonderen@uva.nl

[1] Amsterdam Center for Language and Communication, University of Amsterdam, Spuistraat 134, Amsterdam 1012 VB, The Netherlands

[2] Department of Methodology & Statistics, Utrecht University, Utrecht, The Netherlands

on different scales (e.g., continuous vs. binary), transformations are needed to translate the effect sizes into one common effect-size metric. Although such transformations exist (see, e.g., Cooper et al., 2019; Lipsey & Wilson, 2001), they do not exist between all effect-size metrics and many of these transformations "make strong or even untenable assumptions" (van Assen et al., 2022, p. 1), begging the question of whether effect sizes that require such transformations should be combined in the first place. Furthermore, even small differences in design often result in different population effects being estimated, which means their effect sizes cannot be directly compared or aggregated (Morris & DeShon, 2002, and references cited therein). Finally, meta-analysis is impossible when studies have measured different parts of a larger overarching hypothesis. Such a situation is illustrated by Kevenaar et al. (2021), who investigated children's self-control ratings obtained from multiple informants across four different cohorts. The authors wanted to test the overarching hypothesis that children themselves report most problem behaviors, followed by their mothers and fathers, and that teachers report the fewest problems. However, in each cohort, ratings from only two or three (non-overlapping) informant groups were available, making it impossible to investigate the hypothesis of interest using meta-analysis.

In situations where meta-analysis is difficult or impossible, Bayesian evidence synthesis (BES) may provide a feasible and flexible alternative (Klugkist & Volker, 2023; Kuiper et al., 2013). As we will further elaborate below, BES consists of three steps. First, in each study, statistical hypotheses are formulated that reflect the theories of interest, but that incorporate data and design characteristics unique to that study. Then, Bayes factors are computed to quantify the evidence for the hypotheses in that study. Finally, the study-specific Bayes factors are aggregated to determine which hypothesis best accounts for each study's results when all studies are considered simultaneously. Crucially, BES poses less constraints on the studies to be combined than meta-analysis because studies are combined at the hypothesis level rather than at the level of the effect size. The effect sizes across studies therefore do not need to have the same statistical form. In addition, BES allows for differences in study design and operationalization of variables, as the study-specific hypotheses do not have to be identical. The key idea is that if the study-specific hypotheses test the same underlying (i.e., latent) effect, the Bayes factors for these hypotheses can be meaningfully combined. BES is thus a flexible tool that allows the aggregation of highly diverse studies. However, in contrast to meta-analysis, BES is solely concerned with hypothesis testing. It does not allow researchers to estimate the size of an effect or test whether there is systematic heterogeneity in effect sizes across studies. In addition, unlike meta-analysis, BES is not intended to increase the statistical power for detecting an effect (as will become clear in the remainder of this paper and is also explained in Klugkist and Volker, 2023). For these reasons, meta-analysis will be the preferred method when studies have similar designs and measures. However, hypotheses can often be tested in many different ways (e.g., experiments, tests, surveys, vignette studies) with many different models or parameter estimates (e.g., logistic regression, multilevel models, ANOVA). Hypotheses are often assumed to hold regardless of these (sometimes arbitrary) design choices, as long as all outcomes are considered to measure the same underlying (latent) construct and the sample is considered to be drawn from the population of interest. In such situations, BES can provide us with the global support for each hypothesis across all available studies. In addition, as we will further explain below, BES allows for (i) formulating and testing informative hypotheses that unlike the conventional null hypothesis can directly test a specific hypothesis, and (ii) evaluating multiple (i.e., 2+) hypotheses simultaneously, which allows researchers to directly compare all hypotheses of interest.

The goal of the current paper is to introduce BES as an alternative to meta-analysis when the latter is difficult or impossible and to assess how BES performs in comparison to meta-analysis under various conditions. This will help researchers who are familiar with meta-analysis to correctly evaluate BES results and to understand when and why these results may diverge from what would be expected in a meta-analysis. To investigate the performance of BES in comparison to meta-analysis, we conducted a Monte Carlo simulation study in which we mimicked various scenarios that may be relevant to applied researchers. To also provide readers with a real-world example, we included an empirical demonstration in which we applied BES to a published meta-analysis on statistical learning in people with and without developmental language disorder (Lammertink et al., 2017). Before we turn to the simulation study, we will first explain BES in more detail.

Bayesian evidence synthesis

As mentioned above, BES proceeds in three steps: (i) formulation of study-specific hypotheses, (ii) evaluation of these hypotheses in each study separately using Bayes factors, and (iii) the aggregation of these study-specific Bayes factors to yield the support for the overall theory over all studies combined. We now explain each step in turn.

Formulation of study-specific hypotheses

In the null hypothesis significance testing (NHST) framework, a null hypothesis (H_0: no effect) is tested against the complement hypothesis (not H_0). However, in the context of

BES, it is also possible to evaluate informative hypotheses that represent an explicit theory or expectation by posing constraints on model parameters (e.g., Hoijtink, 2012; Klugkist et al., 2011; van de Schoot et al., 2011). For instance, in an experimental design, specific conditions are included because it is *a priori* expected that in certain conditions participants will score higher or lower than in other conditions. This expectation can be represented by order constraints on the means and could, for example, lead to the informative hypothesis $H_i : \mu_1 < \mu_2 < \mu_3$ (where μ_j is the mean of the j^{th} group or condition). Finding support for H_i (or not) is then more informative than the evaluation of the usual null (all means equal) and complement (not all means equal). Order constraints are just one example of useful constraints to represent specific expectations. Also fitting in the framework of informative hypothesis evaluation are equality constraints (e.g., $\mu = 0.5$, or, $\mu_1 = \mu_2$), range constraints (e.g., $-0.1 < \mu < 0.1$, or, $\mu_1 > (\mu_2 + 0.5)$) and constraints on functions of parameters (e.g., $(\mu_1 + \mu_2)/2 > \mu_3$, or, $\mu_1 > 3\mu_2$). A special hypothesis is the unconstrained hypothesis which poses no constraints on the parameters (in the examples above, this means that all means can take on any value). As we explain below, this hypothesis is used in the computation of Bayes factors comparing two informative hypotheses. In addition, the unconstrained hypothesis is often included in the set of hypotheses under consideration to avoid choosing between different competing hypotheses when none of these hypotheses represent the data well (Hoijtink et al., 2019).

In the context of BES, it is furthermore possible to formulate study-specific hypotheses that test the overarching theory while also incorporating data characteristics and research methodology unique to that study. For example, say that two studies have investigated the effect of age on willingness to take risks where one study has measured the participant's age in years and the other study has divided the participants into three age groups. The overarching hypothesis is that age decreases the willingness to take risks. This theory can be translated into the study-specific hypothesis $H_{i,1} : \beta_{age} < 0$ in Study 1, where β_{age} is the regression estimate for the effect of age on willingness of taking risks; and into $H_{i,2} : \mu_1 > \mu_2 > \mu_3$ in Study 2, where μ_j is the mean willingness to take risks in the jth age group. Another example is provided by the study of Kevenaar et al. (2021) that was briefly mentioned above, where the authors wished to aggregate four cohort studies that provided ratings of children's self-control problems by different informants. One of the overarching hypotheses the authors wished to test was $H_i : \mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}$. However, as the cohort studies obtained ratings from different sets of informants, the authors specified three different study-specific hypotheses, namely $H_{i,1} : \mu_{mother} > \mu_{father} > \mu_{teacher}$; $H_{i,2} : \mu_{self} > \mu_{mother}$; and $H_{i,3} : \mu_{mother} > \mu_{teacher}$. The idea here is that these study-specific hypotheses should receive the most support in each study if H_i is true because these hypotheses are all compatible with H_i even though they only test part of it.

Hypothesis evaluation using the Bayes factor

The relative support for a given hypothesis (or model) can be expressed with the Bayes factor (BF). Note that in the Bayesian testing framework, a hypothesis is formulated as a statistical model; throughout the remainder of this paper we will therefore use the terms *hypothesis* and *model* interchangeably. The BF comparing hypotheses H_i and $H_{i'}$ is given by

$$BF_{ii'} = \frac{P(\boldsymbol{X}|H_i)}{P(\boldsymbol{X}|H_{i'})} = \frac{\int P\left(\boldsymbol{X}|\boldsymbol{\beta}, H_i\right) P\left(\boldsymbol{\beta}|H_i\right) \partial \boldsymbol{\beta}}{\int P\left(\boldsymbol{X}|\boldsymbol{\beta}, H_{i'}\right) P\left(\boldsymbol{\beta}|H_{i'}\right) \partial \boldsymbol{\beta}}, \tag{1}$$

where $P(\boldsymbol{X}|H_i)$ and $P(\boldsymbol{X}|H_{i'})$ denote the marginal likelihood of the observed data under hypothesis H_i and $H_{i'}$, respectively (Kass & Raftery, 1995). These marginal likelihoods are defined as the product of the likelihood function, $P(\boldsymbol{X}|\boldsymbol{\beta}, H)$, and the prior, $P(\boldsymbol{\beta}|H)$, integrated with respect to the parameter vector $\boldsymbol{\beta}$. The BF can be directly interpreted as the evidence in the data for hypothesis H_i versus the evidence in the data for hypothesis $H_{i'}$. As such, a $BF_{ii'} = 10$ for example indicates that hypothesis H_i receives 10 times more support than hypothesis $H_{i'}$.

Calculation of the BF based on its mathematical definition presented in Eq. 1 is typically difficult. However, building on work by Klugkist et al. (2005), Gu et al. (2018) showed that the BF comparing a hypothesis H_i to the unconstrained hypothesis H_u can be approximated by the Savage-Dickey density ratio in Eq. 2 for equality-constrained hypotheses (e.g., $\beta = 0$) when (i) using normal approximations of the prior and posterior distributions of the unconstrained hypothesis, (ii) centering the prior distribution on the boundary of the hypotheses under consideration, and (iii) using a fraction b of the information in the data to construct a proper prior distribution. This yields

$$BF_{i_0 u} = \frac{f_{i_0}}{c_{i_0}} = \frac{P_u(\boldsymbol{\beta} = \boldsymbol{B}_{i_0}|\boldsymbol{X})}{p_u^*(\boldsymbol{\beta} = \boldsymbol{B}_{i_0}|\boldsymbol{X}^b)}, \tag{2}$$

where f_{i_0} is the density of the unconstrained posterior distribution $P_u(\boldsymbol{\beta} = \boldsymbol{B}_{i_0}|\boldsymbol{X})$ (denoted fit) and c_{i_0} is the density of the adjusted unconstrained prior distribution $p_u^*(\boldsymbol{\beta} = \boldsymbol{B}_{i_0}|\boldsymbol{X}^b)$ (denoted complexity) evaluated at the location of the hypothesized values $\boldsymbol{B}_{i_0}$.

For inequality-constrained hypotheses (e.g., $\beta > 0$), BF_{iu} can be approximated by

$$BF_{i_1u} = \frac{f_{i_1}}{c_{i_1}} = \frac{\int_{\boldsymbol{\beta}\in\boldsymbol{B}_{i_1}} P_u(\boldsymbol{\beta}|\boldsymbol{X})\partial\boldsymbol{\beta}}{\int_{\boldsymbol{\beta}\in\boldsymbol{B}_{i_1}} p_u^*(\boldsymbol{\beta}|\boldsymbol{X}^b)\partial\boldsymbol{\beta}}, \quad (3)$$

where the fit (f_{i_1}) is the proportion of the unconstrained posterior distribution that is in line with hypothesis H_i, and the complexity (c_{i_1}) is the proportion of the unconstrained prior distribution for H_u in line with hypothesis H_i. For the computation of BF_{iu} for hypotheses with both equality and inequality constraints see Gu et al. (2018, p. 241).

The BF comparing two informative hypotheses H_i and $H_{i'}$ is then given by

$$BF_{ii'} = \frac{BF_{iu}}{BF_{i'u}} = \frac{f_i/c_i}{f_{i'}/c_{i'}}. \quad (4)$$

The support expressed by the BF thus balances the fit and complexity of the hypotheses under consideration. The fit of a hypothesis is a measure of how well the hypothesis describes the observed data, while the complexity indicates how specific (parsimonious) the hypothesis is. The higher the fit, and the lower the complexity, the higher the BF in favor of the hypothesis at hand relative to an alternative hypothesis. This means that whenever two hypotheses have an equal fit, the most parsimonious hypothesis will be preferred. However, even when a given hypothesis H_i has a *lower* fit than an alternative hypothesis $H_{i'}$, H_i will still be preferred if the decrease in complexity for H_i compared to $H_{i'}$ is larger than the decrease in fit. When evaluating the set of hypotheses under consideration, it is thus important to consider how the relative complexities of these hypotheses will influence the results. The most specific (least complex) hypotheses are equality-constrained hypotheses (e.g., $\mu_1 = \mu_2$). The least specific (most complex) hypothesis is the unconstrained hypothesis H_u which poses no constraints on the parameter values. The unconstrained hypothesis will therefore only be preferred if none of the other hypotheses provide a good fit to the data. Note that per their definition, the complexity and fit of H_u are always 1, which means there is an upper limit of BF_{iu} that is determined by the complexity of H_i (Klugkist & Volker, 2023). For example, the hypothesis $H_1 : \beta > 0$ has a complexity of 0.5 since this hypothesis covers half of the unconstrained prior distribution. This means that BF_{1u} has an upper limit of 2: when the fit of H_1 is perfect, $BF_{1u} = 1/0.5 = 2$. In contrast, when testing a hypothesis against its complement or against another constrained hypothesis, the resulting BF does not have an upper limit. If a researcher is interested in only one informative hypothesis, then testing against the complement is the most powerful, because the two hypotheses cover mutually exclusive regions of the parameter space (Klugkist & Volker, 2023). Note, finally, that when testing an equality-constrained hypothesis (e.g., $H_0 : \mu_1 = \mu_2$), the unconstrained hypothesis H_u is statistically equivalent to the complement hypothesis H_c (not H_0; Hoijtink et al., 2019).[1] We will further illustrate the interplay between the fit and complexity of the hypotheses under consideration in the simulation results.

When comparing a set of hypotheses, it is useful to translate the BFs into posterior model probabilities (PMPs; Kass & Raftery, 1995). PMPs facilitate interpretation as they have values between 0 and 1 (with values closer to 1 indicating more support) that add up to 1 over all hypotheses under consideration; they thus express the relative support for each of the tested hypotheses. Translating BFs into PMPs is simple, given that the BF is a multiplicative factor that transforms the prior odds of two hypotheses (i.e., the ratio of the probabilities of each hypothesis before any data is collected) into the posterior odds (i.e., the ratio of the probabilities of each hypothesis after seeing the data), as shown in Eq. 5:

$$\frac{P(H_i)}{P(H_u)} \times BF_{iu} = \frac{P(H_i|\boldsymbol{X})}{P(H_u|\boldsymbol{X})}. \quad (5)$$

The PMP for hypothesis H_i can thus be computed as

$$PMP(H_i) = \frac{P(H_i) \times BF_{iu}}{\sum_{i=1}^{m} P(H_i) \times BF_{iu}}, \quad (6)$$

where $P(H_i)$ is the prior model probability of hypothesis H_i with $i = 1, 2, \ldots, m$. Typically, equal prior probabilities are assigned to each hypothesis, which means that each hypothesis receives a prior model probability of $1/m$.

Synthesis of Bayes factors

Once the study-specific BFs (or PMPs) are obtained, the final step is to aggregate them to yield the joint support for each hypothesis across all studies. The joint support for each hypothesis is obtained by updating the model probabilities with each new study. In other words, the posterior model probability of study k can be used as the prior model probability for study k+1. Irrespective of the order of the studies, this process can be repeated for a total of K studies, assuming all studies are independent (Kuiper et al., 2013). The aggregated PMP for hypothesis H_i is then given by

$$PMP(H_i)^K = \frac{P^0(H_i) \times \prod_{k=1}^{K} BF_{iu}^k}{\sum_{i=1}^{m} P^0(H_i) \times \prod_{k=1}^{K} BF_{iu}^k}, \quad (7)$$

[1] As explained in Hoijtink et al. (2019, Footnote 1) this is "because, loosely spoken, among the infinite number of possible combinations of values for μ_1, μ_2, and μ_3 that are in agreement with H_u, $\mu_1 = \mu_2 = \mu_3$ has a 'zero probability' of occurring." So whether $\mu_1 = \mu_2 = \mu_3$ is included in the hypothesis (as in H_u) or not (as in H_c) will not affect the Bayes factor.

where $P^0(H_i)$ indicates the prior model probability for hypothesis H_i before any study has been conducted. The numerator represents the joint probability of the data from all studies under the assumption that the constraints of the target hypothesis hold separately in each study, whereas the denominator sums the joint probabilities of the data under each of the hypotheses under consideration. The aggregated PMP therefore provides the joint evidence for a hypothesis in each study relative to the other hypotheses considered. Note that in order to compute the aggregated PMP it is not necessary for the study-specific BFs to have used the same priors for the *model estimates*. BES assumes that studies provide independent pieces of evidence, which means that if the prior used within a study is deemed appropriate to estimate the parameters and/or compute the BFs, then the evidence from this study can be aggregated with the evidence from other studies regardless of whether these other studies used the same prior on the model estimates (see Klugkist & Volker, 2023).

With equal prior *model probabilities* for each hypothesis (i.e., $P^0(H_i) = 1/m$), Eq. 7 can be rewritten as

$$PMP(H_i)^K = \frac{\prod_{k=1}^{K} PMP(H_i)^k}{\sum_{i=1}^{m} \prod_{k=1}^{K} PMP(H_i)^k}, \tag{8}$$

where $PMP(H_i)^k$ is the posterior model probability of hypothesis H_i in study k. When only two hypotheses are tested, this formula simplifies to

$$PMP(H_i)^K = \frac{\prod_{k=1}^{K} PMP(H_i)^k}{\prod_{k=1}^{K} PMP(H_i)^k + \prod_{k=1}^{K} (1 - PMP(H_i)^k)}. \tag{9}$$

Note that with equal prior model probabilities, the information provided by the aggregated PMP is the same as the product of Bayes factors.

Difference with meta-analysis

As mentioned above, the aggregated PMPs obtained by BES give the relative joint probability of the data from all studies under the assumption that the constraints of the target hypothesis hold in each study separately. The aggregated PMP can therefore be used to indicate which hypothesis best describes each study. This is different from inferences based on data-pooling techniques such as meta-analysis, which indicate whether the target hypothesis is supported by the *pooled* data. In a meta-analysis (e.g., Hedges & Olkin, 1985; Hedges & Vevea, 1998), it is assumed that for a set of $k = 1, \ldots, K$ independent studies, the observed effect in study k is given by

$$y_k \sim N(\theta_k, v_k), \tag{10}$$

where θ_k is the (unknown) true effect, and v_k is the sampling variance (which is assumed to be known). Since most meta-analyses are based on sets of studies that are not identical, it is typically assumed that there is variability among the true effects. If this variability is not systematic (i.e., between-study differences do not systematically predict effect size), then this variability can be modeled as purely random with the random-effects model given by

$$\theta_k \sim N(\mu, \tau^2), \tag{11}$$

that is, the true effects θ_k are assumed to be normally distributed with mean μ and variance τ^2. In contrast to BES, where the model parameters are estimated independently in each study and are therefore allowed to vary, meta-analysis assumes that the mean population effect μ is identical for each study. To yield a better estimate of this common population parameter, a weighted average of the observed effect sizes y_k is computed, with weights typically equal to the inverse variance (i.e., $1/[v_k + \hat{\tau}^2]$, where $\hat{\tau}^2$ denotes the estimate of τ^2). Inferential tests and confidence intervals then indicate whether the estimated common population parameter significantly differs from zero.

It can sometimes happen that BES does not yield the same results as data-pooling techniques like meta-analysis, as for example illustrated by Regenwetter et al. (2018), who used both BES (which they called the "group BF") and a data-pooling technique (i.e., BFs computed on the pooled data which they called the "pooled BF") in the context of aggregating single participant data. Diverging results may, for example, occur when the hypothesis best supported by the aggregated data is not well supported in any of the individual studies. Another example of when the methods may not converge on the same hypothesis is when a given hypothesis, H_i, describes most studies reasonably well but provides a *very poor* fit for a few studies that are better described by other hypotheses. In this case, H_i will typically not be selected as the best hypothesis by BES but might still be selected as the best hypothesis by data-pooling techniques. Although both meta-analysis and BES can thus be used for testing hypotheses across multiple studies, the results may sometimes differ because the methods answer a different synthesis question. This will be further illustrated in the simulation study.

Simulation

In the simulation study, we evaluated the performance of BES compared to meta-analysis as a function of true population effect size, total sample size per study, level of variability among the study-specific true effects, and number

of studies. We also assessed how much influence one study with an extremely small sample size or opposite effect size has on the aggregated result. For BES, we show how the results depend on which hypotheses are considered. The simulation was conducted in R (R Core Team, 2021, Version 4.1.0). All R scripts, simulated datasets, and (supplementary) figures are available in the Open Science Framework repository at https://osf.io/gbtyk/.

Data generation

We used the standardized mean difference as an effect size as this is a very common effect-size metric in meta-analyses of experimental studies. For each artificial study, k, with $k = 1, \ldots, K$, we generated data for a total of N participants divided equally across two groups, which we refer to here as the experimental group and the control group. Let $\mathbf{Y}_k^E$ be the $N/2 \times 1$ vector of outcomes for the experimental group and $\mathbf{Y}_k^C$ the $N/2 \times 1$ vector of outcomes for the control group. Assuming normality of the data, these outcomes can be generated as

$$\mathbf{Y}_k^E \sim N\left(\delta_k, 1\right) \text{and } \mathbf{Y}_k^C \sim N(0, 1), \tag{12}$$

where δ_k is the true standardized mean difference for study k. Given the relatively large amount of heterogeneity in effect sizes found in psychological meta-analyses (Linden & Hönekopp, 2021; van Erp et al., 2017), we modeled variability among the study-specific true effect sizes by sampling them from a normal distribution, that is,

$$\delta_k \sim N\left(\delta, \tau^2\right), \tag{13}$$

where δ is the true mean population effect size and τ^2 is the true between-study variance in effect sizes. After simulating the data, we estimated the standardized mean difference $\widehat{\delta}_k$ and its variance v_k following standard formulas, that is,

$$\widehat{\delta}_k = \frac{\overline{Y}_k^E - \overline{Y}_k^C}{\sqrt{(SD_1^2 + SD_2^2)/2}} \tag{14}$$

and

$$v_k = \frac{4}{N} + \frac{\widehat{\delta}_k^2}{2(N-2)}, \tag{15}$$

The different simulation conditions were created by varying the true mean population effect size (δ), the true between-study standard deviation (τ), and the total sample size per study (N). For δ, we chose a small ($\delta = 0.2$) and medium effect ($\delta = 0.5$), given that small-to-medium effect sizes are the most common in the field of psychology (Lovakov & Agadullina, 2021; Open Science Collaboration, 2015). We also included a null effect ($\delta = 0$) to investigate the behavior of BES when the null hypothesis is true. For τ, we chose the first ($\tau = 0.1$) and third quartile ($\tau = 0.3$) of estimated τ values found in 189 meta-analyses of standardized mean differences published in *Psychological Bulletin* (van Erp et al., 2017), thus representing relatively small and relatively large variation in effect sizes within the field of psychology. Finally, for N we chose a range between 20 and 200 with a step size of 20. The lower limit represents the absolute minimal sample size typically considered for experimental studies in psychology, whereas the upper limit represents the maximum sample size found in ~ 80% of 2642 experimental studies collected by Lovakov & Agadullina (2021). For each of the 3 ($\delta \in \{0, 0.2, 0.5\}$) x 2 ($\tau \in \{0.1, 0.3\}$) x 10 ($N \in \{20, 40, \ldots, 200\}$) = 60 conditions, we simulated a set of $K = 30$ studies. We ran 1000 replications per condition, yielding a total of 60 x 30 x 1000 = 1,800,000 simulated datasets. To investigate the effect of the number of studies, we synthesized the first two studies of each replication by means of meta-analysis and BES and then cumulatively added one study at a time until all $K = 30$ studies were synthesized. This way, we could directly investigate the effect of adding more studies to the existing set of studies.[2]

We also investigated how many studies one needs to still yield aggregated support for the target hypothesis when one study provides strong evidence *against* this hypothesis. This may for example happen when one of the studies is underpowered (and therefore provides more support for the null hypothesis), or when one of the studies is sampled from a population with an opposite effect size (e.g., when this is the only study investigating older participants, and the effect turns out to be reversed for older vs. younger participants). To investigate this, we focused on the subset of studies with $\delta = 0.5$ and $N = 140$, such that all studies had sufficient power to detect the true population effect; detecting a standardized mean difference of 0.5 with 80% power requires a total sample size of 128 (calculation performed with G*power; Faul et al., 2007). Then, we replaced the first study in each replication by a newly simulated study with a total sample size of $N = 30$ (representing an underpowered study) or by a newly simulated study with a population effect of $\delta = -0.5$ (representing a study that tested a sample from a different population).

[2] Alternatively, we could have iteratively resampled a new set of studies for each value of k. Note, however, that since we conducted 1000 replications per simulation condition, the mean and variance of the observed effect sizes very closely reflect the values we specified when generating the data. The results would therefore be virtually identical if we had resampled a new set of studies for each value of k. Here, we chose to cumulatively add the K studies because this nicely shows what happens if new studies are added to the existing set of studies. This is fitting in the context of research synthesis since all systematic reviews, including meta-analyses, are advised to be updated (more or less) regularly as new evidence becomes available (Garner et al., 2016).

Meta-analysis

Given that effect sizes in psychology are typically assumed to be heterogenous, we conducted a random-effects meta-analysis (henceforth meta-analysis) with the widely used rma() function from the *metafor* package (Viechtbauer, 2010). This function fits a random-effects model using a two-step approach. First, the amount of between-study variance (i.e., τ^2) is estimated using a restricted maximum likelihood estimator (Raudenbush, 2009; Viechtbauer, 2005). Then, the average true effect is estimated via weighted least squares, with weights equal to the inverse variance. Once the parameter estimates have been obtained, confidence intervals are computed based on a standard normal distribution.

Bayesian evidence synthesis

For BES, we formulated our hypothesis of interest as $H_1 : \delta > 0$. This hypothesis can be tested against three possible alternative hypotheses, namely the null hypothesis, $H_0 : \delta = 0$, the complement hypothesis, H_c: not H_1 (in this case $\delta < 0$), and the unconstrained hypothesis, $H_u : \delta$ (i.e., δ can take on any value). We tested H_1 against each of these alternative hypotheses in turn to demonstrate how the individual characteristics of each alternative hypothesis affect the results. In practice, however, researchers may often want to test multiple hypotheses simultaneously (for examples of empirical studies that used BES to test multiple hypotheses simultaneously, see Kevenaar et al., 2021; Veldkamp et al., 2021; Zondervan-Zwijnenburg, Richards et al., 2020a, Zondervan-Zwijnenburg, Veldkamp et al., 2020b) and we will show how to do this in the empirical demonstration. Nevertheless, it may sometimes happen that researchers are *only* interested in testing a hypothesis H_i against H_c, for example when there are two competing theories that make opposite predictions, and the null hypothesis is considered very unlikely. Likewise, a researcher may opt to only test H_i against H_u when H_i is the only theoretically relevant hypothesis.

To obtain the study-specific PMPs for H_1 we used the R-package *bain* (Gu et al., 2020), which computes the approximate adjusted fractional BFs given in Eqs. (2–4) based on the effect size estimates ($\widehat{\delta}_k$), their variance (v_k) and the study-specific sample size, and translates these BFs into PMPs (see Eq. 6). We assumed equal prior probabilities for each hypothesis and used bain's default priors for the effect-size estimates. After computing the study-specific PMPs, we calculated the aggregated PMPs according to Eq. (9).

Performance measure

We compared meta-analysis and BES by looking at the proportion of times H_1 was "accepted" across simulations. In theory, we cannot accept H_1 within the frequentist framework (we can only reject H_0), but we considered the meta-analysis results to be compatible with H_1 if the meta-analytic effect size was positive and the lower bound of the confidence interval (CI) was greater than zero. We calculated the H_1 acceptance rate based on both 95% CIs (as this is what is typically reported in the literature) as well as 90% CIs (given that H_1 is a directional hypothesis).

For BES, we adopted the decision rule used by Klaassen et al. (2018) who considered a hypothesis H_i the best of a set of m hypotheses if the evidence for H_i was at least $m - 1$ times (with a minimum value of 2) stronger than for any other hypothesis. This ensures that the aggregated PMP of the best hypothesis is at least .50 when all hypotheses are equally likely *a priori*. Since we only tested two hypotheses at a time, this means that we accepted H_1 if its aggregated PMP was at least twice as high as for the alternative hypothesis (i.e., $BF_{1.} = 2$), corresponding to an aggregated PMP of .67 or higher for H_1, and .33 or lower for the alternative hypothesis.

The advantage of this decision rule is that it considers the number of hypotheses under consideration. The larger the number of hypotheses, the less support any one hypothesis will receive (Hoijtink et al., 2019), meaning that using the same cut-off point for comparing two or three (or more) hypotheses is not appropriate. However, which value of the aggregated PMP can be seen as "strong" evidence remains a question for future research and will likely vary as a function of the research field, the number of hypotheses evaluated, and the characteristics of the hypotheses under consideration. Although often-cited guidelines in the literature consider a BF of 10 to provide strong evidence (corresponding to a PMP of .91; Jeffreys, 1961), these guidelines were made for evaluating the unconstrained hypothesis against the null hypothesis in a single study and do not necessarily generalize to the context of evaluating other, or multiple, hypotheses *across* studies.

To investigate the sensitivity of the results to the employed decision rule, we also used a cut-off value of .91 for the aggregated PMP, as shown in Figs. S1 to S4 (available at https://osf.io/gbtyk/). These figures show that this higher cut-off value mainly affects the results when testing against H_u. We will come back to this finding in the Discussion.

Results

In Figs. 1, 2 and 3 we show the H_1 acceptance rate across simulation replications as a function of (i) the synthesis method (meta-analysis vs. BES), (ii) the total sample size per study, (iii) the between-study standard deviation τ, and (iv) the number of studies synthesized. We show separate lines for the synthesis of 5, 10, and 30 studies, as these represent the first, second, and third quartile of the number of effect sizes synthesized per meta-analysis in 747 meta-analyses

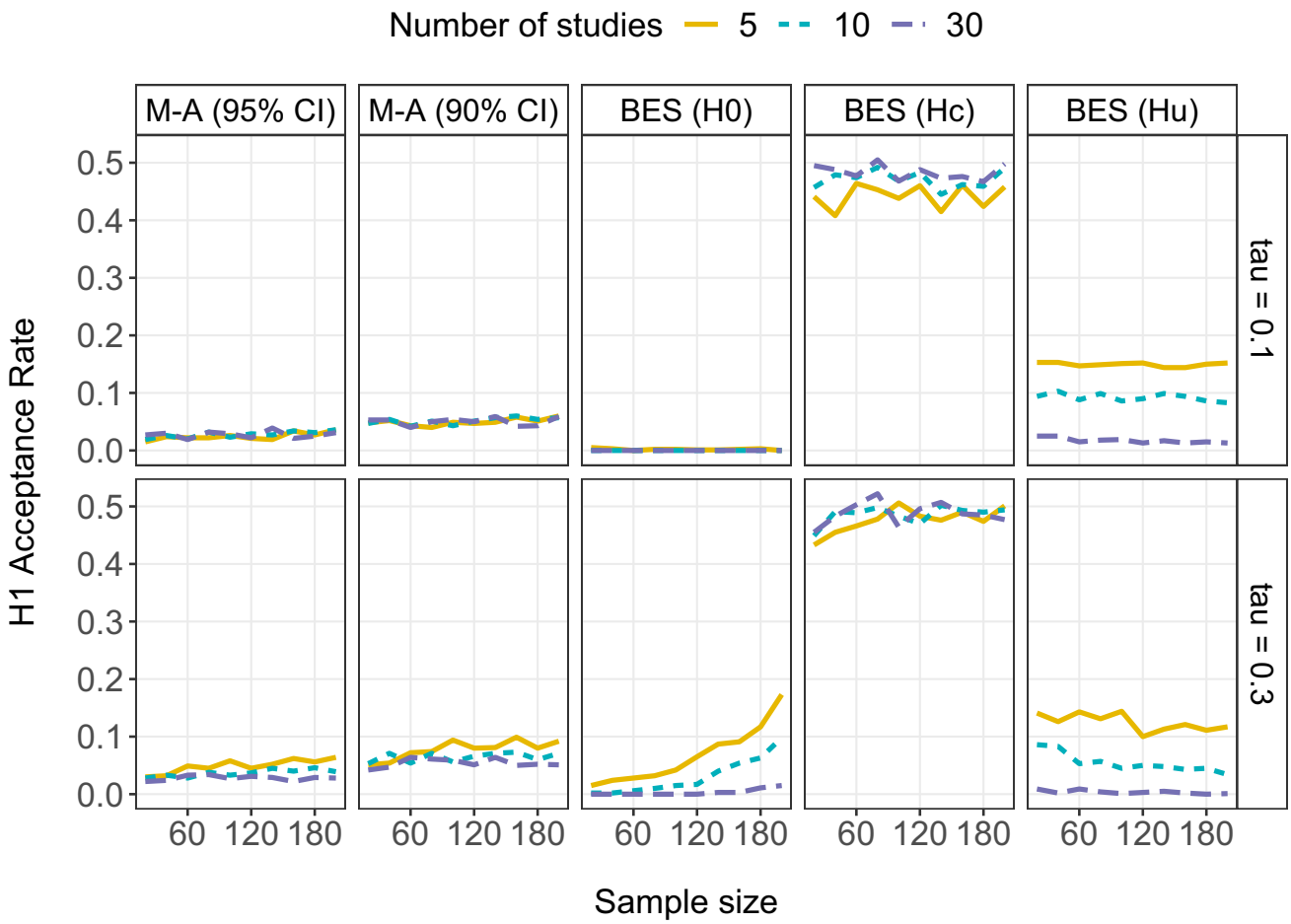

Fig. 1 H_1 acceptance rate for random-effects meta-analysis (M-A) with either 95% or 90% confidence intervals and Bayesian evidence synthesis (BES) as a function of the between-study standard deviation in population effect sizes τ, total sample size per study, and number of studies when the overall mean population effect size is zero. Note that for BES, H_1 ($\delta > 0$) is tested against three alternative hypotheses: H_0 ($\delta = 0$), H_c ($\delta < 0$), or H_u (δ)

reported in *Psychological Bulletin* (van Erp et al., 2017).[3] In these figures, the true mean population effect size δ is 0, 0.2 and 0.5, respectively, representing a null, small, and medium effect. Note that when the true effect is zero (Fig. 1), a low H_1 acceptance rate indicates better performance of the method. Finally, in Fig. 4, we focus on the subset of studies with $\delta = 0.5$ and $N = 140$ and show how including one underpowered study or one study with an opposite population effect influences the results.

[3] Although it is uncommon to see meta-analytic papers with only five studies, many meta-analytic papers report on separate meta-analyses for different outcome measures (e.g., the 747 meta-analyses collected by van Erp et al. were published in only 61 papers). A set of five studies per meta-analytic outcome is not uncommon, as illustrated by the fact that 25% of the meta-analyses reported by van Erp et al. synthesized five studies or less.

Null effect ($\delta = 0$)

Figure 1 shows the results when the null hypothesis is true. Note that in this figure the H_1 acceptance rate runs from 0 to 0.5, rather than to 1. As expected, H_1 is accepted in less than 5% of the meta-analyses based on 95% CIs, and less than 10% of the meta-analyses based on 90% CIs, irrespective of sample size, number of studies and τ values (although the H_1 acceptance rate slightly increases when between-study variance is large and few studies are combined). For BES, the results depend on the alternative hypothesis. When testing against H_0, H_1 is consistently rejected when there is relatively little variation (i.e., $\tau = 0.1$) in population effect sizes across studies. In contrast, when there is relatively large variation in population effect sizes (i.e., $\tau = 0.3$), the H_1 acceptance rate slightly increases with larger sample sizes, especially when combining less than 30 studies. This can be

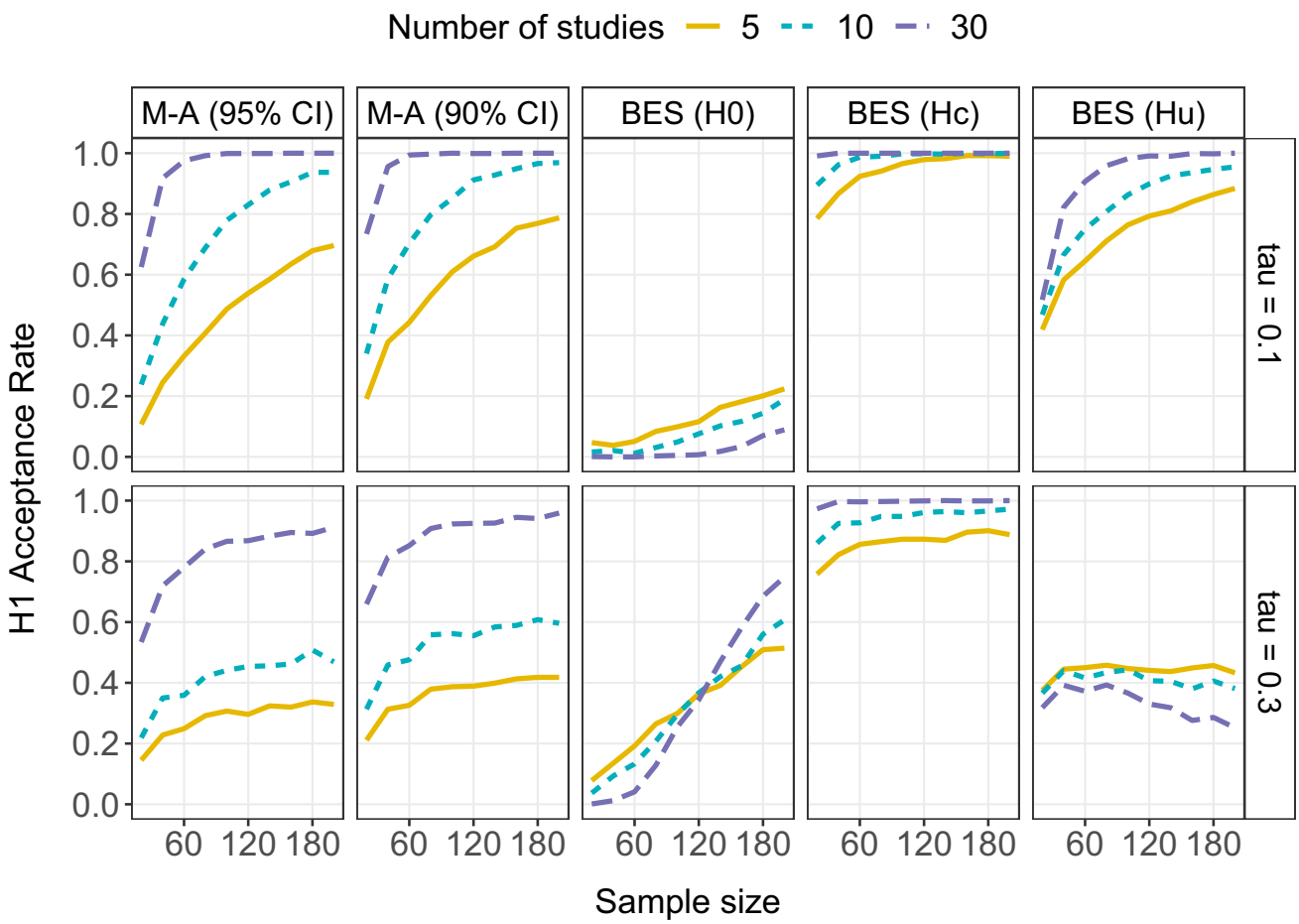

Fig. 2 H_1 acceptance rate for random-effects meta-(M-A) with either 95% or 90% confidence intervals and Bayesian evidence synthesis (BES) as a function of the between-study standard deviation in population effect sizes τ, total sample size per study, and number of studies when the overall mean population effect size is 0.20 (i.e., a small effect). Note that for BES, H_1 ($\delta > 0$) is tested against three alternative hypotheses: H_0 ($\delta = 0$), H_c ($\delta < 0$), or H_u (δ)

explained as follows: When the between-study variance in true effect sizes is large, there will be more studies with true-positive effects as well as with true-negative effects. Positive effects support the hypothesis of interest H_1 (especially when the sample size is large), whereas (large) negative effects do not support the alternative hypothesis H_0. Therefore, the H_1 acceptance rate increases with larger sample sizes when there is large between-study variation among the true effects. However, this trend is countered by combining more studies, as the true effect of most studies will be near the mean population effect size ($\delta = 0$). With more studies, the aggregated support for H_0 will eventually outweigh the few studies that support H_1.

When testing H_1 against H_u, we expect H_u to receive more support than H_1 when H_1 is not true, as is the case here. We see that H_1 is indeed rejected at least 80% of the time when we combine five studies, and almost 100% of the time when we combine 30 studies. When we use a higher cut-off value for accepting H_1, we reject H_1 more often: When the cut-off value is .91, we consistently reject H_1 regardless of τ value, sample size or number of studies combined (see Fig. S1 at https://osf.io/gbtyk/).

Finally, when testing H_1 against H_c, the H_1 acceptance rate is around .50, which is to be expected since the true mean population effect (i.e., $\delta = 0$) is on the boundary of the considered hypotheses. This shows the importance of including all relevant hypotheses: When the true hypothesis (in this case H_0) is not in the set of considered hypotheses, one runs the risk of selecting a hypothesis that provides a poor fit to the data.

Small effect ($\delta = 0.2$)

Figure 2 shows the results when the mean population effect is small (i.e., $\delta = 0.2$). Note that this represents a situation in which all studies are underpowered, as detecting a standardized mean difference of 0.2 with 80% power would require a total sample size of 788 per study (calculation performed with

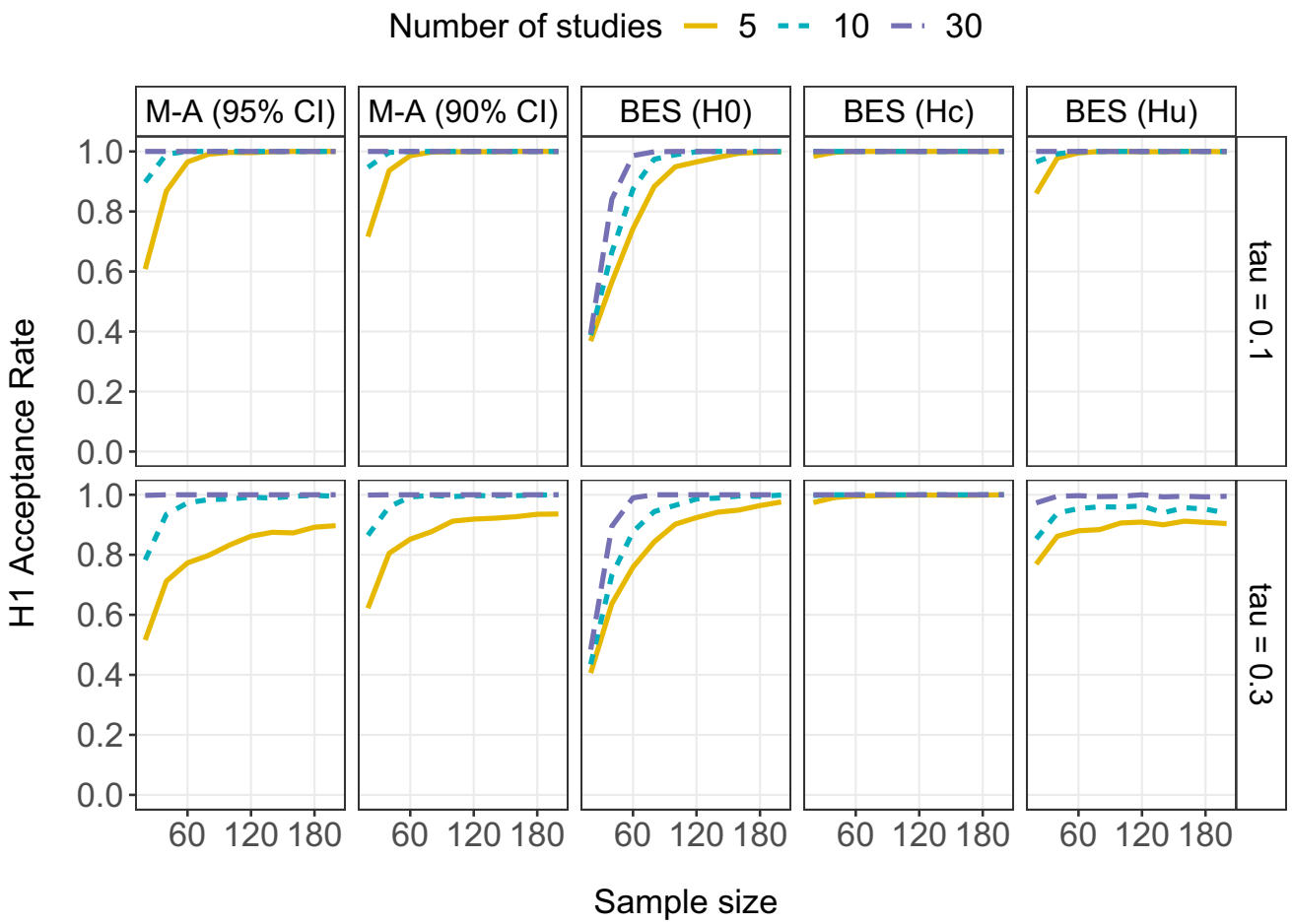

Fig. 3 H_1 acceptance rate for random-effects meta-analysis (M-A) with either 95% or 90% confidence intervals and Bayesian evidence synthesis (BES) as a function of the between-study standard deviation in population effect sizes τ, total sample size per study and number of studies when the overall mean population effect size is 0.50 (i.e., a medium effect). Note that for BES, H_1 ($\delta > 0$) is tested against three alternative hypotheses: H_0 ($\delta = 0$), H_c ($\delta < 0$), or H_u (δ)

G*power; Faul et al., 2007). Given the prevalence of small-to-medium effects and the rarity of sample sizes over 200, this is expected to be a typical situation in the field of experimental psychology (see Linden & Hönekopp, 2021; Lovakov & Agadullina, 2021; Open Science Collaboration, 2015).

When the mean population effect is small, the H_1 acceptance rate is rather low for meta-analysis when combining less than 30 studies, especially when variation among the true effects is large. However, for both small and large τ values, the H_1 acceptance rate increases when combining more studies. This exemplifies how meta-analysis helps mitigate power issues, especially when combining a relatively large number of studies.

For BES, the results again depend on the alternative hypothesis. When testing against H_0 and between-study variation is small, the H_1 acceptance rate is very low. This is because underpowered studies each provide stronger evidence for H_0 than for H_1, and BES answers the question of which hypothesis best describes each individual study, rather than which hypothesis best describes the pooled data. Moreover, combining more studies only strengthens, rather than weakens, the support for H_0, as the more studies we combine that all show support for H_0, the more confident we become that H_0 best describes each individual study (Klugkist & Volker, 2023). In other words, BES does not solve power issues when testing against the null. That said, the H_1 acceptance rate does slightly increase with larger sample sizes, even though studies with $N = 200$ are still very underpowered (recall that a minimum sample size of $N = 788$ is needed here).

A different picture emerges for testing against H_0 when between-study variation is large. Here we see that the H_1 acceptance rate increases more steeply as a function of sample size. As in Fig. 1, this is because large between-study variance in true effects yields effect sizes that are further from the mean effect in both directions, and positive effects provide support for H_1 whereas (large) negative effects do not provide evidence for H_0. We also see that when τ is large, support for H_1 starts to increase again with a greater number

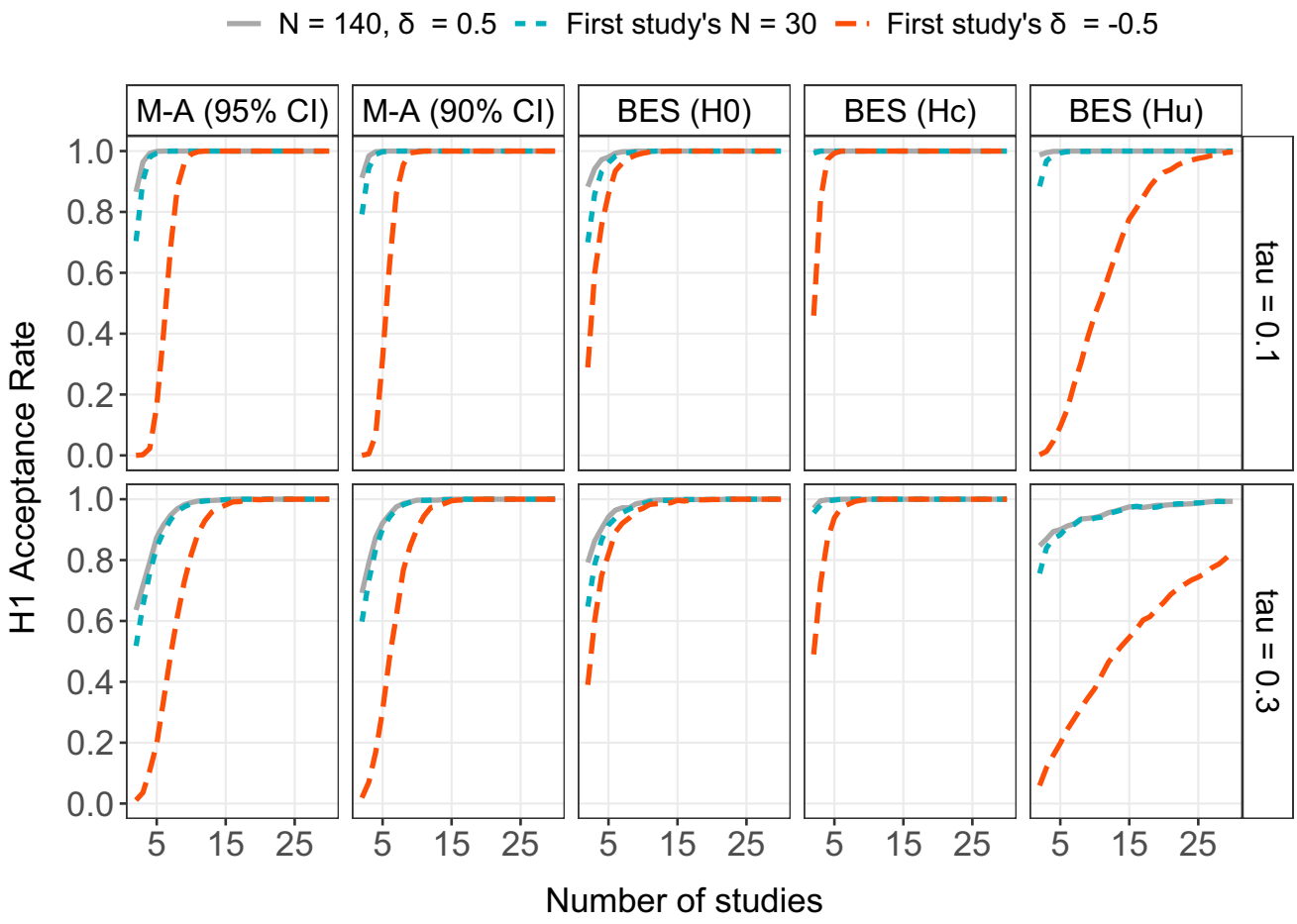

Fig. 4 H_1 acceptance rate for random-effects meta-analysis (M-A) with either 95% or 90% confidence intervals and Bayesian evidence synthesis (BES) as a function of the between-study standard deviation in population effect sizes τ and number of studies (between 2 and 30). The *lines* represent three scenarios: (i) all studies had an N of 140 and a δ of 0.5 (*grey solid line*); (ii) same as the grey line but the first study had an N of 30 (*blue short-dashed line*); (iii) same as the grey line but the first study had a δ of – 0.5 (*red long-dashed line*). Note that for BES, H_1 ($\delta > 0$) is tested against three alternative hypotheses: H_0 ($\delta = 0$), H_c ($\delta < 0$), or H_u (δ)

of studies once the study-specific sample sizes become large enough (i.e., when $N > 120$).

When testing against H_c, BES consistently yields evidence in favor of H_1, almost regardless of sample size or the number of studies combined. This is not surprising, given that positive effects will always render more evidence for H_1 (i.e., $\delta > 0$) than for H_c (i.e., $\delta < 0$), even when they are small. This shows that testing against H_c is a very powerful strategy whenever the null hypothesis is not of interest or is considered very unlikely.

Finally, testing against H_u yields a similar pattern of results as meta-analysis when between-study variation is small, in that the H_1 acceptance rate quickly increases as a function of sample size and the number of studies (with the exact H_1 acceptance rate depending on the employed decision rule, cf. Fig. 2 and Fig. S2). In contrast, the H_1 acceptance rate remains low when between-study variation is large, as large variation around a small mean effect will yield at least some studies that provide very strong evidence *against* H_1. In contrast, *all* studies are in line with H_u by definition. H_u will therefore typically receive more evidence than H_1 when the true mean effect is small and between-study variance is large. Moreover, the aggregated support for H_1 decreases, rather than increases, with the number of studies in this situation; with more studies, there is a higher chance that some of these studies provide strong evidence against H_1, which decreases the aggregated support for H_1.

Medium effect ($\delta = 0.5$)

Figure 3 shows the results when all studies are drawn from a population with a medium mean effect (i.e., $\delta = 0.5$). Here, we see that the H_1 acceptance rate is very high across all synthesis methods. We still need some power in the individual studies when testing against H_0 with either meta-analysis or BES, but once the total sample size per study is larger

Table 1 Study-specific and aggregated posterior model probabilities for H_1 against each of the alternative hypotheses (H_0, H_c, H_u) for the studies reported in the meta-analysis by Lammertink et al. (2017)

Study	PMP_{10}	PMP_{1c}	PMP_{1u}
1. Evans et al. (2009)	.77	.99	.66
2. Evans et al. (2010)	.94	>.99	.67
3. Lukacs & Kemeny (2014)	.29	.91	.65
4. Mayor-Dubois et al. (2014)	.95	>.99	.67
5. Hsu et al. (2014)[a]	.20	.70	.58
6. Hsu et al. (2014)[b]	.35	.87	.64
7. Hsu et al. (2014)[c]	.88	.99	.67
8. Haebig et al. (2017)	.84	.99	.66
9. Grunow et al. (2006)	.22	.75	.60
10. Torkildsen (2010)	.95	>.99	.67
Aggregated PMP	> .99	>.99	> .99

Note. [a]Low variability (X = 2) condition. [b]Mid variability (X = 12) condition. [c]Low variability (X = 24) condition

than 60, the H_1 acceptance rate is above .80 regardless of between-study variation, alternative hypothesis tested, or number of studies combined.

Effect of one study with a small sample size or opposite effect size

In Fig. 4, we show the subset of studies for which $\delta = 0.5$ and $N = 140$ (grey line) and compare this to the same subset of studies for which the first study is replaced by a newly simulated study with a very small total sample size (i.e., $N = 30$, blue line) or with an opposite mean population effect (i.e., $\delta = -0.5$, red line). Whereas including one underpowered study hardly has an effect on the H_1 acceptance rate (and no effect whatsoever when combining at least five studies), including a study with an opposite population effect drastically decreases the H_1 acceptance rate for both meta-analysis and when testing against H_u with BES (but not when testing against H_0 or H_c). This effect is countered, however, by combining more studies.

For meta-analysis, approximately ten studies that support H_1 need to be included to again yield aggregated support for H_1 at least 80% of the time. When testing against H_u with BES, one needs to include at least 15 studies that support H_1 when between-study variation is small, and at least 30 studies when between-study variation is large (more studies are needed when using a higher cut-off value for accepting H_1, see Fig. S4). In other words, testing against H_u is the most sensitive to including studies that show strong support *against* the hypothesis of interest. Again, this is not surprising given that H_u is always true and therefore describes each study well, whereas H_1 in this case provides a very poor fit to one of the included studies. However, since H_1 is more parsimonious than H_u, we eventually still obtain more aggregated support for H_1 if we combine enough studies.

Empirical demonstration

The data for this demonstration come from a meta-analysis on auditory verbal statistical learning in people with and without developmental language disorder (DLD), previously known as specific language impairment (SLI; Lammertink et al., 2017). In this study, ten effect sizes from eight studies were analyzed by means of a random-effects meta-analysis to investigate whether people with DLD show a statistical learning deficit compared to people without DLD. An important feature of this meta-analysis was that only effect sizes were included that came from non-overlapping participant samples. This was important given that BES relies on the assumption that studies provide independent pieces of evidence for the hypotheses under consideration.

Prior to conducting the meta-analysis, Lammertink et al. (2017) computed a standardized mean difference (SMD) for each study based on summary or test statistics reported in the primary studies (i.e., means and standard deviations per group or an *F*-/*t*-statistic). The overall standardized mean difference between participants with and without DLD was 0.54, which was significantly different from zero ($p < .001$, 95% confidence interval [0.36, 0.71]). The authors therefore concluded that there is a robust difference between people with and without DLD in their detection of statistical regularities in the auditory input, congruent with the hypothesis that people with DLD are less effective in statistical learning.

Below we demonstrate how BES can be performed on the effect sizes included in this meta-analysis. For illustrative purposes, we show how to test H_1 (SMD > 0; that is, participants without DLD score higher on statistical learning than participants with DLD) against each alternative hypothesis (H_0, H_c, H_u) in turn, like we did in the simulations. Afterwards we show how multiple hypotheses can be tested simultaneously and how a different type of hypothesis (i.e., a range-constrained hypothesis) can be formulated. Note, however, that if we were to perform BES on this set of studies for an empirical paper, we would probably only test H_1 against H_0, as these seem to be the two hypotheses of interest and H_c is theoretically very unlikely.

We start by specifying the effect size, variance, and total sample size per study as computed by Lammertink et al. (2017; see https://osf.io/4exbz/).

```
ES <- c(0.48, 1.05, 0.3, 0.85, 0.17, 0.36, 0.79, 0.68, 0.2, 1.12)
var <- c(0.04, 0.15, 0.05, 0.08, 0.1, 0.1, 0.1, 0.08, 0.09, 0.16)
N <- c(113, 28, 115, 79, 40, 40, 40, 49, 44, 28)
```

This is all the information we need to compute the study-specific PMPs with the *bain* package.[4] Note that bain() always returns three different sets of PMPs: PMPa contains the PMPs of the hypotheses specified by the user, PMPb adds the unconstrained hypothesis H_u to the set of specified hypotheses, and PMPc adds H_c, that is, the complement of the union of the hypotheses specified by the user. To test H_1 separately against both H_c and H_u, we therefore only need to specify H_1 in the call to bain(), as PMPb will then test H_1 against H_u, and PMPc will test H_1 against H_c. In contrast, to test H_1 against H_0, we must specify both hypotheses; PMPa then contains the result we need. The code below computes the study-specific PMPs.

```
library(bain)

PMP10 <-c(); PMP1c <-c(); PMP1u <-c()

for(i in 1:length(ES)){
  SMD <- ES[i]
  names(SMD) <- "SMD"

  # test H1 against H0
  res1 <- bain(SMD,
       hypothesis = "SMD > 0; SMD = 0",
       n = N[i],
       Sigma = var[i])
  PMP10[i] <- res1$fit["H1","PMPa"]

  # test H1 against Hc/Hu
  res2 <- bain(SMD,
       hypothesis = "SMD > 0",
       n = N[i],
       Sigma = var[i])
  PMP1c[i] <- res2$fit["H1","PMPc"]
  PMP1u[i] <- res2$fit["H1","PMPb"]
}
```

Now that we have computed the study-specific PMPs, we can compute the aggregated PMP for H1 against each of the alternatives with the code below (corresponding to Eq. 9).

```
BES10 <- prod(PMP10)/(prod(PMP10) + prod(1-PMP10))
BES1c <- prod(PMP1c)/(prod(PMP1c) + prod(1-PMP1c))
BES1u <- prod(PMP1u)/(prod(PMP1u) + prod(1-PMP1u))
```

Table 1 shows the individual PMPs for H_1 for each study as well as the aggregated PMP. We do not show the (aggregated) PMPs for the alternative hypotheses, as these are simply 1-PMP_1. The results show that regardless of the considered alternative hypothesis, we obtain overwhelming evidence in favor of H_1, congruent with the result from the meta-analysis reported by Lammertink et al. (2017). Notably, we see that this is the case despite variation in the *study-specific* PMPs across the alternative hypotheses. Due to the relatively small sample sizes per study, H_0 receives more evidence than H_1 in those studies where the estimated effect size is small (note that each study only had 36–79% power to detect a medium effect). In contrast, H_c never receives more evidence than H_1, since all effect sizes are positive and thus provide no evidence in favor of a negative effect size. Finally, H_1 receives more support than H_u in each study, but the maximum study-specific PMP is .67, as this is the upper limit of PMP_{1u} given the complexity of H_1 (see Introduction). Despite these differences in the study-specific PMPs, the aggregated evidence across all studies is clearly in favor of H_1 with all aggregated PMPs being greater than .99.

As mentioned above, it is also possible to test multiple hypotheses simultaneously. For example, if we were to test H_1, H_0, and H_u simultaneously, we would run the following code to compute the study-specific PMPs.

```
PMP1 <-c(); PMP0 <-c(); PMPu <-c()

for(i in 1:length(ES)){
  SMD <- ES[i]
  names(SMD) <-"SMD"
  res <-bain(SMD,
       hypothesis="SMD>0;SMD=0",
       n=N[i],
       Sigma=var[i])
  PMP1[i] <-res$fit["H1","PMPb"]
  PMP0[i] <-res$fit["H2","PMPb"]
  PMPu[i] <-res$fit["Hu","PMPb"]
}
```

The aggregated PMPs are then computed according to Eq. (8). In the code below, BES1 gives the evidence in favor of H_1 relative to both H_0 and H_u, BES2 gives the evidence for H_0 relative to both H_1 and H_u, and BESu gives the evidence for H_u relative to both H_1 and H_0.

```
sum_of_products <- prod(PMP1) + prod(PMP0) + prod(PMPu)

BES1 <- prod(PMP1)/sum_of_products
BES0 <- prod(PMP0)/sum_of_products
BESu <- prod(PMPu)/sum_of_products
```

The results show that also when we test H_1, H_0 and H_u simultaneously, H_1 (aggregated PMP = .995) clearly receives more support than both the null hypothesis (aggregated PMP < .001) and the unconstrained hypothesis (aggregated PMP = .005).

[4] Note that for multiple-parameter hypotheses (i.e., hypotheses that involve more than one effect, such as two main effects and an interaction), *bain* requires the variance–covariance matrix of the parameter estimates rather than just the variance of each estimate.

Finally, to illustrate that different types of hypotheses can be formulated and tested, we now show with the code below how to evaluate a range-constrained hypothesis (i.e., H_1: 0.5 < SMD < 0.8) against its complement (not H_1). H_1 now tests a more specific hypothesis than before: Whereas before we only tested whether DLD had a negative effect on auditory verbal statistical learning, we now test whether this is a *medium* negative effect.

```
# study-specific PMPs
PMP1c <- c()
for(i in 1:length(ES)){
  SMD <- ES[i]
  names(SMD) <- "SMD"
  res <- bain(SMD,
       hypothesis = "0.5 < SMD < 0.8",
       n = N[i],
       Sigma = var[i])
  PMP1c[i] <- res$fit["H1","PMPc"]
}

# aggregated PMP
BES1c <- prod(PMP1c)/(prod(PMP1c) + prod(1-PMP1c))
```

In line with the conclusion from the meta-analysis, there is compelling evidence that there is a medium negative effect of DLD on statistical learning when this hypothesis is tested against its complement (aggregated PMP > .99).

To facilitate the use of BES for new users, we have converted the code in this empirical demonstration into a wrapper function called bes(), which can be downloaded from our OSF page (https://osf.io/gbtyk/). Note, however, that this is not a general function for BES, as it can only handle single-parameter hypotheses; it uses *bain* and *bain*'s default priors for the parameter estimates to compute the study-specific PMPs; and it assumes equal prior model probabilities.

Discussion

In the current study we introduced Bayesian evidence synthesis as a flexible alternative to meta-analysis for situations in which meta-analysis is difficult or impossible. When the set of studies a researcher wishes to combine is heterogeneous in terms of research design, participant characteristics or operationalization of key variables, it may not be possible to combine these studies by means of meta-analysis. In these situations, BES can be applied to investigate which hypothesis receives the most aggregated support across studies. As explained in the Introduction, the main advantage of BES is that it poses less constraints on the studies to be combined, as support for each hypothesis is first estimated in each study separately. This means that as long as each study provides independent evidence for the same overarching theory, BES allows for differences in the parameter estimates across studies and for study-specific hypotheses that include design and data characteristics unique to that study. Additional benefits of BES are that it allows for (i) testing multiple hypotheses simultaneously, and (ii) formulating and testing informative hypotheses that, unlike the conventional null hypothesis (H_0: no effect) and its complement (not H_0), can directly test a specific theory or expectation. BES comes with the disadvantage, however, that unlike meta-analysis, it is only concerned with hypothesis testing and therefore does not allow for quantifying the effect size or the level of heterogeneity among effect sizes across studies. In addition, given that BES is still a relatively novel technique, no methods currently exist within the context of BES to deal with dependent effect sizes, assess and correct for publication bias, or test study-level predictors of degree of evidence across studies (analogous to meta-regression). Further developing BES is part of our research agenda, so future developments of the method may address some of these issues.

The goal of our simulation study was to compare the performance of BES and meta-analysis to illuminate under which conditions the two methods behave similarly and under which conditions their results diverge. Results were expected to sometimes differ, given that BES and meta-analysis answer a slightly different synthesis question: Whereas meta-analysis indicates whether the target hypothesis is supported by the pooled data, BES indicates the hypothesis that best describes each study. The results showed that in most scenarios, BES behaves similar to meta-analysis, in that the acceptance rate of the correct hypothesis increases with larger sample sizes and more studies. The two main exceptions were (i) when all individual studies were underpowered and (ii) when the true parameter value was on the boundary of the tested hypotheses. We will now discuss both situations in turn.

When all individual studies are underpowered, testing against the null or the unconstrained hypothesis can be problematic. As underpowered studies each provide more evidence for the null hypothesis than for the hypothesis of interest, BES only aggravates this issue as combining more studies will then increase our confidence that the null hypothesis best describes each individual study. Similarly, we saw that if studies are underpowered and there is large variation in study-specific parameter values, the unconstrained hypothesis will typically receive more evidence than the hypothesis of interest as the unconstrained hypothesis always describes the data well, whereas the hypothesis of interest will then provide a poor fit to at least some of the studies. This is an important limitation of BES, as underpowered studies are prevalent in the field of experimental psychology and one of the goals of synthesizing multiple studies may therefore be to reduce power issues. On the other hand, we found that including one underpowered study hardly impacted the results, even when combining as few as three studies. This suggests that power issues are mainly a problem when all or most studies are underpowered, but future studies should further investigate this.

When testing a target hypothesis against its complement when the true parameter value is on the boundary of these hypotheses, BES can yield strong support for either hypothesis (in the simulation, the target hypothesis was accepted in approximately 50% of the simulation replications). This issue extends beyond BES, as strong support for either hypothesis can already be found within a single study. However, BES does not help in solving this issue. Three potential solutions are discussed by Volker (2022), namely testing against an equality-constraint hypothesis (e.g., H_0) instead of or in addition to the complement, evaluating a hypothesis with a boundary on the minimum relevant effect size, or testing against both the complement and the unconstrained hypothesis in turn and see if both render support for (or against) the target hypothesis. We are currently still investigating which solutions work best in this scenario.

An additional finding was that testing against the unconstrained hypothesis was the most sensitive to the cut-off value used for accepting the hypothesis of interest (cf. Figs. 1, 2, 3 and 4 to Figs. S1-S4 on https://osf.io/gbtyk/). Because the PMP testing a hypothesis H_i against the unconstrained hypothesis H_u within a single study has an upper limit that is determined by the complexity of H_i, multiple studies will need to be combined to reach an *aggregated* PMP that is above this upper limit, even when all studies provide a perfect fit for H_i. This is not the case when testing against the null or complement hypothesis, as these PMPs do not have an upper limit. This once again shows the importance of considering both the fit and complexity of the hypotheses under consideration when interpreting the results. It also shows that it may be inappropriate to use the same cut-off values across different (sets of) hypotheses to decide what constitutes "strong" evidence for a hypothesis.

Regarding this last point, it is also important to note that using any kind of cut-off value for the aggregated PMP may give way to publication bias and questionable research practices in the same way as has been described for p values (e.g., Ioannidis, 2005; John et al., 2012; Simmons et al., 2011). In this study, we only used a cut-off value so that we could compute a common performance measure for meta-analysis and BES. However, when conducting BES, using a cut-off value is not strictly necessary because PMPs are interpretable by themselves. For a single study, a PMP of .90 means there is a conditional error probability of 10% that one of the other considered hypotheses is more appropriate (Hoijtink et al., 2019). The interpretation of an *aggregated* PMP is a bit more complicated, however, as different scenarios can result in the same aggregated PMP. For instance, when half of the studies support one hypothesis and half of the studies support the alternative hypothesis to the same degree (e.g., with study-specific PMPs of .90 vs .10), this leads to an aggregated PMP of .50 for both hypotheses. However, if all study-specific PMPs are .50 for both hypotheses, this also leads to an aggregated PMP of .50. In both cases we would conclude that neither hypothesis is a good description of all studies, but in the first case this is because the support for the hypotheses varies across studies, whereas in the second scenario there is no strong support for either hypothesis in *any* study. For this reason, it is important to always interpret the aggregated PMP in relation to the study-specific PMPs (cf. Kevenaar et al., 2021). If the study-specific PMPs show large variation, researchers may then try to explain this variation based on study characteristics. Of course, researchers may already expect variation in study-specific PMPs based on certain study characteristics *a priori* and may therefore not be very interested in the global support for a hypothesis. In that case, researchers could opt to perform multiple Bayesian evidence syntheses on specific subsets of studies.

Finally, we would like to point out a few limitations of the current study and provide some suggestions for future research. Because we wanted to show when BES results diverge from what researchers may expect based on meta-analysis, we focused on simple situations in which both methods are feasible. Therefore, we only considered single-parameter hypotheses and evaluated only two hypotheses at a time, and we did not vary data characteristics across studies such as different operationalizations of key variables. However, some of these factors were investigated by Volker (2022), who showed how the performance of BES varies as a function of the complexity of the study-specific hypotheses by varying the operationalization of key variables across studies. Volker also investigated how BES behaves for single- versus multiple-parameter hypotheses and when hypotheses are only partially correct (rather than either correct or incorrect). A second limitation is that we only looked at the acceptance rate as a performance measure. We did this to have a comparable performance measure for both BES and meta-analysis, but for BES it is more insightful to look at the value of the aggregated PMP directly, as this may show more nuanced differences than what we were able to show here. However, we refer the reader to Klugkist and Volker (2023) and Volker (2022) for partially similar simulation conditions where the authors did report the value of the aggregated BFs/PMPs. Future research could test how the number of considered hypotheses affects the BES results and develop guidelines for interpreting the value of the aggregated PMP that take the complexity of the tested hypotheses into account. Moreover, BES would greatly benefit from methods that allow for testing study-level predictors of degree of support and from methods that can handle dependent effect sizes. Finally, future studies may further explore possible solutions to deal with underpowered studies in the context of BES.

Recommendations

Based on this study, we provide the following recommendations for potential users of BES:

- BES can be considered as a possible alternative to meta-analysis if model estimates across studies cannot be converted into a comparable measure, or if heterogeneity in the study designs or samples calls into question whether the effect sizes test the same underlying true effect, and, therefore, whether these effect sizes can be meaningfully aggregated or compared. Secondly, BES can be considered if the researcher wishes to evaluate an informative hypothesis directly (rather than only rejecting the null hypothesis or not), or if there are more than two competing hypotheses and the researcher wishes to evaluate all relevant hypotheses simultaneously.
- As BES provides joint support for a hypothesis relative to the other hypotheses considered, it is important to include all plausible hypotheses. It is also generally recommended to either include the unconstrained or the complement hypothesis to avoid choosing between hypotheses that do not represent the data well (i.e., the best hypothesis out of a set of bad hypotheses is still a bad hypothesis).
- When (some of) the studies are underpowered, it is important to be aware that BES is not a data-pooling approach and thus, does not increase power by aggregating studies. When equality-constrained hypotheses (e.g., H_0) are deemed sufficiently unlikely, one can evaluate the informative hypothesis of interest against its complement; the comparison that is least affected by low power.
- When interpreting the strength of support for a hypothesis, it is important to acknowledge that the complexity of the hypotheses under consideration impact the degree of (study-specific) support for each hypothesis, as well as the number of studies needed to achieve a certain level of aggregated support (see Volker, 2022). More generally, it is advised to always examine and report the individual study results in addition to the aggregated results from BES.
- If researchers are primarily interested in study-level factors that affect the support for a given hypothesis, then there are currently two options: (i) decide *a priori* which study-level factors are expected to impact the support for the hypotheses of interest and then perform BES on specific subgroups of studies, or (ii) try to detect patterns post hoc by carefully examining the study-specific PMPs.

On a final note, we would like to stress once more that BES is a relatively novel technique that is still under development. Some of the limitations mentioned in this paper may therefore be addressed by future developments of the method. Likewise, as we further investigate the behavior of BES under different circumstances, the current recommendations may be adjusted.

Acknowledgements We thank Nikola Sekulovski for his comments on our description of the Bayes factor, and we thank Sharon Unsworth and Chantal van Dijk for their comments on the understandability of a previous version of this manuscript from the perspective of applied researchers. We also thank Imme Lammertink for kindly allowing us to use her meta-analysis as an empirical example in our paper. The authors did not receive support from any organization for the submitted work.

Authors' contributions All three authors contributed to this manuscript. EvW designed the study, conducted the analyses, wrote a first draft of the manuscript, and revised the manuscript. MZZ and IK supervised the study and contributed to the revision. All authors approved the final version.

Open Practices Statement

In accordance with the Peer Reviewers' Openness Initiative (https://opennessinitiative.org, Morey et al., 2016), all simulated data, code, and supplementary figures associated with this manuscript were available during the review process and remain available on OSF project "Bayesian evidence synthesis" at https://osf.io/gbtyk/. None of the analyses were preregistered.

References

Berkey, C. S., Hoaglin, D. C., Mosteller, F., & Colditz, G. A. (1995). A random-effects regression model for meta-analysis. *Statistics in Medicine, 14*(4), 395–411. https://doi.org/10.1002/sim.4780140406

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365–376. https://doi.org/10.1038/nrn3475

Cooper, H. M., Hedges, L. V., & Valentine, J. C. (Eds.). (2019). *The handbook of research synthesis and meta-analysis* (3rd ed.). Russell Sage Foundation.

Evans, J. L., Hughes, C., Hughes, D., Jackson, K., & Fink, T. (2010, June). *SLI - A domain specific or domain general implicit learning deficit? Modality-constrained statistical learning of auditory*

and perceptual motor sequences in SLI. Poster presented at the symposium on research in child language disorders, Madison, WI.

Evans, J. L., Saffran, J. R., & Robe-Torres, K. (2009). Statistical learning in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 52*(2), 321–335. https://doi.org/10.1044/1092-4388(2009/07-0189)

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Garner, P., Hopewell, S., Chandler, J., MacLehose, H., Schünemann, H. J., Akl, E. A., …, & Panel for updating guidance for systematic reviews (PUGs). (2016). When and how to update systematic reviews: Consensus and checklist. *BMJ*, i3507. https://doi.org/10.1136/bmj.i3507

Grunow, H., Spaulding, T. J., Gómez, R. L., & Plante, E. (2006). The effects of variation on learning word order rules by adults with and without language-based learning disabilities. *Journal of Communication Disorders, 39*, 158–170. https://doi.org/10.1016/j.jcomdis.2005.11.004

Gu, X., Hoijtink, H., Mulder, J., Lissa, C. J. van, Camiel, V. Z., Jones, J., & Waller, N. (2020). *bain: Bayes Factors for Informative Hypotheses* (0.2.4). https://CRAN.R-project.org/package=bain

Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology, 71*(2), 229–261. https://doi.org/10.1111/bmsp.12110

Haebig, E., Saffran, J., & Weismer, S. (2017). Statistical word learning in children with autism spectrum disorder and specific language impairment. *The Journal of Child Psychology and Psychiatry, 58*(11), 1251–1263. https://doi.org/10.1111/jcpp.12734

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*(4), 486–504. https://doi.org/10.1037/1082-989X.3.4.486

Hoijtink, H. J. A. (2012). *Informative Hypotheses*. Chapman and Hall/CRC.

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods, 24*(5), 539–556. https://doi.org/10.1037/met0000201

Hsu, H. J., Tomblin, J. B., & Christiansen, M. H. (2014). Impaired statistical learning of non-adjacent dependencies in adolescents with specific language impairment. *Frontiers in Psychology, 5*, 1–10. https://doi.org/10.3389/fpsyg.2014.00175

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jeffreys, H. (1961). *Theory of probability*. Clarendon.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association, 90*(430), 773–795. https://doi.org/10.2307/2291091

Kevenaar, S. T., Zondervan-Zwijnenburg, M. A. J., Blok, E., Schmengler, H., Fakkel, M (Ties)., de Zeeuw, E. L., …, & Oldehinkel, A. J. (2021). Bayesian evidence synthesis in case of multi-cohort datasets: An illustration by multi-informant differences in self-control. *Developmental Cognitive Neuroscience, 47*, 100904. https://doi.org/10.1016/j.dcn.2020.100904

Klaassen, F., Zedelius, C. M., Veling, H., Aarts, H., & Hoijtink, H. (2018). All for one or some for all? Evaluating informative hypotheses using multiple N = 1 studies. *Behavior Research Methods, 50*(6), 2276–2291. https://doi.org/10.3758/s13428-017-0992-5

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods, 10*, 477–493. https://doi.org/10.1037/1082-989X.10.4.477

Klugkist, I., Van Wesel, F., & Bullens, J. (2011). Do we know what we test and do we test what we want to know? *International Journal of Behavioral Development, 35*(6), 550–560. https://doi.org/10.1177/0165025411425873

Klugkist, I., & Volker, T. B. (2023). Bayesian evidence synthesis for informative hypotheses: An introduction. *Psychological Methods.* https://doi.org/10.1037/met0000602

Kuiper, R. M., Buskens, V., Raub, W., & Hoijtink, H. (2013). Combining statistical evidence from several studies: A method using Bayesian updating and an example from research on trust problems in social and economic exchange. *Sociological Methods & Research, 42*(1), 60–81. https://doi.org/10.1177/0049124112464867

Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2017). Statistical learning in Specific Language Impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research, 60*(12), 3474–3486. https://doi.org/10.1044/2017_JSLHR-L-16-0439

Linden, A. H., & Hönekopp, J. (2021). Heterogeneity of research results: A new perspective from which to assess and promote progress in psychological science. *Perspectives on Psychological Science, 16*(2), 358–376. https://doi.org/10.1177/1745691620964193

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Sage Publications Inc.

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology, 51*(3), 485–504. https://doi.org/10.1002/ejsp.2752

Lukács, Á., & Kemény, F. (2014). Domain-general sequence learning deficit in specific language impairment. *Neuropsychology, 28*(3), 472–483. https://doi.org/10.1037/neu0000052

Mayor-Dubois, C., Zesiger, P., Van der Linden, M., & Roulet-Perez, E. (2014). Nondeclarative learning in children with specific language impairment: Predicting regularities in the visuomotor, phonological, and cognitive domains. *Child Neuropsychology, 20*(1), 1–9. https://doi.org/10.1080/09297049.2012.734293

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., …, & Zwaan, R. A. (2016). The Peer Reviewers' Openness Initiative: Incentivizing open research practices through peer review. *Royal Society Open Science, 3*(1), 150547. https://doi.org/10.1098/rsos.150547

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*(1), 105. https://doi.org/10.1037/1082-989X.7.1.105

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

R Core Team. (2021). *R: A Language and Environment for Statistical Computing* (4.1.0). https://www.R-project.org/

Raudenbush, S. W. (2009). Analyzing effect sizes: Random effects models. In L. V. Cooper & J. C. Hedges (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 295–315). Russell Sage Foundation.

Regenwetter, M., Cavagnaro, D. R., Popova, A., Guo, Y., Zwilling, C., Lim, S. H., & Stevens, J. R. (2018). Heterogeneity and parsimony in intertemporal choice. *Decision, 5*(2), 63–94. https://doi.org/10.1037/dec0000069

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Torkildsen, J. v. K. (2010, November). *Event-related potential correlates of artificial grammar learning in preschool children with specific*

language impairment and controls. Poster presented at the Second Annual Neurobiology of Language Conference, San Diego, CA.

van Assen, M. A. L. M., Stoevenbelt, A. H., & van Aert, R. C. M. (2022). The end justifies all means: Questionable conversion of different effect sizes to a common effect size measure. *Religion, Brain & Behavior, 13*(3), 345–347. https://doi.org/10.1080/2153599X.2022.2070249

van Calster, B., Steyerberg, E. W., Collins, G. S., & Smits, T. (2018). Consequences of relying on statistical significance: Some illustrations. *European Journal of Clinical Investigation, 48*(5), e12912. https://doi.org/10.1111/eci.12912

van Erp, S., Verhagen, J., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in *Psychological Bulletin* from 1990–2013. *Journal of Open Psychology Data, 5*(1), 4. https://doi.org/10.5334/jopd.33

van Houwelingen, H. C., Arends, L. R., & Stijnen, T. (2002). Advanced methods in meta-analysis: Multivariate approach and meta-regression. *Statistics in Medicine, 21*(4), 589–624. https://doi.org/10.1002/sim.1040

van de Schoot, R., Hoijtink, H., & Romeijn, J.-W. (2011). Moving Beyond Traditional Null Hypothesis Testing: Evaluating Expectations Directly. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00024

Veldkamp, S. A. M., Zondervan-Zwijnenburg, M. A. J., van Bergen, E., Barzeva, S. A., Tamayo-Martinez, N., Becht, A. I., …, & Hartman, C. (2021). Parental age in relation to offspring's neurodevelopment. *Journal of Clinical Child & Adolescent Psychology, 50*(5), 632–644. https://doi.org/10.1080/15374416.2020.1756298

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics, 30*(3), 261–293. https://doi.org/10.3102/10769986030003261

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3). https://doi.org/10.18637/jss.v036.i03

Volker, T. B. (2022). *Combining support for hypotheses over heterogeneous studies with Bayesian Evidence Synthesis: A simulation study* [Unpublished master's thesis, Utrecht University]. Retrieved February 6, 2024, from https://github.com/thomvolker/bes_master_thesis_ms/blob/main/manuscript/manuscript_volker.pdf

Zondervan-Zwijnenburg, M. A. J., Richards, J. S., Kevenaar, S. T., Becht, A. I., Hoijtink, H. J. A., Oldehinkel, A. J., …, & Boomsma, D. I. (2020a). Robust longitudinal multi-cohort results: The development of self-control during adolescence. *Developmental Cognitive Neuroscience, 45*, 100817. https://doi.org/10.1016/j.dcn.2020.100817

Zondervan-Zwijnenburg, M. A. J., Veldkamp, S. A. M., Neumann, A., Barzeva, S. A., Nelemans, S. A., Beijsterveldt, C. E. M., …, & Boomsma, D. I. (2020b). Parental age and offspring childhood mental health: A multi-cohort, population-based investigation. *Child Development, 91*(3), 964–982. https://doi.org/10.1111/cdev.13267