Behavior Research Methods (2020) 52:2020–2030
https://doi.org/10.3758/s13428-020-01382-8

A group-specific prior distribution for effect-size heterogeneity in meta-analysis

Christopher G. Thompson[1] · Betsy Jane Becker[2]

Published online: 10 March 2020

Abstract
While both methodological and applied work on Bayesian meta-analysis have flourished, Bayesian modeling of differences between groups of studies remains scarce in meta-analyses in psychology, education, and the social sciences. On rare occasions when Bayesian approaches have been used, non-informative prior distributions have been chosen. However, more informative prior distributions have recently garnered popularity. We propose a group-specific weakly informative prior distribution for the between-studies standard-deviation parameter in meta-analysis. The proposed prior distribution incorporates a frequentist estimate of the between-studies standard deviation as the noncentrality parameter in a folded noncentral t distribution. This prior distribution is then separately modeled for each subgroup of studies, as determined by a categorical factor. Use of the new prior distribution is shown in two extensive examples based on a published meta-analysis on psychological interventions aimed at increasing optimism. We compare the folded noncentral t prior distribution to several non-informative prior distributions. We conclude with discussion, limitations, and avenues for further development of Bayesian meta-analysis in psychology and the social sciences.

Keywords Bayesian meta-analysis · Prior distribution · Between-studies heterogeneity

Meta-analysis is a statistical tool for combining results from sets of related studies (Glass, 1976). Beyond presenting overall features of the effect sizes (e.g., central tendency, variability), a critical component of meta-analysis is the exploration of effect-size heterogeneity. All collections of effects exhibit sampling error because effects are based on sample data. However, sampling error alone rarely accounts for all effect-size heterogeneity. Other sources of variation are often present, including random error (which is not directly explainable), systematic error due to moderator(s), or both.

Several approaches are available for the exploration of systematic error in meta-analysis. The most popular are "meta-regression" methods (see Thompson & Higgins, 2002). Meta-regressions specify effect sizes as outcomes and can incorporate both continuous and categorical moderators (see Berkey, Hoaglin, Mosteller, & Colditz, 1995; Greenland, 1987). Another class of heterogeneity-investigation methods relies on weighted analogues to ANOVA (Hedges, 1982). Such ANOVA-like methods again use effect sizes as outcomes and attempt to explain possible systematic error using categorical moderators. ANOVA-like models can either incorporate group-specific between-studies heterogeneity components for subgroups of effects or use a single heterogeneity estimate that is common to all subgroups of studies. For a comparison of these frequentist pooled- versus separate-variance approaches, see Rubio-Aparicio, Sánchez-Meca, López-López, Botella, and Marín-Martínez (2017).

The advantages of Bayesian methods for meta-analysis are well known (e.g., Lewis & Nair, 2015; Sutton & Abrams, 2001). Bayesian modeling allows for a more flexible and transparent means to incorporate and estimate uncertainty. Furthermore, Bayesian-analysis results are typically presented and described in terms of (marginal) posterior distributions. Such distributions allow for direct statements of probability about results, something that frequentists may desire but are restricted in making.

One disadvantage of Bayesian techniques is their possible sensitivity to the choice of prior distributions. In certain circumstances, the choice of prior distribution can have a large effect on

✉ Christopher G. Thompson
cgthompson@tamu.edu

[1] Department of Educational Psychology, Texas A&M University, College Station, TX, USA

[2] Department of Educational Psychology & Learning Systems, Florida State University, Tallahassee, FL, USA

the respective posterior distribution, and thus on inferences about parameters. Furthermore, computations and model creation in Bayesian analyses often use specialized software, and require some knowledge of computer code and programming on the part of the meta-analyst. However, with the development of software such as Stan (Carpenter et al., 2017), and Bayesian packages in existing software such as Stata (StataCorp, 2017), this disadvantage is becoming less of a worry.

In this paper, we explore a Bayesian method for ANOVA-like modeling in meta-analysis which uses weakly informative prior distributions for heterogeneity parameters. The choice of heterogeneity parameter is the standard deviation. Each group-specific between-studies standard deviation will have its own data-driven prior distribution. Each prior is a folded noncentral *t* distribution with a frequentist estimate of the residual between-studies standard deviation as its noncentrality parameter.

The paper proceeds as follows. First, we provide background on group-specific modeling and informative prior distributions for meta-analysis. Next, we present a Bayesian model for the analogue to ANOVA, and an overview of the folded noncentral *t* distribution. We then propose the new prior distribution for Bayesian meta-analytic ANOVA. This is followed by two in-depth examples of use of the new prior distribution, based on data from a meta-analysis on psychological interventions for increasing optimism (Malouff & Schutte, 2017). Our examples are chosen with the applied psychological or social-science researcher in mind. Two different categorical factors are examined: participant payment (yes or no) and timing of post-intervention assessment (immediate or delayed). Both examples include comparisons of results based on our priors to results based on four non-informative prior distributions. We conclude with discussion, limitations, and avenues for further development of Bayesian meta-analysis.

Group-specific modeling in meta-analysis

Though to date it has been somewhat rare in practice, the use of group-specific modeling in meta-analysis has several advantages. Prior experience or expectations about a research domain might suggest *a priori* that subgroups of studies should differ in terms of their within-group variability. For example, studies of low and high quality have been found to exhibit different degrees of variability (e.g., Schulz, Chalmers, Hayes, & Altman, 1995). Furthermore, allowing for between-groups differences in variation may provide better model fit, regardless of whether it was anticipated or not. That is, partitioning studies and allowing for different variances for each group may reveal true differential heterogeneity. Last, descriptive statistics for the study results or possibly exploratory graphics (e.g., a forest plot) may suggest the appropriateness of group-specific modeling. This latter scenario involves a preliminary assessment of the effect sizes before making a modeling decision, which is inherently different from choosing a model based on theoretical reasons. However, this approach suffers from being subject to a researcher's choice of descriptive statistics and may lend itself to "data shopping" (i.e., seeking specific patterns or values).

In both frequentist and Bayesian meta-analysis, group-specific modeling of within-group variance has received little attention. This may have resulted from the widespread use of meta-regression, which does not easily allow for group-specific variances. The typical frequentist approach uses the same point estimate of heterogeneity for all individual subgroups (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2009).

Group-specific modeling in Bayesian meta-analysis is more complex than a frequentist approach because prior distributions are required for all group-specific parameters, such as means and standard deviations, or means and variances. Larose and Dey (1997) and Prevost, Abrams, and Jones (2000) have used Bayesian group-specific models in medical settings. In the earlier of two works, Larose and Dey (1997) used "grouped random-effects models" to synthesize 15 studies on the anti-epileptic drug, Progabide. Studies were partitioned into two groups: studies that were double blinded and those that were single blinded. Using both fully Bayesian estimation and data-driven prior distributions, each group received separate prior distributions for means and between-studies variances.

Prevost, Abrams, and Jones applied a similar approach, which they termed "group-specific heterogeneity," to a set of studies on breast-cancer screening. Studies were again partitioned into two groups: randomized controlled trials and observational studies. Prevost et al. (2000) argued for this partition by stating that observational studies are likely to be more biased than randomized control trials, leading to increased variability among studies. Furthermore, for another Bayesian approach, see Röver, Wandel, and Friede (2019) which focuses on the use of mixture priors. Our approach to group-specific Bayesian modeling uses weakly informative prior distributions, rooted in the work of Gelman (2006).

Informative prior distributions for meta-analysis

For any Bayesian analysis, the choice of prior distribution(s) is a key decision. Non-informative prior distributions, such as the half normal with a large variance or the uniform, are common for heterogeneity parameters in meta-analysis. Such prior distributions "…aim at attenuating the impact of the prior on the resulting inference" (Marin & Robert, 2014, p. 35). One example of comparing non-informative priors to frequentist

estimators of the between-studies standard deviation can be found in Bodnar, Link, Arendacká, Possolo, and Elster (2017). However, current practice is shifting towards the use of more informative prior distributions for scale parameters (e.g., Pullenayegum, 2011; Rhodes, Turner, & Higgins, 2015; Rhodes et al., 2016; Steel, Kammeyer-Mueller, & Paterson, 2015; Turner, Davey, Clarke, Thompson, & Higgins, 2012; Turner, Jackson, Wei, Thompson, & Higgins, 2015); see Williams, Rast, and Bürkner (2018) for an overview. Some authors argue for data-driven prior distributions created from observed between-studies variability found in previously published meta-analyses or meta-analytic databases (e.g., the Cochrane Database of Systematic Reviews). For instance, Turner et al. (2012) extracted 14,886 meta-analyses from the Cochrane Database of Systematic Reviews and created what they call "off-the-shelf" prior distributions for several common medical settings. As another approach for creating an informative prior distribution, Rhodes et al. (2016) implemented data augmentation (Tanner & Wong, 1987) to create a conjugate prior distribution characterized by "pseudo data" (in contrast to, and then combined with, observed data).

Previous approaches are not without limitations. The sets of studies used to develop such priors may have little to do with the target collection of studies. Even if similar studies exist, the number of studies available to create suitable prior distributions may be small. This constraint can hinder meta-analysts who want to combine studies in fields with only a limited amount of related research to create an informative prior distribution.

Notation, model, and folded noncentral *t* distribution

We begin with a collection of K independent effect sizes. These K effect sizes are partitioned into $J \geq 2$ disjoint groups based on some categorical factor, X (e.g., short and long interventions). As both examples in our study (described later) analyze treatment effects by comparing two groups, methods shown here used the standardized-mean-difference effect size. The kth effect-size estimate in the jth group is represented as d_{jk}, where $k = 1, \ldots, K_j$ and $j = 2, \ldots, J$. The total number of effect sizes is $K = \sum_j K_j$. Under a random-effects model, each d_{jk} estimates its respective, though not necessarily unique, effect-size parameter, θ_{jk}, with mean μ_j. Last, each effect size has a within-study variance, v_{jk}.

To obtain posterior distributions for the between-studies standard deviations for J groups, we use a conditional group-specific random-effects (CGRE) model. The CGRE model includes a categorical factor, X, as well as allows group means (μ_j) to vary. This is also applied to different between-studies standard deviations across groups, which we denote as τ_j for group j. Distributional assumptions at the sample effect-size level are $d_{jk} \mid \theta_{jk}, v_{jk} \sim \pi_A(\theta_{jk}, v_{jk})$. At the effect-size parameter level, $\theta_{jk} \mid \mu_j, \tau_j \sim \pi_B(\mu_j, \tau_j)$. Here, $\pi(\bullet)$ represents a choice of distributional form. Subscripts for $\pi(\bullet)$ (e.g., $\pi_A(\bullet)$) denote that, if there are more than one prior distribution, they need not be the same parametric form. The specific required parameters (e.g., shape and scale for a Normal distribution or location for a uniform) vary by choice of distributional form. In practice and this paper we model the between-studies variability as a standard deviation (τ_θ) rather than as a variance (τ_θ^2) at the recommendation of previous work (e.g., Gelman, 2006) and for interpretation purposes.

Effect sizes and variances for the CGRE model are represented as $\mathbf{D}_{1k} = (d_{jk}, v_{jk})$, with the complete observed dataset $\mathbf{D}^* = (\mathbf{D}_{11}, \ldots, \mathbf{D}_{1K_1}, \mathbf{D}_{21}, \ldots, \mathbf{D}_{2K_2}, \ldots, \mathbf{D}_{J1}, \ldots, \mathbf{D}_{JK_J})$ and the study-level effect-size parameter vector $\boldsymbol{\theta}^* = (\theta_{11}, \ldots, \theta_{1K_1}, \theta_{21}, \ldots, \theta_{2K_2}, \ldots, \theta_{J1}, \ldots, \theta_{JK_J})$. The CGRE model has multiple means and between-studies standard deviations, both specific to each group of effects. The vector of mean parameters is $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_J)$, and we define the vector of between-studies standard deviations as $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_J)$. The vector of all unknown parameters in the CGRE model is $\boldsymbol{\gamma}^* = (\boldsymbol{\theta}^*, \boldsymbol{\mu}, \boldsymbol{\tau})$. Assuming conditional independence among all unknown parameters, the joint posterior distribution of $\boldsymbol{\gamma}^*$ is

$$p(\boldsymbol{\gamma}^* | \mathbf{D}^*) \propto p(\mathbf{D}^* | \boldsymbol{\gamma}^*) p(\boldsymbol{\gamma}^*) = \prod_{j=1}^{J} \prod_{k=1}^{K_j} \left[p(d_{jk} | \theta_{jk}, v_{jk}) p(\theta_{jk} | \mu_j, \tau_j) \right] \prod_{j=1}^{J} \left[p(\mu_j) p(\tau_j) \right]. \quad (1)$$

For Eq. (1), $2J$ hyper prior distributions must be specified.

Folded noncentral distribution

The noncentral *t* distribution (Johnson, Kotz, & Balakrishnan, 1995) has several common uses in statistics (e.g., power analysis, distributions of standardized mean differences), including in meta-analysis (Becker, 1988; Camilli, de la Torre, & Chiu, 2010; Hedges, 1981). The noncentral *t* distribution is defined as a function of other distributions. If $Z \sim N(0, 1)$ and $W \sim \chi^2(\psi)$, then a random variable, G, follows a noncentral *t* distribution with a noncentrality parameter φ and ψ degrees of freedom, where

$$G = \frac{Z + \varphi}{\sqrt{W/\psi}}. \quad (2)$$

The possibly less familiar folded noncentral *t* distribution (FNT) is related to the noncentral *t* distribution. Using Eq. (2), the FNT is expressed as $H = |G|$. When $\varphi = 0$, H is a centered half *t* distribution. When $\varphi > 0$ the FNT "folds over"

negatively valued mass on the horizontal axis to the positive side of the distribution. This is inherently different from the similar truncated noncentral *t* distribution, which discards any negatively valued mass. Gelman (2006) proposes to use the FNT rather than non-informative prior distributions on scale parameters (e.g., the inverse-gamma distribution with equivalent scale and shape parameters) due to its better performance (e.g., convergence). In the next section, we argue for a group-specific prior distribution for the between-studies standard deviation using the FNT distribution.

Folded noncentral *t* distribution for meta-analysis

We propose using the FNT as a weakly informative prior distribution for the between-studies standard deviation, τ_j, in group-specific Bayesian meta-analysis. Specifically, the prior distribution for each τ_j is a FNT distribution with a noncentrality parameter equal to $\widehat{\tau}_{\theta|X}$ and degrees of freedom ψ, denoted as FNT($\widehat{\tau}_{\theta|X}, \psi$). Here, $\tau_{\theta\,|\,X}$ represents the remaining unexplained effect-size heterogeneity after accounting for sampling error and any heterogeneity explained by the moderator, denoted as *X*. Put another way, given a categorical moderator, *X*, the residual between-studies standard deviation that is considered common across *K* studies is $\tau_{\theta\,|\,X}$. The method presented in this paper uses $\tau_{\theta\,|\,X}$ to inform our prior assumptions about the distribution for each group-specific τ_j. Common choices for the prior distribution for τ_j would be the inverse gamma or uniform distributions. By using a FNT distribution for each prior distribution on τ_j we are expressing the belief that a more suitable assumption for τ_j places the most weight in the neighborhood of $\widehat{\tau}_{\theta|X}$ instead of, for example, close to zero (e.g., with the inverse-gamma distribution), or equally across the parameter space of τ_j (e.g., with the uniform distribution). Such *a priori* arguments for placing the most weight of the prior density at zero without foundation are unrealistic. By placing the greatest mass of the prior distributions in the neighborhood of $\widehat{\tau}_{\theta|X}$, this method uses information from the complete set of *K* effect sizes to provide possibly more realistic specifications of individual τ_j prior distributions. Applying this prior requires a two-step process. First, an estimate of $\tau_{\theta\,|\,X}$ must be computed. The second step uses $\widehat{\tau}_{\theta|X}$ as the noncentrality parameter in the FNT distribution with degrees of freedom that are flexible.

A comprehensive simulation of the FNT approach to group-specific modeling was completed by Thompson (2016). Conditions in Thompson (2016) included prior distributions (half-Cauchy, half-normal, inverse gamma, and uniform – the same that are used here), number of groups within the categorical variable (2 or 3), number of effects within a group (6 or 36), within-study sample size (50 or 500), degree of within-group heterogeneity (varied by number of groups), and degrees of freedom for the FNT (2, 3, 4, or 30). General findings included that the choice of degrees of freedom for the FNT was not critical. Results for conditions with three groups were essentially extensions of results with two groups. As one might expect, more statistical noise was present in estimated densities when the within-study sample size was 50 compared to 500. That being said, the number of studies was more critical to the shape and smoothness than was the within-study sample size.

Model specification

We compare the performance of the FNT prior for τ_j to those of four non-informative prior distributions used in meta-analysis. In addition to the FNT, other choices of priors for τ_j were a) the half-Cauchy as demonstrated in Gelman (2006), b) the half-normal with large variance, c) the inverse-gamma with small and identical shape and scale parameters, and d) the uniform with a wide range. Distributions at the sample effect-size level (d_{jk}) and effect-size parameter level (θ_{jk}) were treated as normal, and the hyper prior distribution on μ_j was a non-informative normal distribution with large variance. Writing the model in hierarchical form,

$$\begin{aligned} d_{jk}|\theta_{jk}, v_{jk} &\sim \mathrm{N}\left(\theta_{jk}, v_{jk}\right), \\ \theta_{jk}|\mu_j, \tau_j &\sim \mathrm{N}\left(\mu_j, \tau_j\right), \\ \mu_j &\sim \mathrm{N}(0, 100), \text{ and} \\ \tau_j &\sim \pi(\bullet). \end{aligned} \tag{3}$$

We now need only to specify the hyper priors on τ_j in Eq. (3). In both examples, the grouping moderator comprises two groups, thus priors for τ_1 and τ_2 need to be specified. The five priors used for $\pi(\bullet)$ are

1. $\tau_j \sim \mathrm{FNT}\left(\widehat{\tau}_{\theta|X}, 4\right)$ (Folded Noncentral *t*),
2. $\tau_j \sim$HC(25) (Half-Cauchy),
3. $\tau_j \sim$HN(10) (Half-Normal),
4. $\tau_j \sim$IG(.001, .001) (Inverse-Gamma), and
5. $\tau_j \sim$U(0, 10) (Uniform).

As previously shown via simulation (Thompson, 2016), the choice of degrees of freedom ψ for FNT($\widehat{\tau}_{\theta|X}, \psi$) is not critical. As such, the degrees of freedom for the FNT prior distribution here are arbitrarily set to $\psi = 4$.

Model assessment and software

For Malouff and Schutte's (2017) meta-analysis examples shown below, several numerical quantities and graphics were examined. First, MCMC convergence is assessed using autocorrelations, the Gelman-Rubin diagnostic (Gelman & Rubin, 1992), running mean plots, and trace plots. Although results for τ_j are of main interest here, convergence should be assessed for all model parameters. All of our models were

run with three chains, each with 500,000 iterations and a 100,000 burn-in period.

Model assessment can be completed in two ways: using individual model fit and model-to-model comparisons across several competing models (i.e., different prior distributions for τ_j). For each model several descriptive statistics from the marginal posterior distributions for τ_j should be inspected, including the median, mean, standard deviation, 2.5th percentile, and 97.5th percentile. Furthermore, the 95% highest posterior density (HPD) intervals for τ_j may be particularly informative. The HPD interval can be considered as the smallest credible interval of a posterior distribution which covers a prespecified $100(1-\alpha)\%$ range. Because we used three chains when estimating the posterior density, for each τ_j we obtained three HPD upper bounds and three HPD lower bounds, with one pair for each chain. In all conditions, these bounds were virtually identical across chains, thus we only report results for the first chain in each case.

When comparing fit across models we used the posterior predictive check (PPC). The PPC (denoted p^*) is a quantity that compares the fit (similarity) of data simulated from a candidate model to the observed data. The PPC ranges from 0 to 1, with an optimal value being $p^* = .50$.

Examples

We demonstrate the use of this new prior distribution with two examples. Both examples draw on Malouff and Schutte's (2017) meta-analysis on psychological interventions for increasing optimism. The effect size is the standardized mean difference, with positive values favoring the intervention group. We examine two grouping variables. The first is time of post-intervention assessment (within 1-day post intervention or more than 1-day post intervention) and the second is whether participants were paid for their participation or not. These examples were chosen based on their differing group-specific between-studies standard deviations. The original Malouff and Schutte (2017) article did not report between-studies standard deviations, but we were able to compute these estimates using their reported effect sizes. For the post-intervention-assessment-time example, the first group of effects (within 1 day post intervention) has an estimated between-studies standard deviation of $\widehat{\tau}_{G1} = 0.17$ and the second group of effects (more than 1 day post intervention) has an estimated between-studies standard deviation of $\widehat{\tau}_{G2} = 0.14$. Thus, both groups have similar and small between-studies standard-deviation estimates. For the participant-pay example, the first group of effects (for unpaid participants) has a small estimated between-studies standard deviation of $\widehat{\tau}_{G1} < .001$, and the second group of effects (from paid participants) has an estimated between-studies standard deviation of $\widehat{\tau}_{G2} = 0.34$. These group-specific between-studies standard-deviation estimates differ more in magnitude, with $\widehat{\tau}_{G1}$ small and $\widehat{\tau}_{G2}$ being moderately large.

All analyses were completed using a combination of R (Microsoft R Open, 2017) and JAGS (Plummer, 2003) via the *rjags* package (Plummer, 2016). Non-Bayesian estimates of $\widehat{\tau}_{\theta|X}$ were obtained using restricted maximum likelihood estimation in the *metafor* package (Viechtbauer, 2010). Forest plots were created using the *meta* package (Schwarzer, 2007).

Example 1: Time-of-assessment moderator

Of the 29 effect sizes in Malouff and Schutte (2017), 23 effects came from studies where the final assessment was administered within 1 day of the conclusion of training (Group 1); the other 6 effects were from studies that administered later final assessments (Group 2). Patterns of effects by group are shown in Fig. 1. The original meta-analysis reported a statistically significant test of group differences in mean effects ($Q(1) = 27.2, p = .025$). The random-effects mean for Group 1 was $\overline{d}_{G1} = 0.46$ and for Group 2 it was $\overline{d}_{G2} = 0.22$, both being statistically significant. As stated before, each group showed some evidence of within-group heterogeneity. The between-studies standard-deviation estimates, $\widehat{\tau}_{G1} = 0.17$ and $\widehat{\tau}_{G2} = 0.14$, suggest that the groups have similar amounts of within-group heterogeneity.

Next, we compare descriptive statistics for τ_{G1} and τ_{G2} across the five choices of prior distributions. The median, mean, SD, 2.5th and 97.5th percentiles (Table 1) and HPD intervals (Table 2) were computed from marginal posterior distributions (when convergence was reached). The model that used the inverse-gamma prior distribution was unable to converge, despite several additional steps to achieve convergence (e.g., increased number of iterations, larger burn-in period), thus we have four models to compare (folded noncentral *t*, half-Cauchy, half normal, and uniform).

Many descriptive statistics were either identical (up to two decimal places) or very similar across choices of prior distributions (for both τ_{G1} and τ_{G2}). Given that both groups had similar between-studies standard-deviation estimates (recall $\widehat{\tau}_{G1} = 0.17$ and $\widehat{\tau}_{G2} = 0.14$, also in Table 1), it is not surprising that both the marginal posterior medians and means are similar as well. In contrast, the SDs and 2.5th and 97.5th percentiles show differences in the overall variability of the marginal posterior distributions, and particularity the length of the right tail (i.e., the 97.5 percentile). These were noticeability larger for the later assessment group (G2) than for the immediate assessment group (G1). This difference in variability and the width of the marginal posterior may stem from the difference in sample sizes, $K_{G1} = 23$ and $K_{G2} = 6$, as the former provides more information (and precision) than the latter. The HPD intervals (Table 2) showed similarity across

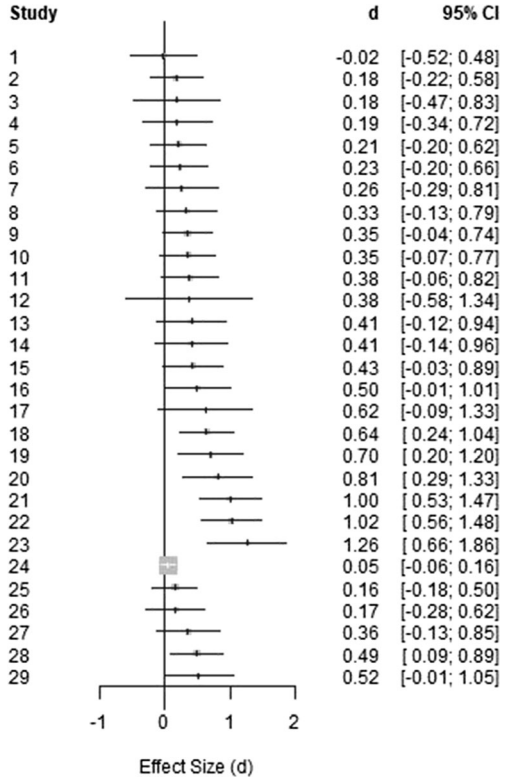

Fig. 1 Forest plot of effects ordered by timing of the assessment and effect size. Studies 1–23 applied an assessment within 1 day after the end of training and studies 24–29 applied an assessment at some point more than 1 day after the end of training

choice of prior distribution and differences between the two groups, and Group 2 always had a wider HPD interval than Group 1. The marginal posteriors for $\widehat{\tau}_{G1}$ and $\widehat{\tau}_{G2}$ using the FNT prior condition are shown in Fig. 2. The overall shape of the marginal densities similar between groups (e.g., peak between $\tau = 0.1$ and $\tau = 0.2$ with a long right-side tail). However, more noise is visible in the lower values of τ (close to zero) for $\widehat{\tau}_{G1}$ compared to $\widehat{\tau}_{G2}$.

Model fit was similar among the four prior-distribution choices that did converge. All PPC results were $p^* = .49$ (with values close to $p^* = .50$ being preferable), indicating good predictive capability.

Example 2: Participant-payment moderator

Again using the 29 effect sizes in Malouff and Schutte (2017), 18 effects came from studies where participants were not paid (Group 1) and the remaining 11 effects were from studies that did pay their participants (Group 2). Confidence intervals for effects by group are shown in Fig. 3. Malouff and Schutte (2017) reported a non-significant between-groups test ($Q(1) = 1.4, p = .23$), indicating that participant payment status did not significantly explain the overall effect-size variability. The random-effects mean effect for the unpaid group was $\overline{d}_{G1} = 0.36$, and for those who were paid was $\overline{d}_{G2} = 0.53$, both being statistically significant. Studies that paid participants showed slightly (but not significantly) larger intervention effects.

We are interested in the within-group variation, regardless of the group means. Between-studies standard-deviation estimates appear quite distinct, with $\widehat{\tau}_{G1} < 0.001$ for studies of unpaid participants, and $\widehat{\tau}_{G2} = 0.34$ for studies with paid participants. This pattern is in contrast to the variances of the timing-of-assessment groups previously discussed. Studies with unpaid participants were very homogeneous compared

Table 1 Descriptive statistics for τ_j marginal posteriors – time-of-assessment moderator

Prior distribution	K_j	Median	Mean	SD	2.5th percentile	97.5th percentile	$\widehat{\tau}_j$
Folded Noncentral *t*							
Assessment ≤ 1 Day of End of Training	23	0.17	0.17	0.09	0.01	0.35	0.17
Later Assessment	6	0.18	0.20	0.15	0.02	0.57	0.14
Half-Cauchy							
Assessment ≤ 1 Day After End of Training	23	0.17	0.17	0.09	0.01	0.35	0.17
Later Assessment	6	0.18	0.21	0.16	0.02	0.62	0.14
Half-Normal							
Assessment ≤ 1 Day After End of Training	23	0.17	0.17	0.09	0.01	0.35	0.17
Later Assessment	6	0.18	0.21	0.16	0.02	0.61	0.14
Inverse Gamma							
Assessment ≤ 1 Day After End of Training	23	—	—	—	—	—	0.17
Later Assessment	6	—	—	—	—	—	0.14
Uniform							
Assessment ≤ 1 Day After End of Training	23	0.17	0.17	0.09	0.01	0.35	0.17
Later Assessment	6	0.18	0.22	0.17	0.02	0.63	0.14

Table 2 HPD Intervals for τ_j – time-of-assessment moderator

Prior distribution	K_j	Lower bound	Upper bound
Folded Noncentral *t*			
Assessment ≤ 1 Day After End of Training	23	0.00	0.32
Later Assessment	6	0.00	0.47
Half-Cauchy			
Assessment ≤ 1 Day After End of Training	23	0.00	0.32
Later Assessment	6	0.00	0.51
Half-Normal			
Assessment ≤ 1 Day After End of Training	23	0.00	0.32
Later Assessment	6	0.00	0.50
Inverse Gamma			
Assessment ≤ 1 Day After End of Training	23	—	—
Later Assessment	6	—	—
Uniform			
Assessment ≤ 1 Day After End of Training	23	0.00	0.32
Later Assessment	6	0.00	0.50

to studies that *did* pay participants. This scenario, with two groups of effects that clearly have differing degrees of within-group effect-size variability, is a prime example for which group-specific modeling may be informative and beneficial.

As with the first example, we compare the descriptive statistics for τ_{G1} and τ_{G2} across the prior distributions. The inverse-gamma prior distribution again did not converge, hence we again have four comparison priors (folded noncentral *t*, half-Cauchy, half normal, and uniform). Medians, means, SDs, 2.5th and 97.5th percentiles are in Table 3 and HPD intervals are given in Table 4.

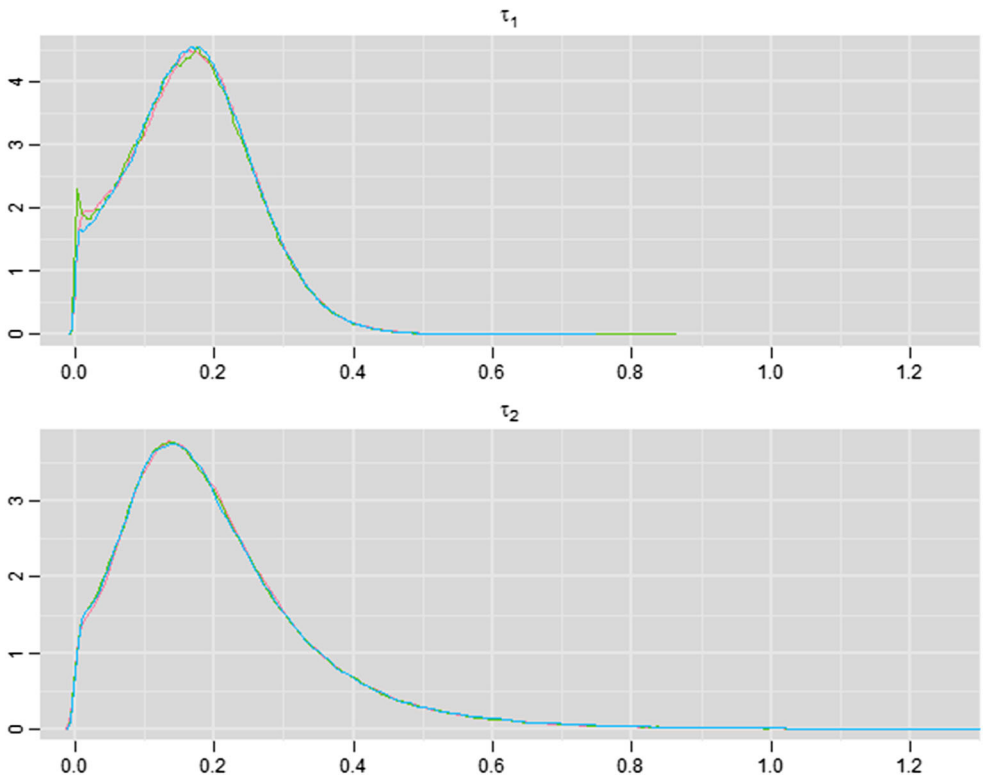

Fig. 2 Marginal posterior distributions of between-studies standard deviations for the time-of-assessment moderator. The *top cell* (τ_1) is the plot for the "assessment ≤ 1 day after end of training" group. The *bottom cell* (τ_2) is the plot for the "later assessment" group. Different color densities represent MCMC-generated chains

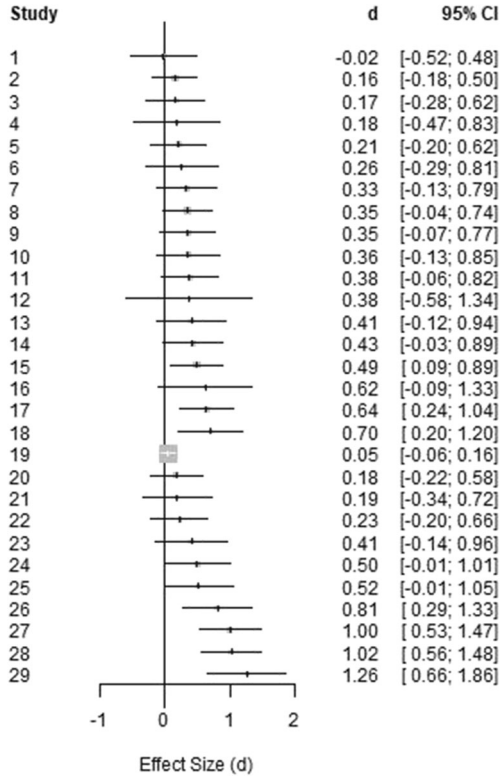

Fig. 3 Forest plot of effects ordered by whether participants were paid and effect size. Participants in studies 1–18 were paid for their participation but others in studies 19–29 were not

Descriptive statistics (Table 3) again were similar across the four prior distributions for both τ_{G1} and τ_{G2}. The medians, means, SDs, and 2.5th and 97.5th percentiles were comparable among prior distributions. This was also the case for the HPD intervals (Table 4) – all HPD intervals were of similar length across prior-distribution choices. Furthermore, the HPD intervals showed almost no overlap between the two groups (unpaid participants and paid participants) for all prior distribution choices, supporting a likely difference in within-group effect-size heterogeneity.

The marginal posteriors for $\widehat{\tau}_{G1}$ and $\widehat{\tau}_{G2}$ using the FNT prior condition are shown in Fig. 4. In contrast to the timing-of-assessment moderator, overall shape of the marginal densities differed between Group 1 (no payment participant) and Group 2 (participants received payment). For Group 1, the marginal posterior of $\widehat{\tau}_{G1}$ has a peak slightly above zero, followed by a step downward slope and elongated tail.

Table 4 HPD intervals for τ_j – participant-payment moderator

Prior distribution	K_j	Lower bound	Upper bound
Folded Noncentral *t*			
Participants Paid: No	18	0.00	0.17
Participants Paid: Yes	11	0.18	0.66
Half-Cauchy			
Participants Paid: No	18	0.00	0.18
Participants Paid: Yes	11	0.18	0.68
Half-Normal			
Participants Paid: No	18	0.00	0.18
Participants Paid: Yes	11	0.17	0.68
Inverse Gamma			
Participants Paid: No	18	—	—
Participants Paid: Yes	11	—	—
Uniform			
Participants Paid: No	18	0.00	0.18
Participants Paid: Yes	11	0.18	0.68

Table 3 Descriptive statistics for τ_j marginal posteriors – participant-payment moderator

Prior Distribution	K_j	Median	Mean	SD	2.5th percentile	97.5th percentile	$\widehat{\tau}_j$
Folded Noncentral *t*							
Participants Paid: No	18	0.06	0.07	0.06	0.00	0.21	< 0.001
Participants Paid: Yes	11	0.37	0.39	0.13	0.20	0.71	0.34
Half-Cauchy							
Participants Paid: No	18	0.06	0.07	0.06	0.00	0.21	< 0.001
Participants Paid: Yes	11	0.38	0.40	0.14	0.20	0.74	0.34
Half-Normal							
Participants Paid: No	18	0.06	0.07	0.06	0.00	0.21	< 0.001
Participants Paid: Yes	11	0.38	0.40	0.14	0.20	0.74	0.34
Inverse Gamma							
Participants Paid: No	18	—	—	—	—	—	< 0.001
Participants Paid: Yes	11	—	—	—	—	—	0.34
Uniform							
Participants Paid: No	18	0.06	0.07	0.06	0.00	0.21	< 0.001
Participants Paid: Yes	11	0.38	0.40	0.14	0.20	0.74	0.34

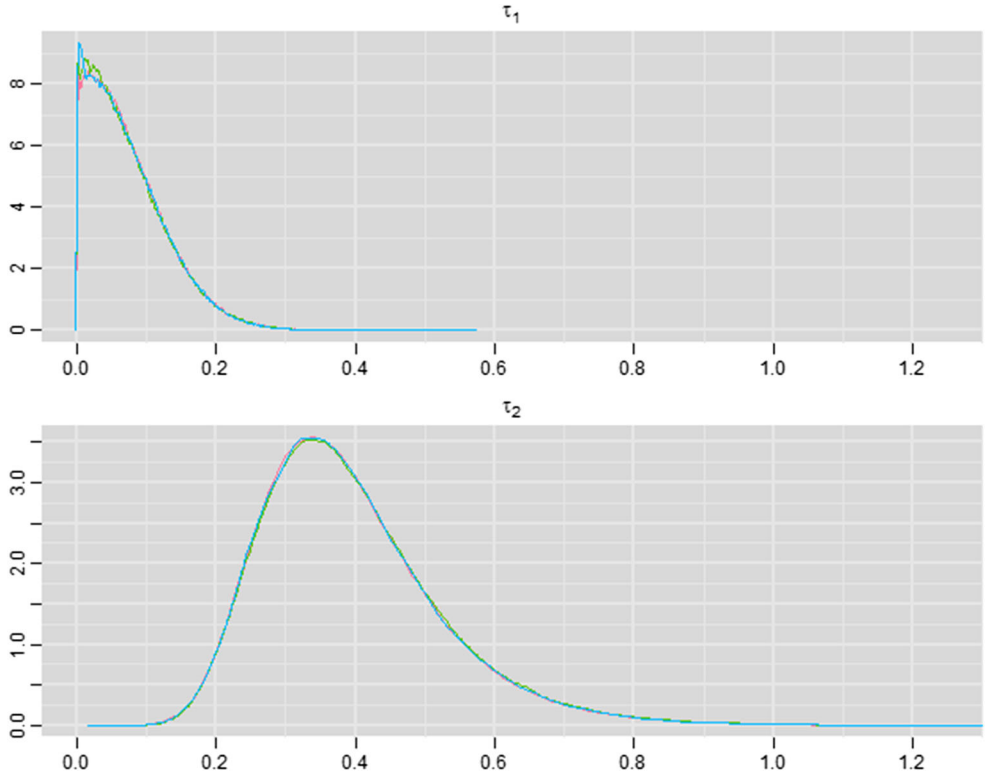

Fig. 4 Marginal posterior distributions of between-studies standard deviation for the participant-payment moderator. The *top cell* (τ_1) is the plot for the "no participant payment" group. The *bottom cell* (τ_2) is the plot for the "participants received payment" group. Different color densities represent MCMC-generated chains

Looking at the marginal posterior of $\widehat{\tau}_{G2}$, we see a more normal density with a slight right skew. The peak of the posterior is between $\tau = 0.3$ and $\tau = 0.4$ The stark contrast between these two densities does align with the descriptive statistics described above.

All PPC results were $p^* = .16$ (again, values close to $p^* = .50$ being preferable), indicating fairly poor capability. This may be at least a partial result of the challenge of finding a suitable prior for a subgroup with a small variance or standard-deviation parameter (as was very likely the case with $\tau_{G1} < 0.001$).

Discussion

One of the most contentious parts of Bayesian data analysis, including Bayesian meta-analysis, is choosing the prior distributions. The main purpose of this paper was to present a new prior distribution, the folded noncentral *t*, for modeling between-studies heterogeneity in Bayesian meta-analysis. After a review of existing weakly-informative prior distributions for Bayesian meta-analysis (including those that are group-specific based on some categorical factor), the new prior distribution was presented and its implementation thoroughly explained. Our example used data from a recent meta-analysis on interventions aimed at increasing a person's optimism, and showed that all but one prior distribution converged and produced reasonably similar estimates.

Though current Bayesian meta-analysis practice mostly uses fully Bayesian (i.e., data independent) prior distributions, the prior distribution presented here is data driven, similar to the log-Cauchy and log-logistic distributions which are partial functions of the harmonic mean of the individual study effect-size variances (see DuMouchel, 1994; Larose & Dey, 1997). The noncentrality parameter in the folded noncentral *t* distribution is specified using data from the set of studies, specifically $\widehat{\tau}_{\theta|X}$ which reflects the impact of a specific moderator. This is not the case when using a function of harmonic means of individual study effect-size variances, as for the log-logistic and log-Cauchy priors.

As has been previously stated, the folded noncentral *t*, with $\widehat{\tau}_{\theta|X}$ as its noncentrality parameter for the distributions of the τ_j in Eq. (3), produced similar results compared to three fully Bayesian prior distributions, the half-Cauchy, half-Normal, and uniform. The inverse gamma prior distribution case failed to converge as previously discussed. The results we examined included central-tendency measures (means and medians), standard deviations, upper and lower tail regions, HPD intervals, and two model-fit indices. The differences were minor and would have little-to-no practical impact on inferences and

interpretations. With that being said, in line with Lambert, Sutton, Burton, Abrams, and Jones (2005) we recommend using several competing prior distributions, and see the FNT as one good competitor distribution. In most instances, meta-analysts seek to achieve robustness by assessing several prior distributions. Showing that the choice of prior distribution does not influence posterior-distribution descriptive statistics and model fit provides substantive evidence that the posterior distribution is not entirely subject to the choice of prior distribution.

Limitations

We do acknowledge limitations of our work. First, the examples presented here were chosen for their clear link to the field of behavioral research. Although the chosen examples provide several realistic features such as unbalanced within-group numbers of studies and within-group between-studies variability, we do not attempt to generalize all results in this paper to all other fields of research or meta-analytic situations. This includes, but is not limited to, within-study and between-studies sample sizes and group-specific weighted means. For the latter point, one similar avenue of further research would be to formulate and assess group-specific weakly-informative prior distributions for within-group means (instead of within-group between-studies standard deviations, as demonstrated here). Last, the FNT is a trivariate distribution (noncentrality parameter, degrees of freedom, and scale parameter), as described in Gelman (2006). In our work, we chose to set the scale parameter to unity and manipulate the noncentrality parameter. This decision was based on the goal of developing an empirically driven prior distribution (i.e., $\widehat{\tau}_{\theta|X}$). Another area of future research would be assessing different combinations of conditions of the FNT by manipulation of other parameters.

Implications

The use of Bayesian meta-analysis in behavioral research and the social sciences in general is limited, but has recently gained traction (e.g., Kim, Belland, & Walker, 2018; Perret & Bonin, 2019). The advantages of Bayesian meta-analysis compared to a frequentist approach are well understood (e.g., Sutton & Abrams, 2001) and in our examples are seen most clearly in the comparisons of HPD intervals. With recent computational advances, the obstacle of complex Bayesian-model computation is less and less a problem. A call for social-science fields, particularly in behavioral research and education, to adopt Bayesian meta-analysis methods is warranted. Our work here supports that.

Through the examples in this paper we demonstrate how a more nuanced story can be told with Bayesian methodology. Particularly for between-studies heterogeneity, applying the group-specific Bayesian model provided some evidence that 1) Using a single point estimate to represent the between-studies standard-deviation parameter is strongly inadvisable for the participant-payment moderator, and 2) Once the between-studies standard deviation was modeled as a random variable, its posterior distribution was seen to be quite wide, particularly for the smaller of the two groups.

One particularly strong advantage of Bayesian meta-analysis methods is in the interpretation of results. Quantitative behavioral research is traditionally filled with null-hypothesis statistical testing, including p values, confidence intervals, and similar quantities. Results are often misinterpreted using direct probability statements (Goodman, 2008). We do not claim that Bayesian meta-analysis methods are "better" than non-Bayesian meta-analysis methods, rather the two approaches provide two different sets of tools. When used correctly, Bayesian meta-analysis provides more flexibility, allows for direct probabilistic statements, and increases transparency about the uncertainty in our estimates.

Author Notes This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. We declare that there is no conflict of interest with respect to the research, authorship, and/or publication of this work.

References

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology*, *41*(2), 257–278.

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.

Bodnar, O., Link, A., Arendacká, B., Possolo, A., & Elster, C. (2017) Bayesian estimation in random effects meta-analysis using a non-informative prior. *Statistics in Medicine*, *36*(2), 378–399.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Subgroup analyses. *Introduction to meta-analysis* (pp. 149–186). West Sussex: John Wiley & Sons.

Camilli, G., de la Torre, J., & Chiu, C. Y. (2010). A noncentral *t* regression model for meta-analysis. *Journal of Educational and Behavioral Statistics*, *35*(2), 125–153.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., …, Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32.

DuMouchel, W. (1994). *Hierarchical Bayes linear models for meta-analysis* (tech. rep. No. 27). National Institute of Statistical Sciences.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Data Analysis*, *1*(3), 515–534.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3–8.

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology*, *45*, 135–140.

Greenland, S. (1987). Quantitative methods in the review of epidemiologic literature. *Epidemiologic Reviews*, *9*(1), 1–30.

Hedges, L. V. (1981). Distribution theory of Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128.

Hedges, L. V. (1982). Fitting categorical models to effect sizes from a series of experiments. *Journal of Educational Statistics*, *7*(2), 119–137.

Johnson, N. L., Kotz, S., & Balakrishnan, B. (1995). Noncentral *t*-distributions. *Continuous univariate distributions* (pp. 508–544; 2nd ed.). New York: John Wiley & Sons.

Kim, N. J., Belland, B. R., & Walker, A. E. (2018). Effectiveness of computer-based scaffolding in the context of problem-based learning for stem education: Bayesian meta-analysis. *Educational Psychology Review*, *30*(2), 397–429.

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., & Jones, D. R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, *24*(15), 2401–2428.

Larose, D. T., & Dey, D. K. (1997). Grouped random effects models for Bayesian meta-analysis. *Statistics in Medicine*, *16*(16), 1817–1829.

Lewis, M. G., & Nair, N. S. (2015). Review of applications of Bayesian meta-analysis in systematic reviews. *Global Journal of Medicine and Public Health*, *4*(1), 1–9.

Malouff, J. M., & Schutte, N. S. (2017). Can psychological interventions increase optimism? A meta-analysis. *The Journal of Positive Psychology*, *12*(6), 594–604.

Marin, J. M., & Robert, C. P. (2014). Bayesian essentials with R (2nd ed.). New York: Springer.

Microsoft R Core Team (2017). *Microsoft R open*. Redmond, WA: Microsoft. Retrieved from https://mran.microsoft.com

Perret, C., & Bonin, P. (2019). Which variables should be controlled for to investigate picture naming in adults? A Bayesian meta-analysis. *Behavior Research Methods, 51*(6), 2533–2545.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (dsc 2003)*.

Plummer, M. (2016). *Rjags: Bayesian graphical models using MCMC*. R package version 4–6. Retrieved from https://CRAN.R-project.org/package=rjags

Prevost, T. C., Abrams, K. R., & Jones, D. R. (2000). Hierarchical models in generalized synthesis of evidence: An example based on studies of breast cancer screening. *Statistics in Medicine*, *19*(24), 3359–3376.

Pullenayegum, E. M. (2011). An informed reference prior for between-study heterogeneity in meta-analyses of binary outcomes. *Statistics in Medicine*, *30*(26), 3082–3094.

Rhodes, K. M., Turner, R. M., & Higgins, J. P. T. (2015). Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *Journal of Clinical Epidemiology*, *68*(1), 52–60.

Rhodes, K. M., Turner, R. M., White, I. R., Jackson, D., Spiegelhalter, D. J., & Higgins, J. P. T. (2016). Implementing informative priors for heterogeneity in meta-analysis using meta-regression and pseudo data. *Statistics in Medicine*, *35*(29), 5495–5511.

Röver, C., Wandel, S., & Friede, T. (2019). Model averaging for robust extrapolation in evidence synthesis. *Statistics in Medicine*, *38*(4), 674–694.

Rubio-Aparicio, M., Sánchez-Meca, J., López-López, J. A., Botella, J., & Marín-Martínez, F. (2017). Analysis of categorical moderators in mixed-effects meta-analysis: Consequences of using pooled versus separate estimates of residual between-studies variances. *British Journal of Mathematical and Statistical Psychology*, *70*(3), 439–456.

Schulz, K. F., Chalmers, I., Hayes, R. J., & Altman, D. G. (1995). Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of the American Medical Association*, *273*(5), 408–412.

Schwarzer, G. (2007). meta: An R package for meta-analysis. *R News*, *7*(3), 40–45.

StataCorp. (2017). *Stata Statistical Software: Release 15*. College Station: StataCorp LLC

Steel, P., Kammeyer-Mueller, J., & Paterson, T. A. (2015). Improving the meta-analytic assessment of effect size variance with an informed Bayesian prior. *Journal of Management*, *14*(1), 718–743.

Sutton, A. J., & Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, *10*(4), 277–303.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*(398), 528–540.

Thompson, C. G. (2016). *A weakly-informative group-specific prior distribution for meta-analysis* (Unpublished doctoral dissertation). Florida State University. Retrieved from http://purl.flvc.org/fsu/fd/FSU_2016SP_Thompson_fsu_0071E_13051

Thompson, S. G., & Higgins, J. P. T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*, *21*(11), 1559–1573.

Turner, R. M., Davey, J., Clarke, M. J., Thompson, S. G., & Higgins, J. P. T. (2012). Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. *International Journal of Epidemiology*, *41*(3), 818–827.

Turner, R. M., Jackson, D., Wei, Y., Thompson, S. G., & Higgins, J. P. T. (2015). Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Statistics in Medicine*, *34*(6), 984–998.

Viechtbauer, W. (2010). Conducting meta-analysis in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

Williams, D. R., Rast, P., & Bürkner, P. (2018). Bayesian meta-analysis with weakly informative prior distributions. https://doi.org/10.17605/OSF.IO/7TBRM

Open Practices Statement Data for all analyses in this paper are from a previously published meta-analysis. These data are available in the original meta-analysis.