



RESEARCH ARTICLE

Estimation of an overall standardized mean difference in random-effects meta-analysis if the distribution of random effects departs from normal

María Rubio-Aparicio¹ | José Antonio López-López²  | Julio Sánchez-Meca¹  |
Fulgencio Marín-Martínez¹ | Wolfgang Viechtbauer³ | Wim Van den Noortgate⁴

¹Department of Basic Psychology and Methodology, University of Murcia, Murcia, Spain

²School of Social and Community Medicine, University of Bristol, Bristol, UK

³Department of Psychiatry and Neuropsychology, Maastricht University, Maastricht, The Netherlands

⁴Faculty of Psychology and Educational Sciences, University of Leuven, Leuven, Belgium

Correspondence

Julio Sánchez-Meca, Department of Basic Psychology and Methodology, University of Murcia, Murcia, Spain.
Email: jsmecca@um.es

Funding information

Ministerio de Economía y Competitividad of the Spanish Government; Fondo Europeo de Desarrollo Regional (FEDER) Project No. PSI2016-77676-P

The random-effects model, applied in most meta-analyses nowadays, typically assumes normality of the distribution of the effect parameters. The purpose of this study was to examine the performance of various random-effects methods (standard method, Hartung's method, profile likelihood method, and bootstrapping) for computing an average effect size estimate and a confidence interval (CI) around it, when the normality assumption is not met. For comparison purposes, we also included the fixed-effect model. We manipulated a wide range of conditions, including conditions with some degree of departure from the normality assumption, using Monte Carlo simulation. To simulate realistic scenarios, we chose the manipulated conditions from a systematic review of meta-analyses on the effectiveness of psychological treatments. We compared the performance of the different methods in terms of bias and mean squared error of the average effect estimators, empirical coverage probability and width of the CIs, and variability of the standard errors. Our results suggest that random-effects methods are largely robust to departures from normality, with Hartung's profile likelihood methods yielding the best performance under suboptimal conditions.

KEYWORDS

confidence interval, meta-analysis, overall effect size, random-effects model

1 | INTRODUCTION

Meta-analysis is a form of systematic review that allows the integration of the results of a set of primary studies on a given topic by applying statistical methods. When the dependent variable is continuous and the aim of the meta-analysis is to compare the performance between two groups (eg, interventions) across studies, standardized mean differences are the effect size indices most commonly used.^{1,2} This paper focuses on various methods for computing an estimate of the average standardized mean difference together with its confidence

interval (CI) when some assumptions of the underlying statistical model are not met.

Two general statistical models are available for meta-analysis, namely, fixed-effect (FE) and random-effects models. Model choice is crucial, as it determines the statistical procedures used to estimate the mean effect and its CI as well as the generalizability of the meta-analysis results.^{1,3,4}

The FE model assumes that all studies included in the meta-analysis share a common effect parameter such that the only source of variability is sampling error in the selection of participants.⁵ This assumption might apply

if all included studies were similarly designed and conducted and used highly similar samples. In contrast, the random-effects model assumes that each study estimates a different effect parameter. Therefore, the estimation of the overall effect in a random-effects model is affected by sampling error both in the random selection of participants for each study and in the selection of studies.⁶

In this paper, we focused on the performance of the random-effects model, which allows for a broader generalization of results and conclusions and is currently assumed in most meta-analyses.^{3,6}

1.1 | The random-effects model

Let k denote the number of studies included in a meta-analysis and $\hat{\theta}_i$ indicate the effect size estimate from the i th study. The underlying statistical model can be written as follows:

$$\hat{\theta}_i = \theta_i + e_i, \quad (1)$$

where θ_i is the effect parameter for the i th study and e_i is the sampling error of $\hat{\theta}_i$. Usually, e_i is assumed to be normally distributed, ie, $e_i \sim N(0, \sigma_i^2)$, with σ_i^2 as the within-study variance for the i th study.

The random-effects model assumes that the effect parameters θ_i are randomly selected from a population of parameters. Thus, θ_i can be defined as follows:

$$\theta_i = \mu_\theta + \varepsilon_i, \quad (2)$$

where μ_θ is a parameter representing the overall mean of the effect parameters and ε_i denotes the difference between the effect parameter of the i th study θ_i and the overall mean μ_θ . It is assumed that $\varepsilon_i \sim N(0, \tau^2)$, with τ^2 as the between-studies variance. Therefore, combining Equations (1) and (2) enables us to formulate the random-effects model as follows:

$$\hat{\theta}_i = \mu_\theta + e_i + \varepsilon_i, \quad (3)$$

where ε_i and e_i are assumed independent and, as a result, the effect size estimates $\hat{\theta}_i$ are assumed to be normally distributed with mean μ_θ and variance $\sigma_i^2 + \tau^2$, ie, $\hat{\theta}_i \sim N(\mu_\theta, \sigma_i^2 + \tau^2)$.^{6,7}

Although the normality of the distribution of effect parameters is a common assumption in the random-effects model, it might not be realistic or even approximate in a wide range of applied situations including meta-analyses including a small number of studies.⁷⁻¹³ Departures from normality might affect the estimation of key model parameters such as μ_θ and τ^2 . This scenario has important practical implications because a

substantial proportion of the meta-analyses conducted over the last two decades assumed a random-effects model to analyze databases with small-to-moderate numbers of studies. Therefore, assessing the consequences of a violation of the normality assumption constitutes a relevant question in meta-analysis.

To the best of our knowledge, the works of Kontopantelis and Reeves^{11,12} are the only simulation studies that compared the performance of several statistical methods for random-effects meta-analysis under nonnormal scenarios. Eight statistical methods were examined, and a wide range of scenarios was considered. In particular, Kontopantelis and Reeves manipulated the distribution of the effect parameters (normal, skew normal, and *extremely* nonnormal), the number of studies in the meta-analysis, and the heterogeneity. Most methods were found to be highly robust against violations of the assumption of normality. These previous studies focused on the field of epidemiology, and the set of simulated scenarios and outcome measures and the effect size index (odds ratios) were selected accordingly, following the results of a survey of meta-analyses published in the medical field.¹⁴

Furthermore, Kontopantelis and Reeves^{11,12} generated the individual effect estimates using the method for log odds ratios developed by Brockwell and Gordon.⁸ This approach has two major limitations: It is not realistic because it does not start from 2×2 tables,¹⁵ and it is also not appropriate for other effect metrics.

In the current study, we aimed to assess the consequences of violating the normality assumption in random-effects meta-analyses conducted in the psychological field and particularly in meta-analyses on the effectiveness of psychological treatments for various psychological or psychiatric disorders.

In summary, the purpose of our study was to compare the performance of various random-effects meta-analysis methods for the computation of an average effect size and its CI when the normality assumption is not met. For this purpose, a wide range of scenarios was considered, including conditions with some degree of departure from normality. A Monte Carlo simulation was conducted using the standardized mean differences as the effect size index. To avoid the problems in the Kontopantelis and Reeves^{11,12} studies, the standardized mean differences were individually generated in our simulations by assuming a noncentral t distribution.¹⁶ Although our study focused on the random-effects model, the FE model was also included for comparison purposes.

In the following section, we outline the statistical methods considered in this study and describe the residual heterogeneity variance estimators. A simulation study comparing the performance of the methods is detailed.

Finally, a description of the results is presented, and considerations arising from the results are discussed.

1.2 | Methods for estimation of an overall effect size

1.2.1 | FE model

The uniformly minimum variance unbiased estimator of the mean effect size under a FE model is given by the expression.¹⁶

$$\hat{\mu}_{UMVU}^{FE} = \frac{\sum_i w_i^{FE} \hat{\theta}_i}{\sum_i w_i^{FE}}, \quad (4)$$

with

$$w_i^{FE} = 1/\sigma_i^2, \quad (5)$$

and σ_i^2 as the within-study variance of $\hat{\theta}_i$. Since the σ_i^2 are unknown, w_i^{FE} are usually replaced by \hat{w}_i^{FE} based on the estimated within-study variances $\hat{\sigma}_i^2$, as follows:

$$\hat{w}_i^{FE} = 1/\hat{\sigma}_i^2. \quad (6)$$

Thus, in practice, the overall effect size is estimated by the following

$$\hat{\mu}_{FE} = \frac{\sum_i \hat{w}_i^{FE} \hat{\theta}_i}{\sum_i \hat{w}_i^{FE}}. \quad (7)$$

The sampling variance of $\hat{\mu}_{FE}$ is usually estimated as shown:

$$\hat{V}_{FE} = \frac{1}{\sum_i \hat{w}_i^{FE}}. \quad (8)$$

Additionally, a $100(1 - \alpha)\%$ CI for $\hat{\mu}_{FE}$ can be calculated as follows:

$$\hat{\mu}_{FE} \pm z_{1-\alpha/2} \sqrt{\hat{V}_{FE}}, \quad (9)$$

where $z_{1-\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the standard normal distribution.

1.2.2 | Random-effects model

In a random-effects model, the uniformly minimum variance unbiased estimator of μ_θ is given by the following.^{17,18}

$$\hat{\mu}_{UMVU}^{RE} = \frac{\sum_i w_i^{RE} \hat{\theta}_i}{\sum_i w_i^{RE}}, \quad (10)$$

with w_i^{RE} as the optimal weights, defined as $w_i^{RE} = 1/(\sigma_i^2 + \tau^2)$. The variance for $\hat{\mu}_{UMVU}^{RE}$ is given by the formula $V_{UMVU} = 1/\sum_i w_i^{RE}$.

However, σ_i^2 and τ^2 are unknown in practice, and hence, they must be estimated from the studies. The overall mean μ_θ can be estimated using the following equation (11),

$$\hat{\mu}_{RE} = \frac{\sum_i \hat{w}_i^{RE} \hat{\theta}_i}{\sum_i \hat{w}_i^{RE}} \quad (11)$$

where

$$\hat{w}_i^{RE} = 1/(\hat{\sigma}_i^2 + \hat{\tau}^2), \quad (12)$$

where $\hat{\sigma}_i^2$ is the estimated within-study variance of $\hat{\theta}_i$ and $\hat{\tau}^2$ is an estimate of the between-studies variance. Several estimators of the between-studies variance are described in the further section.

In the current study, we compare four alternative random-effects methods to construct a CI around the mean effect size estimate: the standard method (SM), Hartung's method (HM), the profile likelihood (PL) method, and the bootstrapping method.

Standard method

The method most frequently used to obtain a CI around the mean effect size estimate $\hat{\mu}_{RE}$ in a random-effects meta-analysis assumes a normal sampling distribution for $\hat{\mu}_{RE}$. Its sampling variance is usually estimated by the following:

$$\hat{V}_{RE} = \frac{1}{\sum_i \hat{w}_i^{RE}}. \quad (13)$$

Therefore, a $100(1 - \alpha)\%$ CI around $\hat{\mu}_{RE}$ can be computed as shown:

$$\hat{\mu}_{RE} \pm z_{1-\alpha/2} \sqrt{\hat{V}_{RE}}. \quad (14)$$

Hartung's method

Although the SM is the usual procedure for calculating a CI around the mean effect size, this method assumes a normal distribution and does not consider the uncertainty derived from the estimation process of the variance

parameters. As a consequence, the CI based on the z distribution has been shown to yield CIs that are too narrow, resulting in empirical coverage below the nominal level in some scenarios, especially as the between-studies variance increases and the number of studies decreases.⁸ To solve this limitation, Hartung¹⁹ proposed assumption of a t distribution instead of the standard normal distribution and use of an improved variance estimator.^{20,21} A $100(1 - \alpha)\%$ CI for this method is supplied by the expression

$$\hat{\mu}_{RE} \pm t_{k-1; 1-\alpha/2} \sqrt{\hat{V}_{HA}}, \quad (15)$$

where $t_{k-1; 1-\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the t distribution with $k - 1$ degrees of freedom, $\hat{\mu}_{RE}$ is computed by Equation (11), and \hat{V}_{HA} is an estimate of the sampling variance of $\hat{\mu}_{RE}$ with a weighted extension of the usual formula given by

$$\hat{V}_{HA} = \frac{\sum_i \hat{w}_i^{RE} (\hat{\theta}_i - \hat{\mu}_{RE})^2}{(k-1) \sum_i \hat{w}_i^{RE}}. \quad (16)$$

Compared with the standard random-effects method, HM has been found to yield wider CIs with better coverage probabilities, especially under suboptimal scenarios,^{17,22} including scenarios with violation of the normality assumption.¹²

PL method

The PL is an iterative and computationally intensive method that can be used to obtain a likelihood-based CI around an overall estimate obtained with the random-effects model, considering the fact that μ_{θ} and τ^2 must be estimated simultaneously.¹⁰ The PL method provides two alternatives to calculate a CI around $\hat{\mu}_{RE}$, namely, the first-order likelihood method and the higher-order Skovgaard's method. In a simulation study, Guolo²³ showed that the Skovgaard's method produces far more accurate results than the first-order method, especially with small sample sizes. The R code for this method is provided in Data S1.

It is expected that likelihood approaches might improve the performance of standard random-effects methods under nonnormal scenarios.^{10,23,24} Although SMs unrealistically assume that the between-studies variance is known, the likelihood approach allows derivation of the likelihood-based CIs for the between-studies variance and for the overall effect. The iterative and joint estimation of both parameters considers the fact that the other parameters are also unknown and must be estimated.

Bootstrapping

Bootstrapping methods are increasingly applied in the meta-analytic arena if the assumptions of the random-effects model are not met. These methods are free from theoretical distribution assumptions and therefore are expected to be more robust to violations of the normality assumption than standard meta-analytic techniques.^{25,26} In particular, a bootstrapping approach consists of generating a distribution of mean effect size estimates by resampling a large number of samples, eg, 1000 samples.²⁷⁻²⁹ Thus, a 95% CI is given by the 2.5th and 97.5th percentiles of the distribution of mean effect estimates. We examined two methods for the interval estimation of the mean effect size: the percentile method and the bias-corrected and accelerated (BCa) method. The percentile method yields confidence limits that are directly extracted from the percentiles of the distribution. However, the BCa method is preferred in practice because it adjusts for both bias and skewness in the bootstrap distribution.²⁷ See Data S1 for additional computational details.

1.3 | Heterogeneity variance estimators

An estimate of τ^2 is required to obtain the mean effect size estimate and its CI under a random-effects model, at least for the standard and Hartung's approaches. Several methods have been proposed to estimate the between-studies variance τ^2 in random-effects meta-analysis.^{17,18,30} In this section, we present formulas for the three estimators considered in this study.

DerSimonian and Laird estimator

The most commonly used estimator was proposed by DerSimonian and Laird³¹ and is derived from the moments method and computed with the following expression:

$$\hat{\tau}_{DL}^2 = \frac{Q - (k-1)}{c}, \quad (17)$$

where

$$Q = \sum_i \hat{w}_i^{FE} (\hat{\theta}_i - \hat{\mu}_{FE})^2, \quad (18)$$

with $\hat{\mu}_{FE}$ and \hat{w}_i^{FE} defined in Equations (7) and (6), respectively, and c given by the following:

$$c = \sum_i \hat{w}_i^{FE} - \frac{\sum_i (\hat{w}_i^{FE})^2}{\sum_i \hat{w}_i^{FE}}. \quad (19)$$

When $Q < (k - 1)$, $\hat{\tau}_{DL}^2$ is usually set to zero. When the estimated weights \hat{w}_i^{FE} are used instead of the optimal

values, the Q statistic no longer follows the chi-squared distribution usually assumed and this may negatively affect the performance of the $\hat{\tau}_{DL}^2$ estimator.^{32,33}

Restricted maximum likelihood estimator

Another alternative for estimating the between-studies variance component is based on restricted maximum likelihood (REML) estimation. The REML estimator is obtained iteratively from the following.^{17,18}

$$\hat{\tau}_{REML}^2 = \frac{\sum_i (\hat{w}_i^{RE})^2 \left[(\hat{\theta}_i - \hat{\mu}_{RE})^2 - \hat{\sigma}_i^2 \right]}{\sum_i (\hat{w}_i^{RE})^2} + \frac{1}{\sum_i \hat{w}_i^{RE}}, \quad (20)$$

with $\hat{\mu}_{RE}$ and \hat{w}_i^{RE} defined in Equations (11) and (12), respectively, and $\hat{\tau}^2$ initially estimated with any of the noniterative estimators of the heterogeneity variance.

When $\hat{\tau}_{REML}^2 < 0$, it is truncated to zero.

Empirical Bayes estimator

The final estimator of τ^2 that we include is the empirical Bayes (EB) method, which is also an iterative method obtained by replacing $(\hat{w}_i^{RE})^2$ with \hat{w}_i^{RE} in Equation (20) for $\hat{\tau}_{REML}^2$.^{34,35} The EB estimator is obtained as shown:

$$\hat{\tau}_{EB}^2 = \frac{\sum_i \hat{w}_i^{RE} \left[(\hat{\theta}_i - \hat{\mu}_{RE})^2 - \hat{\sigma}_i^2 \right]}{\sum_i \hat{w}_i^{RE}} + \frac{1}{\sum_i \hat{w}_i^{RE}}. \quad (21)$$

Again, negative values of $\hat{\tau}_{EB}^2$ are truncated to zero. The EB estimator is equivalent to the Paule-Mandel estimator.^{30,36}

2 | METHOD OF THE SIMULATION STUDY

In the previous section, we presented two methods for estimating the mean effect size, μ_θ (ie, FE model and standard random-effects model), six methods for computing the CI around an estimate of μ_θ (ie, FE model, standard random-effects model, HM, PL method with higher-order Skovgaard's approach, and bootstrapping with the BCa and percentile methods), and three estimators of τ^2 (ie, the DerSimonian and Laird [DL], REML, and EB estimators) in the context of random-effects meta-analysis. We compared the performance of combinations of these methods using Monte Carlo simulation. However, not all of the methods were combined with each other; in particular, we only combined the PL method with REML estimation and the bootstrapping

method with the DL estimator, whereas the SM and HM were combined with the three τ^2 estimators, and no τ^2 estimators were needed for the FE model. This approach yielded four methods used to estimate the mean effect size and 10 ways to calculate a CI around that estimate.

The simulation was programmed in R using the *metafor*,³⁷ *metaLik*,³⁸ and *boot*³⁹ packages. Data S1 contains the full R code of our simulation study. The standardized mean difference was used as the effect size measure. We simulated designs comparing two groups (experimental and control) with respect to a continuous dependent variable, which is a scenario often found in psychology. Both populations were assumed to be normally distributed with common variance $[N(\mu_E, \sigma^2), N(\mu_C, \sigma^2)]$. For each study, the population standardized mean difference θ was defined as follows.¹⁶

$$\theta = \frac{\mu_E - \mu_C}{\sigma}. \quad (22)$$

In a random-effects model, a distribution of effect parameters θ_i is assumed, with a specific mean μ_θ , heterogeneity variance τ^2 , and shape (details on how the distributions shapes were defined are supplied below). To simulate a meta-analysis, k effect parameters θ_i were randomly selected from the distribution of effect parameters, and an individual parameter θ_i was used in each study.

The effect parameter for the i th study θ_i was estimated using the nearly unbiased estimator proposed by Hedges and Olkin.¹⁶

$$\hat{\theta} = c(m)g, \quad (23)$$

where g is a positively biased estimator computed from the following:

$$g = \frac{\bar{y}_E - \bar{y}_C}{S}, \quad (24)$$

and $c(m)$ is a correction factor for small sample sizes, given by the following:

$$c(m) = 1 - \frac{3}{4m-1}, \quad (25)$$

where $m = n_E + n_C - 2$ and n_E and n_C are the experimental and control group sizes, respectively.

In Equation (24), \bar{y}_E and \bar{y}_C are the sample means of the experimental and control groups, respectively, and S is a pooled standard deviation computed as shown:

$$S = \sqrt{\frac{(n_E-1)S_E^2 + (n_C-1)S_C^2}{n_E + n_C - 2}}, \quad (26)$$

where S_E^2 and S_C^2 are the unbiased variances of the experimental and control groups, respectively.

Equation (23) applies to each study such that $\hat{\theta}_i$ is an estimate of the effect parameter θ_i . The estimates of the sampling variance of $\hat{\theta}$ in each study were obtained by the following:

$$\hat{\sigma}_{\hat{\theta}}^2 = \frac{n_E + n_C}{n_E n_C} + \frac{\hat{\theta}^2}{2(n_E + n_C)}. \quad (27)$$

Hedges and Olkin^{16(p79)} showed that $\sqrt{n_E n_C / (n_E + n_C)} g$ follows a noncentral t distribution with noncentrality parameter $\sqrt{n_E n_C / (n_E + n_C)} \theta$ and $n_E + n_C - 2$ degrees of freedom. The $\hat{\theta}_i$ value for the i th study was simulated from $Z / \sqrt{X/m}$, where Z is a random normal variable with distribution $N(\theta, 1/n_E + 1/n_C)$, and X is a random chi-square variable with $m = n_E + n_C - 2$ degrees of freedom.

When calculating $\hat{\mu}_{FE}$ (Equation (7)) and $\hat{\mu}_{RE}$ (Equation (11)), a potential source of bias is the correlation between the standardized mean difference (Equation (23)) and its sampling variance (Equation (27)), particularly with small sample sizes.

To identify a range of realistic scenarios in this field, the manipulated conditions in the current study were set according to the results of a systematic review of 50 meta-analyses on the efficacy of psychological interventions using three types of standardized mean differences (posttest standardized mean difference, standardized mean change, and standardized mean change difference) as effect size indices.⁴⁰ For the number of studies k , four values were considered, ie, 10, 20, 40, and 60, corresponding to a small-to-large number of studies for the meta-analysis. The overall mean of the distribution of effect parameters μ_{θ} was set to 0, 0.2, 0.5, and 0.8, which reflect conditions of no effect and effects of low, medium, and large magnitude, respectively. Furthermore, a wide range of values for the population between-studies variance τ^2 was considered, namely, 0, 0.03, 0.06, 0.11, 0.18, and 0.39. The simulated conditions for k , μ_{θ} , and τ^2 were within the range of values found in the systematic review of 50 meta-analyses previously mentioned.⁴⁰

The shape of the distribution of the effect parameters θ_i was manipulated through six combinations of the skewness and kurtosis values. First, a normal scenario (ie, zero skewness and kurtosis) was set. Second, five nonnormal conditions were considered based on the results from a previous systematic review.⁴⁰ In that review, the skewness distribution of the 50 meta-analyses presented a median value of 0.52, with 25th and 75th percentiles of 0.18 and 1.1 and minimum and maximum values of -2 and 3.67, respectively. Although the small number of studies in many of those meta-analyses did

not allow accurate estimation of the population skewness and kurtosis, some of the values we found suggest challenging scenarios for random-effects meta-analyses assuming normality. Based on these results, a wide range of skewness values of -2 , -1 , 0 , 1 , and 2 were selected to simulate the effect parameter distribution. The nonlinear relationship exhibited by the 50 pairs of skewness and kurtosis values found in the systematic review was used to predict the kurtosis values. Figure 1 presents the scatter plot relating the skewness and kurtosis values of the 50 meta-analyses. A nonlinear predictive model was fit to this dataset, leading to the predictive equation, $\text{Kurtosis} = -0.581 + 0.023 * \text{Skewness} + 1.069 * \text{Skewness}^2$, and the resulting five nonnormal combinations between skewness and kurtosis values were $(-2, 3.65)$, $(-1, 0.47)$, $(0, -0.58)$, $(1, 0.51)$, and $(2, 3.74)$. Figure 2 presents histograms of the effect parameter distributions for the six simulated combinations of skewness and kurtosis. Data S2 presents five examples of real meta-analyses selected from the previous study⁴⁰ with similar skewness and kurtosis values as each of the five nonnormal scenarios defined in our simulation study. Data S3 presents the individual standardized mean differences and sampling variances of each of the five real meta-analyses.

We applied Fleishman's algorithm⁴¹ to generate distributions of effect parameters with a given mean (μ_{θ}), variance (τ^2), skewness, and kurtosis. In particular, Fleishman's power transformation $X = a + bZ + cZ^2 + dZ^3$ applied on a standard normal distribution $Z \sim N(0, 1)$ allows generation of a nonnormal random variable X with mean 0, variance 1, skewness γ_1 , and kurtosis γ_2 . For a specific combination of γ_1 and γ_2 values, the equations used to

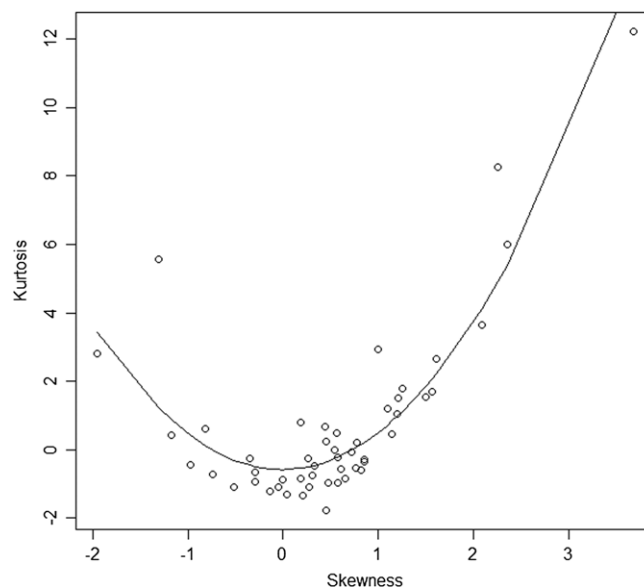


FIGURE 1 Scatter plot of the skewness and kurtosis values found in a systematic review of 50 meta-analyses of on efficacy of psychological interventions⁴⁰

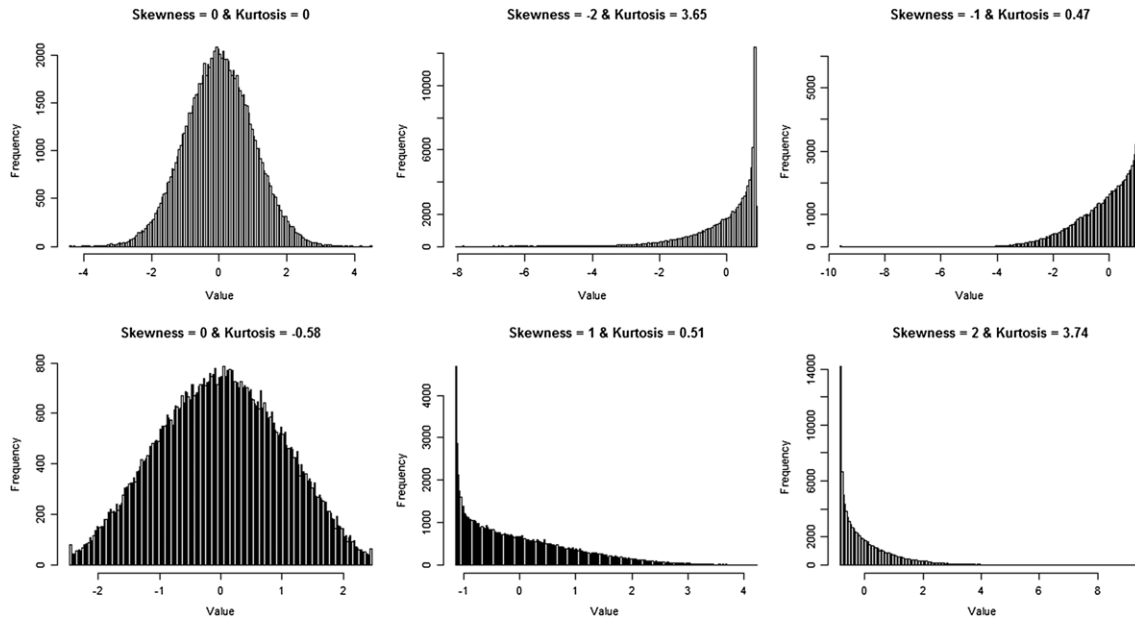


FIGURE 2 Simulated scenarios for the shape of the distribution of effect parameters, assuming $\mu_{\theta} = 0$ and $\tau^2 = 1$

find the a , b , c , and d constants were calculated by solving the equation system presented in Fleishman.^{41(pp522-526)} Table 1 presents the values of a , b , c , and d for the six combinations of γ_1 and γ_2 values in the simulated distributions of the effect parameters. The linear transformation $Y = m + nX$ was subsequently applied to generate distributions with the manipulated values of the mean of the effect parameters ($\mu_{\theta} = 0, 0.2, 0.5$, and 0.8) and the population between-studies variance ($\tau^2 = 0, 0.03, 0.06, 0.11, 0.18$, and 0.39), where $m = \mu_{\theta}$ and $n = \sqrt{\tau^2}$.

Fleishman's algorithm does not yield an exact solution under extreme conditions of skewness and kurtosis.^{41(p526)} Consequently, under the two most extreme conditions in Table 1, ie, $\gamma_1 = -2$, $\gamma_2 = 3.65$ and $\gamma_1 = 2$, $\gamma_2 = 3.74$, the constants a , b , c , and d yielded $\hat{\gamma}_1$ values deviating from the expected values, namely, -1.67 and 1.70 , respectively. Nonetheless, the resulting simulated distributions strongly departed from normality, as intended in our simulation study.

TABLE 1 Values of the a , b , c , and d constants in Fleishman's algorithm for the six combinations of skewness and kurtosis

| Skewness (γ_1) | Kurtosis (γ_2) | a | b | c | d |
|-------------------------|-------------------------|--------|-------|--------|--------|
| 0 | 0 | 0 | 1 | 0 | 0 |
| -2 | 3.65 | 0.349 | 0.862 | -0.349 | -0.018 |
| -1 | 0.47 | 0.267 | 1.124 | -0.267 | -0.071 |
| 0 | -0.58 | 0 | 1.093 | 0 | -0.032 |
| 1 | 0.51 | -0.256 | 1.112 | 0.256 | -0.064 |
| 2 | 3.74 | -0.360 | 0.862 | 0.360 | -0.021 |

The average total sample sizes of the individual studies \bar{N} were 20, 30, 50, and 100. The primary studies were simulated within a two-group design with $n_E = n_C$. The distribution of the individual sample sizes was based on the systematic review reported in a previous study⁴⁰ in which the sample size distributions of the 50 meta-analyses exhibited a clear positive skewness with average skewness = $+1.423$. To emulate such distribution, a chi-square distribution with 4 degrees of freedom was used to simulate the sample sizes (as the expected skewness for the distribution is $\sqrt{8/df} = 1.414$, similar to that obtained empirically). Additionally, values of 16, 26, 46, and 96 were added to achieve the desired average values.

When $\tau^2 = 0$, the number of conditions was 64 [4 (k values) $\times 4$ (μ_{θ} values) $\times 4$ (\bar{N} values)]. For the other values of τ^2 , the number of conditions was 1920 [4 (k) $\times 4$ (μ_{θ}) $\times 4$ (\bar{N}) $\times 6$ (shape of the distribution of θ_i values) $\times 5$ (τ^2 values)]. The total number of conditions was 1984, and for each one, 10 000 meta-analyses were generated. Thus, 19 840 000 meta-analyses were simulated. Furthermore, 1000 samples per iteration were used in the bootstrapping method.

Several criteria were considered. First, the bias of each of the four methods to estimate the mean effect size was assessed as the difference between the mean of the 10 000 empirical values for each method and condition and the parametric mean effect size for that scenario μ_{θ} . Second, the accuracy in the estimates produced by these four methods was assessed by calculating the mean squared error (MSE) with respect to the true value μ_{θ} across the 10 000 replications of one single condition. Third, the CI width of the 10 methods

used to calculate the CI was estimated by averaging the CI widths across 10 000 replications for each condition. Fourth, the empirical coverage probability for the 95% nominal confidence level of each method was calculated as the percentage of CIs that included the true mean effect size μ_θ using the 10 000 replications for each condition. Finally, we examined the variability in the estimation of the standard errors in the standard random-effects, HM, bootstrapping, and FE methods. This effort was accomplished using the following formula:

$$\frac{Md(SE(\hat{\mu})) - SD(\hat{\mu})}{SD(\hat{\mu})} * 100, \quad (28)$$

with $SD(\hat{\mu})$ as the standard deviation of the mean effect estimates obtained in 10 000 replications of a given condition and $Md(SE(\hat{\mu}))$ representing the median of the estimated standard errors for the mean effect estimates through the 10 000 replications of the same condition. The reason for using the median instead of the mean was to avoid the potential influence of extreme values. Negative values for Equation (28) indicate underestimation of the standard errors.

3 | RESULTS

For brevity, we include only the results for $\mu_\theta = 0.5$ and $\bar{N} = 30$ as the patterns were similar for the remaining levels of both factors. Additionally, we discuss only the results for $\tau^2 = 0.39$ since the differences in the performance of the methods were more pronounced for that value, although the trends observed in scenarios with lower between-studies variation were analogous. The full set of results can be found in Data S4.

This section is divided into five subsections corresponding to the comparative criteria: the bias and MSE of the average effect estimators, the empirical coverage probability and width of the CIs, and the variability of the estimated standard errors.

3.1 | Bias of the average effect estimators

Figure 3 shows the bias of the SM with the DL, REML, and EB estimators of τ^2 and the FE method as a function of the number of studies k and the shape of the distribution of θ_i .

All methods showed a small negative bias across all simulated scenarios for the shape of the distribution of effect parameters, regardless of the number of studies. The FE yielded the most negatively biased estimates

across all conditions because this model assumes a null between-studies variance ($\tau^2 = 0$).

Under normal scenarios (skewness = 0 and kurtosis = 0), the biases of DL, REML, and EB were quite similar across conditions with the same number of studies. These methods produced the most negatively biased values with $k = 20$. For skewness = 0 and kurtosis = -0.58, the performance of the four methods was quite similar to the normal condition. When the shape of the distribution of effect parameters was manipulated with skewness = -2 and kurtosis = 3.65, the mean effects calculated under an RE model with the DL, REML, and EB methods were practically unbiased. Similar results were found with skewness = -1 and kurtosis = 0.47, although under this condition, the four methods were more negatively biased. Under conditions with skewness = 1 and kurtosis = 0.51 and with skewness = 2 and kurtosis = 3.74, the differences in bias among the DL, REML, and EB methods were practically negligible, with values of bias close to -0.025 for all conditions of k . The FE model yielded more negatively biased estimates than the random-effects methods.

3.2 | MSE of the average effect estimators

Figure 4 shows a comparison of the MSE of the standard random-effects methods. As expected, an increase in the number of studies led to a decrease in the MSE values of the four estimators of μ_θ , regardless of the shape of the distribution of effect parameters. In addition, the results across different conditions of skewness and kurtosis and number of studies were generally similar across all four methods, without notable differences in their performance. The FE method showed slightly lower MSE values than the methods based on the RE model with a small number of studies ($k = 10$), and the RE methods had higher MSE values for skewness = 0 and kurtosis = -0.58 than in the normal conditions.

3.3 | Coverage probability of the CIs

Figure 5 shows the empirical coverage probability of the six CIs compared. The SM and HM were not influenced by the applied heterogeneity estimator (DL, REML, or EB). Therefore, only results for the REML estimator are presented. Furthermore, the empirical coverages yielded by the FE method were far below the nominal level and outside of the range considered in Figure 5. The full set of results is presented in Data S4.

Most CIs calculated with the SM, HM, BOOT_P, BOOT_Bca, and PL methods offered better coverage as the number of studies increased, and this improvement was especially evident as k increased from 10 to 20. Under

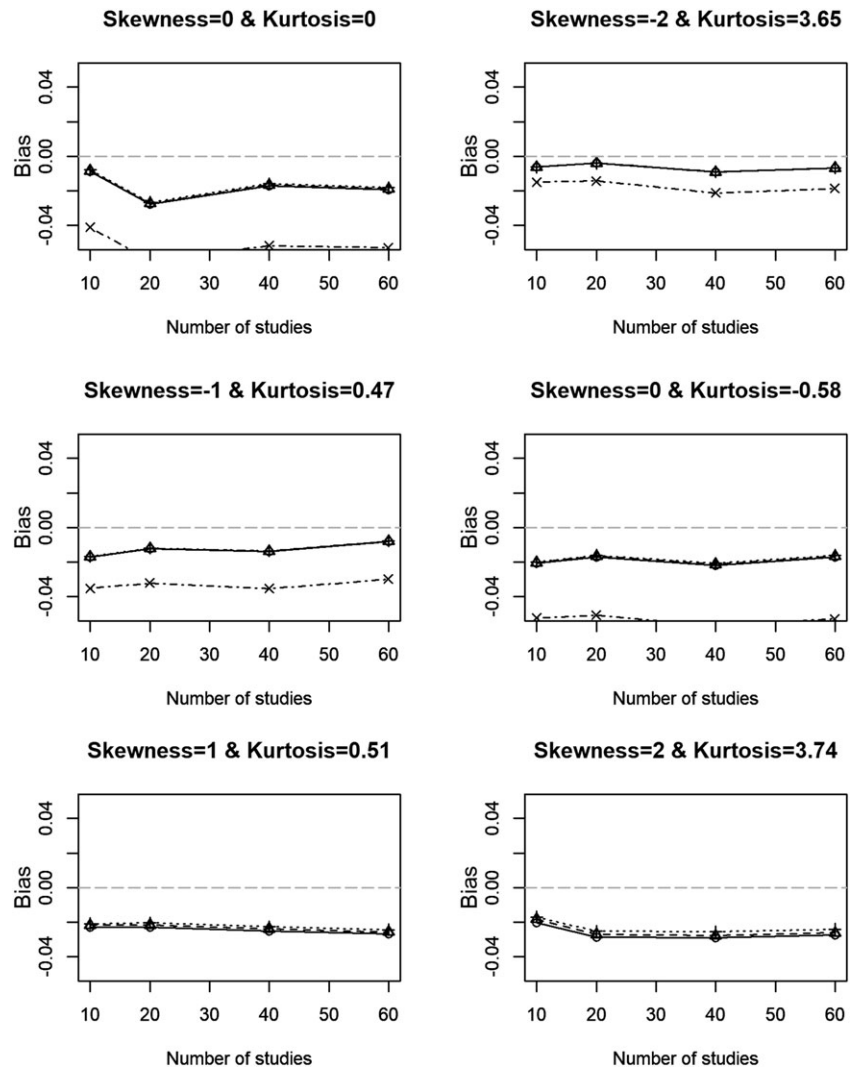


FIGURE 3 Bias of the four methods to estimate μ_{θ} . \circ DL, standard method with DerSimonian and Laird estimator of τ^2 ; Δ REML, standard method with restricted maximum likelihood estimator of τ^2 ; $+$ EB, standard method with empirical Bayes estimator of τ^2 ; \times FE, fixed-effect model. These results are for $\tau^2 = 0.39$, $\mu_{\theta} = 0.5$, and $\bar{N} = 30$. The average standard error of the simulations was 0.0035

normality, some differences in coverage probabilities were found among the CIs obtained by SM, HM, BOOT_P, BOOT_Bca, and PL methods for small numbers of studies ($k = 10$ and 20), with the HM and PL method showing the best coverage. For $k = 10$ and $k = 20$, the HM exhibited observed probabilities of 0.956 and 0.945, respectively, and the PL method obtained values of 0.944 and 0.943. The same trend was found when the effect parameters were nonnormally distributed.

The worst coverage values were found for skewness = 1 and kurtosis = 0.51 and for skewness = 2 and kurtosis = 3.74. Under these two conditions, the CIs obtained by all methods generally showed empirical coverage probabilities slightly below the nominal confidence level, even for a large number of studies.

3.4 | Width of the CIs

Figure 6 shows the width of the five 95% CIs for the compared μ_{θ} . For the standard and Hartung's random-effects methods, only the results for the REML estimator are

presented (see Data S4 for the full set of results). Comparisons of the CI widths are only meaningful between methods with similar coverage probabilities.

The interval width of the five CI procedures uniformly decreased as the number of studies increased. For $k = 10$ and 20 , the CIs obtained with the HM (especially) and PL method were wider than those yielded by the other methods. Although this pattern was consistent across all scenarios, the CIs were narrower in conditions with some degree of departure from normality. This was probably due to a coverage slightly below nominal under nonnormal scenarios. For instance, with $k = 10$ and under the normal scenario, the CI widths for HM and PL were 1.004 and 0.992 with empirical coverage probabilities of 0.956 and 0.944, respectively. Conversely, under the highly nonnormal scenario with skewness = -2 and kurtosis = 3.65, the CI widths for HM and PL were 0.9456 and 0.9306 with empirical coverage probabilities 0.948 and 0.941. The FE method consistently yielded the narrowest CIs at the expense of exhibiting empirical coverages well below nominal.

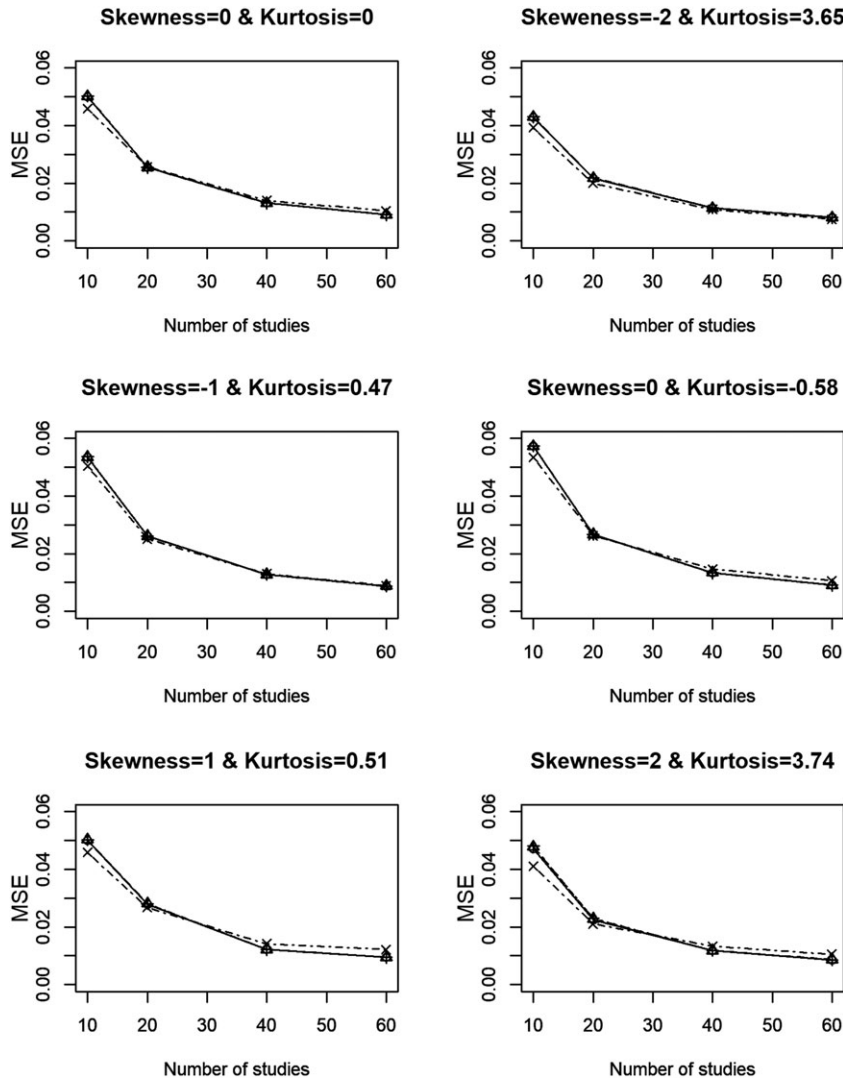


FIGURE 4 Mean squared error of the four methods to estimate μ_{θ} . \circ DL, standard method with DerSimonian and Laird estimator of τ^2 ; Δ REML, standard method with restricted maximum likelihood estimator of τ^2 ; $+$ EB, standard method with empirical Bayes estimator of τ^2 ; \times FE, fixed-effect model. These results are for $\tau^2 = 0.39$, $\mu_{\theta} = 0.5$, and $\bar{N} = 30$. The average standard error of the simulations was 0.0022

3.5 | Variability of the standard errors

Figure 7 shows the variability (in %) of the standard error estimates produced using the REML estimator (see Data S4 for the full set of results). On average, all methods yielded standard error estimates smaller than the standard deviation of the distribution of overall effect estimates empirically constructed through 10 000 replications in a given condition (see Equation (28)). The SM, HM, and BOOT method exhibited standard error estimates very close to the standard deviation of the effect size distribution in all manipulated conditions. In particular, for $k \geq 20$, the percentage underestimation was lower than 5%, with the exception of the condition with skewness = 1 and kurtosis = 0.51. In general, the good performance of the standard error estimates of these methods improved with larger number of studies regardless of shape of the distribution of θ_i , with the exception of conditions with skewness = 1 and kurtosis = 0.51 and skewness = 2 and kurtosis = 3.74, where a slight increase of the percentage underestimation was observed for $k = 60$.

The HM systematically showed the best performance of the standard error estimates in contrast to the BOOT method, which exhibited poor performance (excluding the FE method, not shown in Figure 7). This same trend was found across all conditions of skewness and kurtosis regardless of the number of studies. On average, the percentage departures of the standard errors for SM, HM, and BOOT were -3.52% , -1.89% , and -5.16% , respectively. These differences were larger for small k values. For instance, for $k = 10$, the percentage departures of the standard errors of SM, HM, and BOOT with the conditions of skewness and kurtosis were -5.90% , -4.79% , and -10.18% , respectively.

4 | DISCUSSION

In this study, we examined the performance of various methods for random-effects meta-analysis in terms of bias and MSE of the average effect size estimates, empirical coverage and width of CIs around the average effect size,

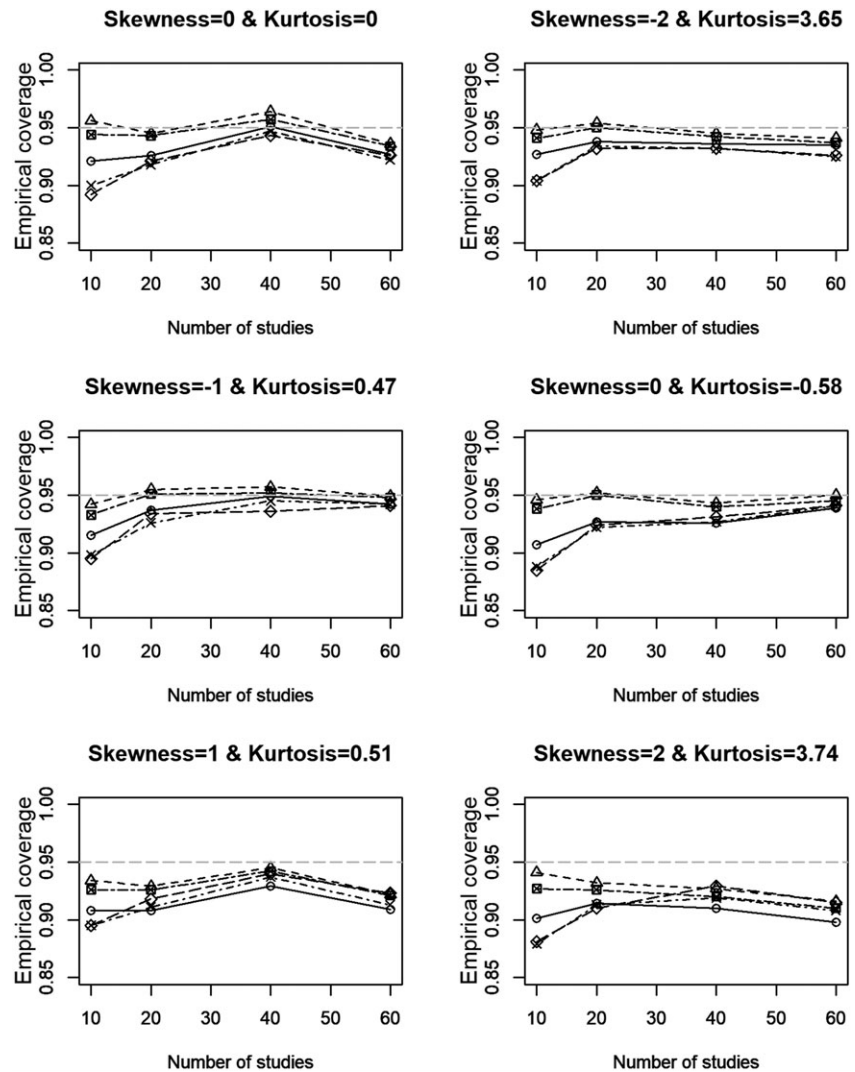


FIGURE 5 Empirical coverage probability for the five confidence interval (CI) methods. \circ SM, standard method; Δ HM, Hartung's method; \times BOOT_P, bootstrapping with the percentile method; \diamond BOOT_Bca, bootstrapping with the BCa method; \square PL, profile likelihood method. The CI methods used REML estimate of τ^2 . These results are for $\tau^2 = 0.39$, $\mu_\theta = 0.5$, and $N = 30$. The average standard error of the simulations was 0.0031

and variability of the standard error estimates, when the normality assumption is not met. We simulated a wide range of scenarios considered to be common in clinical psychology research, using the standardized mean difference as the effect size measure.

Random-effects model typically assumes normality of the effect parameter distribution, and several authors have raised concerns related to the potential impact of nonnormality on the performance of meta-analysis techniques.^{7-9,11,12,21,42} We performed an empirical comparison of several meta-analysis methods using Monte Carlo simulation, and our results suggest that most estimates were not substantially affected by the underlying distribution of effect parameters, even under severe departures from normality. A slightly negative bias of the mean effect size estimates was found across all conditions, even in normal scenarios. This finding has also been reported in previous studies using standardized mean differences (cf, eg, Hedges and Olkin^{16(p125, Ch 6, table 7)} and Marín-Martínez and Sánchez-Meca^{43(p68, fig 1)}), and it is due to a negative relationship between the d estimates and their

weights both for both FE and RE models (Equations (6) and (12), respectively). Such a negative relationship is induced by the inclusion of the effect size estimate, $\hat{\theta}$, in the calculation of the individual sampling variances in Equation (27). As a consequence, the larger the effect size estimate, the lower the weight. An unexpected result was that under normality, the negative bias was slightly larger than for conditions with negatively skewed distributions (skewness = -2 and kurtosis = 3.55, and skewness = -1 and kurtosis = 0.47). For RE methods, the negative bias found in conditions with positive skewness was similar to that observed in normal scenarios. Thus, violation of the normality assumption does not appear to be critical in the estimation of an overall effect in random-effects meta-analysis.

Our findings are largely in agreement with those reported by Kontopantelis and Reeves^{11,12} in the epidemiological field. The conditions manipulated in our study were related to the psychological field, where it is more common to find meta-analyses with a large number of studies and standardized mean differences are often used.

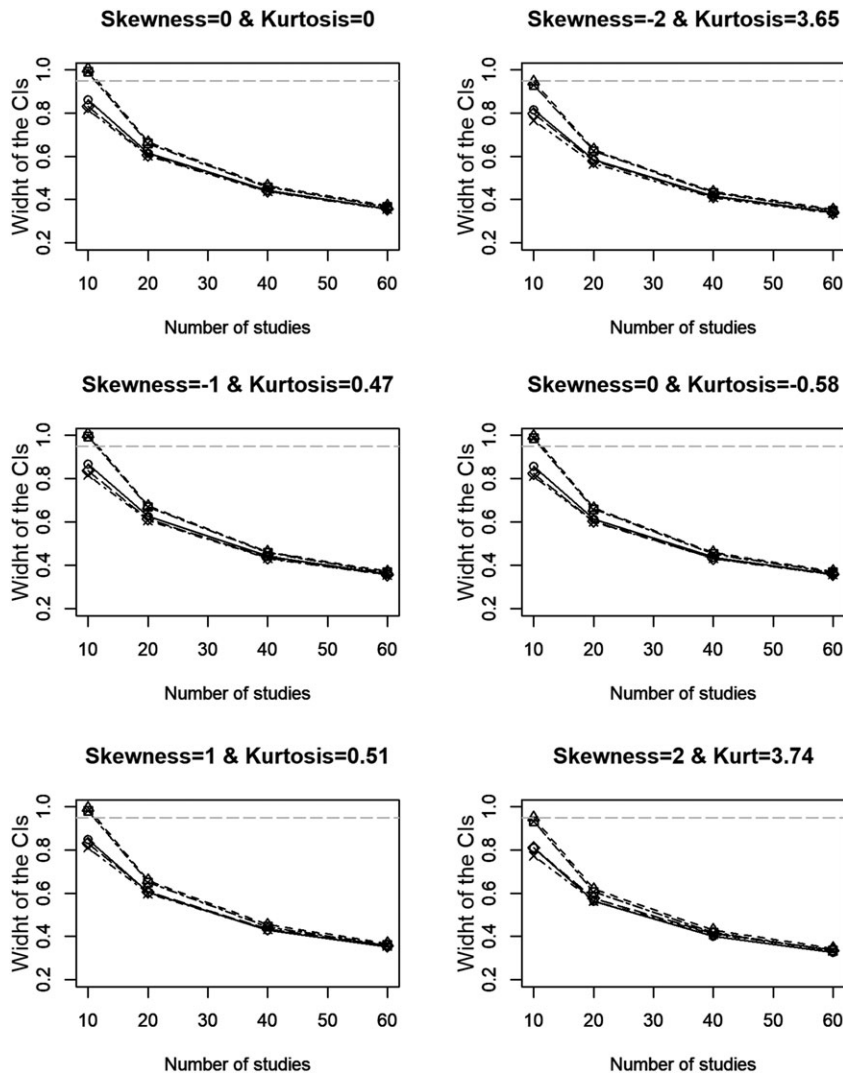


FIGURE 6 Width of the 95% confidence interval (CI) for μ_θ of the five CI methods. \circ SM, standard method; Δ HM, Hartung's method; \times BOOT_P, bootstrapping with the percentile method; \diamond BOOT_Bca, bootstrapping with the BCa method; \square PL, profile likelihood method. The CI methods used REML estimate of τ^2 . These results are for $\tau^2 = 0.39$, $\mu_\theta = 0.5$, and $\bar{N} = 30$. The average standard error of the simulations was 0.0062

We also manipulated the average total sample size of the individual studies and the overall mean of the distribution of effect parameters. Furthermore, we considered several heterogeneity variance estimators and examined the bootstrapping method. A limitation of Kontopantelis and Reeves^{11,12} was that they used an inappropriate method to generate the individual log odds ratios, which cannot be applied to other effect metrics.

As expected, the FE method—which assumes no between-studies variability—provided a poor performance in the estimation of an average effect size in scenarios where $\tau^2 > 0$. For random-effects methods, results were found to be unaffected by the heterogeneity estimator used.

Several authors have criticized the standard random-effects method for not considering the uncertainty due to the variance estimation process, which increases the risk of false positive results.⁴⁴ Our results showed that HM outperformed the SM, with better coverage of the nominal confidence level. This was also reported in

previous simulation studies restricted to normal scenarios.^{17,22,36} Compared with HM, the PL method produced slightly narrower CIs. Both methods yielded coverage probabilities close to the nominal confidence level, with slightly lower values for the PL method.

The final method that we examined was bootstrapping. Despite its theoretical advantage under nonnormal scenarios, this method did not perform better than the SM, HM, or PL method across the set of manipulated conditions and the comparative criteria considered in our study. This method requires substantially more computational resources, and our empirical results (based on the DL estimator) do not encourage its use in this context.

Out of the factors manipulated in this simulation, our results suggest that the number of studies exerts an important influence on the performance of the methods compared. With a small number of studies (less than 20), the performance of the methods was poorer and more notable differences were observed among them

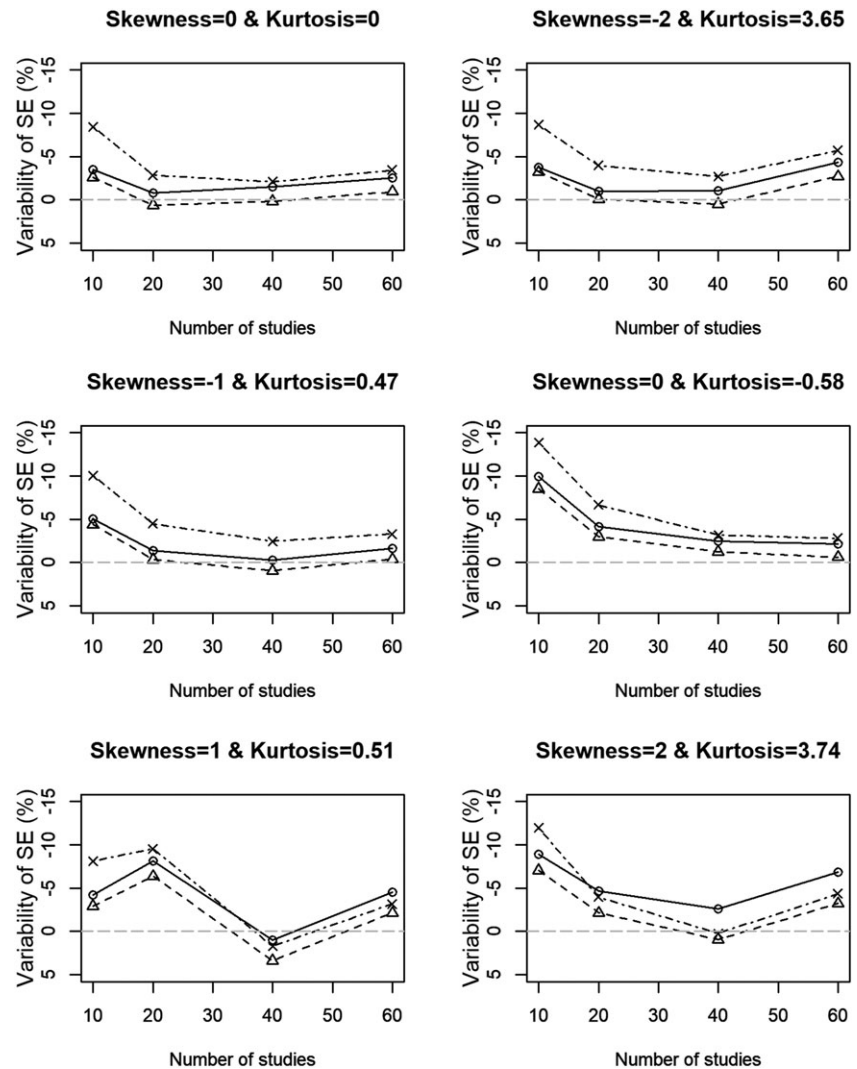


FIGURE 7 Variability of the standard error (SE) of the three methods. ○ SM, standard method; △ HM, Hartung's method; × BOOT, bootstrapping. These results are for $\tau^2 = 0.39$, $\mu_{\theta} = 0.5$, and $\bar{N} = 30$. The average standard error of the simulations was 0.0009%

compared with a moderate-to-large number of studies. Similar results were observed in previous studies that simulated normal scenarios.^{45,46} Many meta-analyses in clinical psychology include fewer than 20 studies, and the situation is even more extreme in other health sciences.⁴⁷ Moreover, our results suggest that large between-studies heterogeneity led to less accurate results and more pronounced differences among methods.

In conclusion, the results of our simulation study suggest that the most commonly used meta-analytic techniques are largely robust to violations of the normality assumption of the effect parameter distribution. All random-effects methods examined, including bootstrapping, yielded similar results under optimal conditions (eg, moderate-to-large number of studies and small between-studies heterogeneity). However, we recommend use of the HM and PL method to construct a CI for the average effect due to their suitability in a wide range of scenarios and their computational simplicity. Nevertheless, the results of our study pertain to the standardized mean

difference and are limited to the manipulated conditions, such that future studies are warranted to improve the generalizability of these findings, extend the manipulated conditions, and consider other effect size indices. Finally, our conclusions apply not only to the estimation of an overall effect size together with its CI under random-effects models but also to the analysis of the influence of moderator variables under mixed-effects models. Indeed, when the influence of a categorical moderator variable on the effect sizes is investigated, the average effect sizes and CIs for each subgroup are calculated. Thus, our recommendation of using HM or PL method for that purpose can also be extended to the estimation of the mean effect parameter of each category of the moderator.

ACKNOWLEDGEMENT

Funding for this study was provided by the Ministerio de Economía y Competitividad of the Spanish Government

and by funds from the Fondo Europeo de Desarrollo Regional (FEDER; Project No. PSI2016-77676-P).

CONFLICT OF INTEREST

The authors reported no conflict of interest.

ORCID

José Antonio López-López  <http://orcid.org/0000-0002-9655-3616>

Julio Sánchez-Meca  <http://orcid.org/0000-0002-8412-788X>

REFERENCES

- Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. *Introduction to Meta-analysis*. Chichester: Wiley; 2009 <https://doi.org/10.1002/9780470743386>.
- Cooper H, Hedges LV, Valentine JC. *The Handbook of Research Synthesis and Meta-analysis*. 2nd ed. New York: Russell Sage Foundation; 2009.
- Hedges LV, Vevea JL. Fixed- and random-effects models in meta-analysis. *Psychol Methods*. 1998;3(4):486-504. <https://doi.org/10.1037/1082-989X.3.4.486>
- Sánchez-Meca J, López-López JA, López-Pina JA. Some recommended statistical analytic practices when reliability generalization (RG) studies are conducted. *Br J Math Stat Psychol*. 2013;66(5):402-425. <https://doi.org/10.3102/1076998612466142>
- Konstantopoulos S, Hedges LV. Analyzing effect sizes: fixed-effects models. In: Cooper H, Hedges LV, Valentine JC, eds. *The Handbook of Research Synthesis and Meta-analysis*. 2nd ed. New York: Russell Sage Foundation; 2009:279-293.
- Raudenbush SW. Analyzing effect sizes: random-effects models. In: Cooper H, Hedges LV, Valentine JC, eds. *The Handbook of Research Synthesis and Meta-analysis*. 2nd ed. New York: Russell Sage Foundation; 2009:295-315.
- Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res Synth Methods*. 2010;1(2):97-111. <https://doi.org/10.1002/jrsm.12>
- Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. *Stat Med*. 2001;20(6):825-840. <https://doi.org/10.1002/sim.650>
- Brockwell SE, Gordon IR. A simple method for inference on an overall effect in meta-analysis. *Stat Med*. 2007;26(25):4531-4543. <https://doi.org/10.1002/sim.2883>
- Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Stat Med*. 1996;15(6):619-629. [https://doi.org/10.1002/\(SICI\)1097-0258\(19960330\)15:6<619::AID-SIM188>3.0.CO;2-A](https://doi.org/10.1002/(SICI)1097-0258(19960330)15:6<619::AID-SIM188>3.0.CO;2-A)
- Kontopantelis E, Reeves D. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: a comparison between DerSimonian-Laird and restricted maximum likelihood. *Stat Methods Med Res*. 2012;21(6):657-659. <https://doi.org/10.1177/0962280211413451>
- Kontopantelis E, Reeves D. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: a simulation study. *Stat Methods Med Res*. 2012;21(4):409-426. <https://doi.org/10.1177/0962280210392008>
- Schmidt FL, Oh IS, Hayes TL. Fixed- versus random-effects models in meta-analysis: model properties and an empirical comparison of differences in results. *Br J Math Stat Psychol*. 2009;62(1):97-128. <https://doi.org/10.1348/000711007X255327>
- Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Stat Med*. 2000;19:1707-1728. [https://doi.org/10.1002/10970258\(20000715\)19:13<1707::AID-SIM491>3.0.CO;2-P](https://doi.org/10.1002/10970258(20000715)19:13<1707::AID-SIM491>3.0.CO;2-P)
- Hoaglin DC. We know less than we should about methods of meta-analysis. *Res Synth Methods*. 2015;6(3):287-289. <https://doi.org/10.1002/jrsm.1146>
- Hedges LV, Olkin I. *Statistical Methods for Meta-analysis*. Orlando, FL: Academic Press; 1985.
- Sánchez-Meca J, Marin-Martínez F. Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychol Methods*. 2008;13(1):31-48. <https://doi.org/10.1037/1082-989X.13.1.31>
- Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *J Educ Behav Stat*. 2005;30(3):261-293. <https://doi.org/10.3102/10769986030003261>
- Hartung J. An alternative method for meta-analysis. *Biom J*. 1999;41(8):901-916. [https://doi.org/10.1002/\(SICI\)1521-4036\(199912\)41:8<901::AID-BIMJ901>3.0.CO;2-W](https://doi.org/10.1002/(SICI)1521-4036(199912)41:8<901::AID-BIMJ901>3.0.CO;2-W)
- Hartung J, Knapp G. On tests of the overall treatment effect in the meta-analysis with normally distributed responses. *Stat Med*. 2001;20(12):1771-1782. <https://doi.org/10.1002/sim.791>
- Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. *Stat Med*. 2002;21(16):3153-3159. <https://doi.org/10.1002/sim.1549>
- Int'Hout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol*. 2014;14(1):25. <https://doi.org/10.1186/1471-2288-14-25>
- Guolo A. Higher-order likelihood inference in meta-analysis and meta-regression. *Stat Med*. 2012;31(4):313-327. <https://doi.org/10.1002/sim.4451>
- Henmi M, Copas JB. Confidence intervals for random effects meta-analysis and robustness to publication bias. *Stat Med*. 2010;29(29):2969-2983. <https://doi.org/10.1002/sim.4029>
- Adams DC, Gurevitch J, Rosenberg MS. Resampling tests for meta-analysis of ecological data. *Ecology*. 1995;78(4):1277-1283. <https://doi.org/10.2307/2265879>
- van den Noortgate W, Onghena P. Parametric and nonparametric bootstrap methods for meta-analysis. *Behav Res Methods*. 2005;37(1):11-22. <https://doi.org/10.3758/BF03206394>
- Efron B. Better bootstrap confidence intervals. *J Am Stat Assoc*. 1987;82(397):171-200.
- Efron B, Hastie T. *Computer Age Statistical Inference*. New York: Cambridge University Press; 2016.

29. Efron B. Bootstrap methods: another look at the jackknife. *Ann Stat*. 1979;7(1):1-26. <https://doi.org/10.1214/aos/1176344552>
30. Veroniki AA, Jackson D, Viechtbauer W, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res Synth Methods*. 2016;7(1):55-79. <https://doi.org/10.1002/jrsm.1164>
31. DerSimonian R, Laird N. Meta-analysis of clinical trials. *Clin Contr Trials*. 1986;7(3):177-188. [https://doi.org/10.1016/0197-2456\(86\)90046-2](https://doi.org/10.1016/0197-2456(86)90046-2)
32. Hoaglin DC. Misunderstandings about Q and 'Cochran's Q test' in meta-analysis. *Stat Med*. 2016;35(4):485-495. <https://doi.org/10.1002/sim.6632>
33. Kulinskaya E, Dollinger MB, Bjørkestøl K. Testing for homogeneity in meta-analysis I. The one-parameter case: standardized mean difference. *Biometrics*. 2011;67(1):203-212. <https://doi.org/10.1111/j.1541-0420.2010.01442.x>
34. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. *Stat Med*. 1995;14(4):395-411. <https://doi.org/10.1002/sim.4780140406>
35. Morris CN. Parametric empirical Bayes inference: theory and applications. *J Am Stat Assoc*. 1983;78(381):47-55. <https://doi.org/10.1080/01621459.1983.10477920>
36. Viechtbauer W, López-López JA, Sánchez-Meca J, Marín-Martínez F. A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychol Methods*. 2015;20(3):360-374. <https://doi.org/10.1037/met0000023>
37. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36:1-48.
38. Guolo A, Varin C. The R package metaLik for likelihood inference in meta-analysis. *J Stat Softw*. 2012;50:1-14.
39. Cauty A, Ripley BD. boot: bootstrap R (S-PLUS) functions, 2012. Available from: <http://CRAN.R-project.org/package=boot>, R Package version 1.3-7.
40. Rubio-Aparicio M, Marín-Martínez F, Sánchez-Meca J, López-López JA. A methodological review of meta-analyses about the effectiveness of clinical psychology treatments. *Behav Res Methods*. 2017. <https://doi.org/10.3758/s13428-017-0973-8>
41. Fleishman AI. A method for simulating non-normal distributions. *Psychometrika*. 1978;43(4):521-532. <https://doi.org/10.1007/BF02293811>
42. Sidik K, Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies. *Stat Med*. 2007;26(9):1964-1981. <https://doi.org/10.1002/sim.2688>
43. Marín-Martínez F, Sánchez-Meca J. Weighting by inverse variance or by sample size in random-effects meta-analysis. *Educ Psychol Meas*. 2010;70(1):56-73. <https://doi.org/10.1177/0013164409344534>
44. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? *Stat Med*. 2002;21(11):1559-1573. <https://doi.org/10.1002/sim.1187>
45. López-López JA, Marín-Martínez F, Sánchez-Meca J, van den Noortgate W, Viechtbauer W. Estimation of the predictive power of the model in mixed-effects meta-regression: a simulation study. *Br J Math Stat Psychol*. 2014;67(1):30-48. <https://doi.org/10.1111/bmsp.12002>
46. Rubio-Aparicio M, Sánchez-Meca J, López-López JA, Marín-Martínez F, Botella J. Analysis of categorical moderators in mixed-effects meta-analysis: consequences of using pooled vs. separate estimates of the residual between-studies variances. *Br J Math Stat Psychol*. 2017;70(3):439-456. <https://doi.org/10.1111/bmsp.12092>
47. Davey J, Turner RM, Clarke MJ, Higgins JPT. Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. *BMC Med Res Methodol*. 2011;11(1):160. <https://doi.org/10.1186/1471-2288-11-160>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Rubio-Aparicio M, López-López JA, Sánchez-Meca J, Marín-Martínez F, Viechtbauer W, Van den Noortgate W. Estimation of an overall standardized mean difference in random-effects meta-analysis if the distribution of random effects departs from normal. *Res Syn Meth*. 2018;9:489–503. <https://doi.org/10.1002/jrsm.1312>