British Journal of Mathematical and Statistical Psychology (2014), 67, 30–48

The British Psychological Society

www.wileyonlinelibrary.com

Estimation of the predictive power of the model in mixed-effects meta-regression: A simulation study

José Antonio López-López[1], Fulgencio Marín-Martínez[1]*, Julio Sánchez-Meca[1], Wim Van den Noortgate[2] and Wolfgang Viechtbauer[3]

[1]Dept. Basic Psychology & Methodology, University of Murcia, Spain
[2]University of Leuven, Belgium
[3]Maastricht University, The Netherlands

Several methods are available to estimate the total and residual amount of heterogeneity in meta-analysis, leading to different alternatives when estimating the predictive power in mixed-effects meta-regression models using the formula proposed by Raudenbush (1994, 2009). In this paper, a simulation study was conducted to compare the performance of seven estimators of these parameters under various realistic scenarios in psychology and related fields. Our results suggest that the number of studies (k) exerts the most important influence on the accuracy of the results, and that precise estimates of the heterogeneity variances and the model predictive power can only be expected with at least 20 and 40 studies, respectively. Increases in the average within-study sample size ($\bar{N}$) also improved the results for all estimators. Some differences among the accuracy of the estimators were observed, especially under adverse (small k and $\bar{N}$) conditions, while the results for the different methods tended to convergence for more optimal scenarios.

1. Introduction

Meta-analysis is a form of research synthesis that allows researchers to quantitatively integrate the results from a set of studies on the same topic (Borenstein, Hedges, Higgins & Rothstein, 2009; Cooper, Hedges & Valentine, 2009). Since the outcomes from the individual studies are often expressed in different measurement units, their results are typically converted into a common metric through a standardized effect size index (such as the standardized mean difference). The main objectives in a meta-analysis are to obtain an overall effect size estimate, to assess the heterogeneity among the individual effect size estimates, and to search for moderators that can account for (at least) part of that heterogeneity (Hedges & Olkin, 1985; Sánchez-Meca & Marín-Martínez, 2010).

*Correspondence should be addressed to Fulgencio Marín-Martínez, Ph. D., Assistant Professor of Methodology of the Behavioral Sciences, Dept. Basic Psychology & Methodology, Faculty of Psychology, Espinardo Campus, University of Murcia, 30100-Murcia, Spain (e-mail: fulmarin@um.es).

DOI:10.1111/bmsp.12002

The results or effect sizes of the individual studies in a meta-analysis usually exhibit some heterogeneity (e.g., Sidik & Jonkman, 2005b; Thompson & Higgins, 2002). This means that, although a set of studies analysing the same phenomenon (e.g., effectiveness of psychological treatments and interventions on a given disorder) are selected, their results are likely to differ to some extent. For that reason, moderator analyses typically constitute a crucial element of a meta-analysis (Lipsey, 2009). In a moderator analysis, the goal is to test the influence of one or more study characteristics (e.g., type and duration of the intervention, severity of the disorder in the sample patients) on the outcome variable (e.g., efficacy of the intervention, assessed through the comparison between a treatment and a control group). Such analyses can be conducted by fitting linear models to the data where the moderators constitute the predictor variables and the effect sizes are employed as the criterion variable (Borenstein *et al.*, 2009). This leads to so-called meta-regression models (Thompson & Higgins, 2002). In a meta-regression model, both continuous and categorical moderators can be included.

When carrying out a meta-analysis, some statistical model must be assumed for the effect size distribution, and the model choice will have an influence on the validity and generalizability of the results from the meta-analysis. Two kinds of statistical models have been employed for the majority of meta-analytic reviews conducted so far, namely the fixed-effects and random-effects models (Hedges & Vevea, 1998; Schmidt, Oh & Hayes, 2009). Nowadays, most researchers agree that the model choice should be made based on the generalizability intended for the results (National Research Council, 1992). Only random-effects models, which include an additional variance component to model the between-studies heterogeneity, allow for generalization to studies different to the ones included in the meta-analysis, which is usually the goal when carrying out such a review. Thus, random-effects models are a suitable option for most meta-analyses (Hedges & Vevea, 1998; Raudenbush, 1994, 2009).

Under a random-effects model, it is assumed that the study outcomes (e.g., treatment efficacy) will fluctuate as a consequence of two sources of variation: the sampling of the participants for each study; and the differential characteristics of the studies (e.g., different conditions of the sample, treatment application, methodology, or context in each individual study). The magnitude of the latter can be analysed through the estimation of the heterogeneity (or between-studies) variance, τ^2, which represents the excess variation among the effects over that expected from sampling error alone (Thompson & Sharp, 1999). In contrast to the sampling variances from each effect size, which quantify the random sampling error, τ^2 denotes systematic differences due to the influence of characteristics from the individual studies. The identification of some of these characteristics (or moderators) is the main objective of the moderator analyses. Since the moderators are usually included as fixed effects in the model, the addition of a random effect (the effect sizes in the studies) to model the heterogeneity among the studies leads to mixed-effects meta-regression models.

There are several parameters of interest in a meta-regression model. One of these is the model predictive power, denoted by P^2 (P denotes the capital Greek letter 'rho'), which can be defined as the proportion of variance among the effect sizes that can be accounted for by the predictors included in the model. Note that only the variance due to differences among the studies, quantified by τ^2, can be explained by the predictors usually included in a mixed-effects meta-regression model. An estimate of the P^2 parameter is usually denoted as an R^2 value. The interpretation of R^2 is identical in ordinary regression and in meta-regression models, in terms of a percentage or proportion of the variability in the outcomes associated with the predictor(s).

When regression models are fitted using ordinary least squares techniques, the R^2 index is computed as the quotient between the sum of squares due to the regression and the total sum of squares, that is, $R^2 = SS_R/SS_T$ (e.g., Pedhazur & Schmelkin, 1991). However, this strategy is not suitable for meta-regression models because part of the total variability, more specifically the sampling error of an observed effect size given the population effect size in that study, cannot by definition be explained by the moderators included in the model (Aloe, Becker & Pigott, 2010; Konstantopoulos & Hedges, 2009; Rodriguez & Maeda, 2006).[1] Thus, a different method is typically proposed for obtaining an R^2 index in meta-regression models (Raudenbush, 1994), where the total variability is an estimate of the between-studies variance, τ^2, and the variability explained by the predictors in the model is estimated as a part of τ^2 (see equation (3)) . This method will be presented, explained, and illustrated in this paper.

In a meta-regression model, an adequate estimate of the magnitude of its predictive power via the R^2 index is an essential complement of the statistical significance of the model. The R^2 index informs us about the practical significance or the degree of influence of a set of moderators in the heterogeneity of the effect sizes in a meta-analysis (e.g., explaining around 20% or 30% of the heterogeneity). However, as far as we know, no studies have yet evaluated in a systematic manner the performance of the R^2 index in the conditions of a meta-regression model. Therefore, the purpose of the present study was to assess the performance of the method proposed by Raudenbush to compute an R^2 index in meta-analysis, by conducting a Monte Carlo simulation with different conditions usually found in the real meta-analyses.

The outline of the present paper is as follows. First, mixed-effects meta-regression models are briefly sketched. Second, various alternatives for computing an R^2 index according to the proposal of Raudenbush (1994) for meta-analysis are considered. After presenting the methods, results from previous simulation studies that pursued part of the objectives of our study are summarized. The performance of the alternative methods here considered is then illustrated by applying them to an example. Next, a simulation study comparing the various estimators is presented and the results obtained are detailed. Finally, the results are discussed and some conclusions provided, where the degree of accuracy of the different methods for the computation of an R^2 index as a measure of the explanatory power of a predictor is assessed as a function of the specific conditions in a meta-analysis (e.g., number of studies, sample size distribution of the studies, effect size distribution, and the true percentage of variance accounted for by the predictor).

2. Mixed-effects meta-regression models

In a meta-analysis with k studies, let $\mathbf{y}$ denote a $k \times 1$ vector of independent effect sizes $\{y_i\}$ that represents the results of the studies, and $\mathbf{X}$ a $k \times (p + 1)$ design matrix of full column rank with p predictor variables, representing some differential characteristics in the studies. Since the predictors are included as fixed effects in the model, assuming a random-effects model for the effect sizes leads to a mixed-effects meta-regression model, which can be expressed by the formula (Raudenbush, 1994)

[1] An exception to this is when meta-analysing the raw data from a set of individual studies, in which case within-study variability can be accounted for. For more details on so-called individual participant data meta-analyses, see, for example, Cooper and Patall (2009).

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{u} + \mathbf{e}, \tag{1}$$

where $\boldsymbol{\beta}$ is a $(p + 1) \times 1$ vector containing the regression coefficients $\{\beta_0, \beta_1, \ldots, \beta_p\}$, $\mathbf{u}$ is a $k \times 1$ vector of independent between-studies errors $\{u_i\}$ with distribution $N(0, \tau^2_{\text{res}})$, and $\mathbf{e}$ is a $k \times 1$ vector of independent within-study errors $\{e_i\}$, each with distribution $N(0, v_i)$. While v_i is the within-study variance (or sampling error) for the ith study, τ^2_{res} represents the residual heterogeneity (or between-studies) variance, that is, the remaining variability in the true effect sizes not accounted for after adding one or more predictors to the model (Viechtbauer, 2007a).

Note that the mixed-effects model presented in equation (1) is actually an extension of the random-effects model and that the latter can be formulated if $\mathbf{X}$ is defined as a $k \times 1$ vector of ones. In this case we would have a model without predictors, where $\boldsymbol{\beta}$ is a scalar containing the hypermean (mean of the population effects) and $\mathbf{u}$ is normally distributed with mean 0 and variance τ^2, the latter denoting the total heterogeneity in the true effects. If, moreover, the error term $\mathbf{u}$ were suppressed from equation (1), then the model would become a fixed-effect model (which is equivalent to setting $\tau^2 = 0$ or assuming that the sampling error is the only source of variability).

The regression coefficients $\{\beta_0, \beta_1, \ldots, \beta_p\}$ can be estimated using the weighted least squares formula

$$\mathbf{b} = (\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{W}}\mathbf{y}, \tag{2}$$

where $\hat{\mathbf{W}}$ is a $k \times k$ diagonal matrix with the inverse variances of the effect sizes as elements, that is, $\{1/(v_i + \hat{\tau}^2_{\text{res}})\}$ for mixed-effects models. Note that an adequate estimate of both the within-study variance for each study, v_i, and the residual between-studies variance, τ^2_{res}, is needed for the estimation of the regression coefficients. For commonly used effect size metrics (e.g., standardized mean differences, correlation coefficients, odds ratios, risk ratios), approximately unbiased estimators are available for v_i and the usual practice in meta-analysis is to substitute those estimates and treat them as known values (e.g., Aloe *et al.*, 2010; Hedges & Pigott, 2004; Knapp, Biggerstaff & Hartung, 2006; Konstantopoulos & Hedges, 2009; Viechtbauer, 2007b; for a different approach, see Malzahn, Böhning & Holling, 2000). A more crucial issue is the choice of estimator for τ^2_{res}, and at least seven different estimators have been described in the literature, as detailed in the next section.

3. Estimating the model predictive power in meta-analysis

A proposal to compute an R^2 index in meta-analysis was presented by Raudenbush (1994, 2009). It is based on the re-estimation of the amount of heterogeneity (i.e., between-studies variance) after adding one or more predictors to the model, resulting in the residual heterogeneity or the heterogeneity that cannot be explained by the predictors. The rationale for this index is that the extent to which the moderators can account for the heterogeneity in the true effects will be reflected in the degree by which the residual heterogeneity, τ^2_{res}, will be smaller than the total amount of heterogeneity, τ^2, as a result of including explanatory variables in the model. In practice, the parameter values are replaced by their estimates, $\hat{\tau}^2$ and $\hat{\tau}^2_{\text{res}}$, allowing for the computation of the R^2 index as (Borenstein *et al.*, 2009)

$$R^2 = 1 - \frac{\hat{\tau}^2_{res}}{\hat{\tau}^2}, \tag{3}$$

denoting the proportion of total heterogeneity accounted for by the moderator(s) included in the model.

Several alternatives have been proposed in the literature to estimate the total heterogeneity variance, τ^2, in random-effects models (DerSimonian & Laird, 1986; Morris, 1983; Sánchez-Meca & Marín-Martínez, 2008; Sidik & Jonkman, 2005b, 2007; Viechtbauer, 2005). Most of these estimators have also been extended to mixed-effects models, allowing for estimating the residual heterogeneity variance, τ^2_{res} (Raudenbush, 1994, 2009; Sidik & Jonkman, 2005a,b). It is important to remark here that, for both parameters, no estimator is expected to provide accurate results unless the number of studies is large enough (e.g., Borenstein *et al.*, 2009; Schulze, 2004).

Seven different estimators of τ^2 and τ^2_{res} can be computed with the formulae gathered in Table 1. The metafor package programmed in R (Viechtbauer, 2010) directly computes these seven estimators from the values of the effect sizes and their corresponding within-study variances in the studies of the meta-analysis. The Hedges (HE), Hunter–Schmidt (HS), DerSimonian–Laird (DL), and Sidik–Jonkman (SJ) methods are non-iterative estimators, while the maximum likelihood (ML), restricted maximum likelihood (REML), and empirical Bayes (EB) methods require iterative computations. All estimators presented in Table 1 can be succinctly expressed after defining the matrix

$$\mathbf{M} = \mathbf{W} - \mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}, \tag{4}$$

where $\mathbf{W}$ is a diagonal weighting matrix whose elements, w_i, can change from one estimator to another. For the iterative estimators, one starts with an initial estimate of τ^2_{res} (e.g., as obtained with one of the non-iterative estimators) and then iterates through the equation

$$\hat{\tau}^2_{new} = \hat{\tau}^2_{old} + \Delta \tag{5}$$

until convergence, where Δ is given in Table 1 for the ML, REML, and EB estimators. Although all the equations gathered in Table 1 include predictors, they also apply for the random-effects model without predictors by setting $p = 0$ and with $\mathbf{X}$ being a $k \times 1$ vector of ones. In a model without predictors, the equations in Table 1 estimate the total heterogeneity variance, τ^2, while the inclusion of predictors in the same equations leads to the estimation of the residual heterogeneity variance, τ^2_{res}.

A value of zero for $\hat{\tau}^2_{res}$ suggests that all the heterogeneity among the effect sizes is accounted for by the predictors included in the model (Viechtbauer, 2007a). Also, due to random sampling error, the estimators in Table 1 (with the exception of the SJ estimator) can provide a negative estimate, which is a value outside of the parameter space for a variance component. The usual practice is to truncate negative values to zero. When an iterative estimator is employed, a simple strategy to avoid negative estimates is the use of step-halving (Jennrich & Sampson, 1976), which implies multiplying the adjustment value, Δ, by 1/2 (e.g., first by 1/2, then by 1/4, then by 1/8, and so on) until it becomes sufficiently small enough for the resulting estimate to stay non-negative.

Both (total and residual) heterogeneity variance estimates employed in equation (3) can be obtained using any of the methods presented in Table 1. As a consequence, there

Table 1. Heterogeneity variance estimators

Estimator	Formula	Weights	Reference
Hedges (HE)	$\hat{\tau}^2_{\text{HE}} = \frac{\mathbf{y}'\mathbf{M}\mathbf{y} - \text{tr}(\mathbf{M}\mathbf{V})}{k-p-1}$	$w_i = 1$	Raudenbush (1994)
Hunter–Schmidt (HS)	$\hat{\tau}^2_{\text{HS}} = \frac{\mathbf{y}'\mathbf{M}\mathbf{y} - k}{\text{tr}(\mathbf{W})}$	$w_i = 1/v_i$	Viechtbauer, López-López, Sánchez-Meca and Marín-Martínez (2012)
DerSimonian–Laird (DL)	$\hat{\tau}^2_{\text{DL}} = \frac{\mathbf{y}'\mathbf{M}\mathbf{y} - (k-p-1)}{\text{tr}(\mathbf{M})}$	$w_i = 1/v_i$	Sidik and Jonkman (2005a)
Sidik–Jonkman (SJ)	$\hat{\tau}^2_{\text{SJ}} = \frac{\hat{\tau}^2_0(\mathbf{y}'\mathbf{M}\mathbf{y})}{k-p-1}$	$w_i = 1/(v_i + \hat{\tau}^2_0)$	Sidik and Jonkman (2005b)
Maximum likelihood (ML)	$\Delta_{\text{ML}} = \frac{\mathbf{y}'\mathbf{M}\mathbf{M}\mathbf{y} - \text{tr}(\mathbf{W})}{\text{tr}(\mathbf{W}\mathbf{W})}$	$w_i = 1/\left(v_i + \hat{\tau}^2_{\text{old}}\right)$	Raudenbush (2009)
Restricted maximum likelihood (REML)	$\Delta_{\text{REML}} = \frac{\mathbf{y}'\mathbf{M}\mathbf{M}\mathbf{y} - \text{tr}(\mathbf{M})}{\text{tr}(\mathbf{M}\mathbf{M})}$	$w_i = 1/\left(v_i + \hat{\tau}^2_{\text{old}}\right)$	Raudenbush (2009)
Empirical Bayes (EB)	$\Delta_{\text{EB}} = \frac{(k/(k-p-1))\mathbf{y}'\mathbf{M}\mathbf{y} - k}{\text{tr}(\mathbf{W})}$	$w_i = 1/\left(v_i + \hat{\tau}^2_{\text{old}}\right)$	Sidik and Jonkman (2005a)

Note. For the HE estimator, $\mathbf{V} = \text{diag}\{v_i\}$; for the SJ estimator, $\hat{\tau}^2_0 = \frac{\Sigma(y_i - \bar{y})^2}{k}$.

are at least seven different methods for computing the R^2 index using this proposal. Aloe *et al.* (2010) recommended using the same method for both estimates. Indeed, it does not seem sensible to mix two estimates obtained using methods with different theoretical assumptions and, furthermore, only the estimates obtained with the same method are readily comparable.

It is important to note that, due to sampling error, the formula proposed by Raudenbush may require or lead to truncation in several situations. First, $\hat{\tau}^2_{res}$ can be larger than $\hat{\tau}^2$ for a given meta-analytic data set, especially with small samples (small number of studies, small sample sizes, or both), leading to a negative R^2 value that is typically truncated to zero in practice (indicating that all of the heterogeneity among the effect sizes remains unaccounted for after including the moderator(s) in the model). Second, a negative value of $\hat{\tau}^2$ truncated to zero leads to division by zero in equation (3), in which case R^2 is undefined. It is then common practice to set (or truncate) the value of R^2 to 0 (indicating that none of the heterogeneity is accounted for by the moderators, given that there appeared to be none to begin with). Finally, with a positive value of $\hat{\tau}^2$, a negative value of $\hat{\tau}^2_{res}$ truncated to zero will lead to an R^2 value of 1 (indicating that all of the heterogeneity is accounted for).

Since an estimate of the heterogeneity variance is included in both the random- and mixed-effects model weights (cf. equation (2)), the accuracy of these estimates might affect the result of other statistical analyses, such as the computation of an overall effect size estimate and its confidence interval in a random-effects model or the estimation and testing of the model coefficients in a mixed-effects meta-regression model. However, getting accurate estimates of τ^2 and τ^2_{res} seems even more crucial for the assessment of the predictive power in meta-regression models since the R^2 index in equation (3) requires estimates both of the total and residual amount of heterogeneity.

4. Previous simulation studies

Several simulation studies have already been conducted with the aim of comparing the accuracy of various estimators of the heterogeneity variance in meta-analysis. Some of these studies employed effect size indices for dichotomous measures (e.g., Malzahn *et al.*, 2000; Sidik & Jonkman, 2005b, 2007), while others considered indices for continuous variables (e.g., Van den Noortgate & Onghena, 2003; Viechtbauer, 2005).

In general, a positive bias has been found in the SJ estimator for small to medium parameter values (Sidik & Jonkman, 2005b, 2007), while a negative bias was reported for the HS and ML estimators, as well as for the DL method when estimating large parameter values (Malzahn *et al.*, 2000; Viechtbauer, 2005). The HE method was found to perform appropriately in terms of bias, although it was less efficient than the HS, DL, ML, and REML estimators (Viechtbauer, 2005). Finally, good performance was observed for both the REML and EB estimators when considering bias and efficiency criteria jointly (Sidik & Jonkman, 2007; Van den Noortgate & Onghena, 2003; Viechtbauer, 2005).

All of these simulation studies focused on random-effects models. Therefore, it is not certain to what extent these results would also carry over to mixed-effects meta-regression models. Moreover, these studies do not indicate whether one of the various estimators for τ^2 and τ^2_{res} would be preferable when computing the R^2 index given by equation (3).

5. Objectives and hypotheses of this study

In the present study, all seven heterogeneity variance estimators presented (i.e., the HE, HS, DL, SJ, ML, REML, and EB estimators) were considered and applied to simulated meta-analyses where the standardized mean difference was the effect size index. This simulation compared the accuracy of the methods under different scenarios for the estimation of the total and residual heterogeneity variances as well as of the model predictive power, as defined by Raudenbush (1994).

A first objective was to check whether the patterns reported in previous studies for the heterogeneity variance estimators under random-effects models also apply for mixed-effects models with one predictor. The second objective was to assess the performance of Raudenbush's proposal for estimating the model predictive power in meta-analysis when computing R^2 with the various estimators for τ^2 and τ^2_{res} described earlier.

Regarding our hypotheses, we expected to find results similar to those reported in previous simulation studies for the different estimators of the total heterogeneity variance under random-effects models. In particular, we expected the HS and ML estimators to show a negative bias and the DL method to provide negatively biased estimates for large parameter values. The SJ estimator was expected to show a large positive bias for small to medium parameter values, while the HE method was expected to provide essentially unbiased estimates, although less efficiently than the remaining methods under comparison. According to our hypotheses, the REML and EB estimators were expected to provide the best performance, as found in previous simulation studies. The same trends observed for the different estimators under random-effects models were also expected to be found when estimating the residual heterogeneity variance under mixed-effects meta-regression models with one moderator. Finally, it was expected that the REML and EB estimators would also provide the best performance for the estimation of the predictive power in mixed-effects meta-regression models, computed with equation (3). We also expected that an increase in the average sample size and (especially) the number of studies would lead to more precise results for all estimators.

6. An illustrative example

Else-Quest, Hyde and Linn (2010) published a meta-analysis integrating results from the Programme for International Student Assessment (PISA) in different countries in 2003. This report evaluated 15-year-old students' performance in several subjects. The authors focused on mathematics and, since they were interested in gender differences, effect sizes were defined as standardized mean differences between the marks achieved by boys and girls (with positive values indicating better performance for boys).

One of the coded characteristics for each country was the share of parliamentary seats held by women (given as a proportion), used as a moderator in this example. Twenty countries from different parts of the world were selected to illustrate the methods described earlier. Table 2 shows the effect size, y_i, sampling variance, v_i, and the moderator value, $Parl_i$, for each of the 20 countries.

All seven variance estimators compared in this study were employed to estimate the total heterogeneity variance in a random-effects model, as well as the slope, the residual heterogeneity variance, and the proportion of variance accounted for by the moderator in a mixed-effects meta-regression model. Results are presented in Table 3.

Table 2. Data from the meta-analysis published by Else-Quest *et al.* (2010)

Country	y_i	v_i	$Parl_i$	Country	y_i	v_i	$Parl_i$
Australia	0.06	0.0003	0.27	Mexico	0.13	0.0001	0.16
Belgium	0.07	0.0005	0.25	Netherlands	0.06	0.0010	0.33
Brazil	0.16	0.0009	0.09	Poland	0.06	0.0009	0.21
Canada	0.13	0.0002	0.24	South Korea	0.25	0.0008	0.06
France	0.09	0.0009	0.12	Spain	0.10	0.0004	0.27
Germany	0.09	0.0009	0.31	Sweden	0.07	0.0009	0.45
Greece	0.21	0.0009	0.09	Thailand	−0.05	0.0008	0.10
Iceland	−0.17	0.0012	0.35	Tunisia	0.15	0.0008	0.12
Italy	0.19	0.0003	0.10	Turkey	0.14	0.0008	0.04
Japan	0.08	0.0009	0.10	USA	0.07	0.0007	0.14

Table 3. Estimates in random- and mixed-effects models using data from Else-Quest *et al.* (2010)

Method	$\hat{\tau}^2$	$\hat{\beta}_1$	$\hat{\tau}^2_{res}$	R^2
HE	0.0077	−0.3870	0.0061	.2120
HS	0.0052	−0.3849	0.0046	.1207
DL	0.0058	−0.3861	0.0054	.0691
SJ	0.0076	−0.3870	0.0061	.1891
ML	0.0069	−0.3858	0.0051	.2544
REML	0.0073	−0.3867	0.0058	.2060
EB	0.0075	–0.3868	0.0059	.2093

As the slope estimates show, a negative relationship was found with all methods, indicating that a higher percentage of women in parliament was associated with decreasing advantages for boys in the mathematics test. Regarding the total heterogeneity variance, the lowest estimates were obtained using HS and DL methods (0.0052 and 0.0058, respectively), while the highest estimates were provided by HE, SJ, and EB methods (0.0077, 0.0076, 0.0075, respectively). Residual heterogeneity variance estimates also showed some variability, with values ranging between 0.0046 (HS estimator) and 0.0061, obtained with the HE and SJ estimators. These differences led to notable variation among the estimates of the model predictive power depending on the estimator used. The R^2 values showed fluctuations from 6.9% of heterogeneity accounted for by the moderator (DL estimator) to the 25.4% obtained with the ML estimator.

7. Simulation study

A simulation study was programmed in R using the metafor (Viechtbauer, 2010) package. Meta-analyses of k studies were generated, obtaining the individual scores for each study from two normal populations (see Marín-Martínez & Sánchez-Meca, 2010) and using the standardized mean difference as the effect size index (Marín-Martínez & Sánchez-Meca, 2010; equation 2).

For each meta-analysis, $\boldsymbol{\theta}$ and $\mathbf{x}$ were defined as $k \times 1$ vectors containing parameter effects and moderator values, respectively. The predictor $\mathbf{x}$ was generated from a standard normal distribution. On the other hand, the $\boldsymbol{\theta}$ values were obtained from the expression $\boldsymbol{\theta} = \beta_0 + \beta_1\mathbf{x} + \mathbf{u}$, where β_0 was set to 0.5, which can be regarded as an effect of medium

size in some psychological areas (Cohen, 1988); the slope β_1 was set as described below, and **u** is an error term with distribution $N(0, \tau^2_{res})$. Note that if the predictor is dropped from the model, the error term **u** will have distribution $N(0, \tau^2)$.

The total heterogeneity variance, τ^2, and the model predictive power, P^2, were manipulated in the simulations. The former was set to values representative of no, low, medium, or large amounts of heterogeneity in psychology and related fields (0, 0.08, 0.16, and 0.32, respectively), similar to the values employed in previous simulation studies (e.g., Knapp & Hartung, 2003; Marín-Martínez & Sánchez-Meca, 2010; Schulze, 2004). For P^2, we used values of 0%, 25%, 50%, or 75% of heterogeneity accounted for, with the aim of reflecting realistic conditions (Thompson & Higgins, 2002). After setting both parameter values, we then assigned a value to β_1 by means of the expression $\beta_1^2 = \tau^2 P^2$. Table 4 gathers the different values considered for these parameters, as well as the resulting values for β_1^2 and the residual heterogeneity variance parameter, τ^2_{res}, which we computed as $\tau^2_{res} = \tau^2 - \beta_1^2$.[2]

Other factors manipulated in this simulation were the number of studies in each meta-analysis (k = 5, 10, 20, 40, and 80) and the average sample size of the k studies ($\bar{N}$ = 30, 50, 100, 150, and 200). Note that, for the ith study, $N_i = n_{iE} + n_{iC}$, with $n_{iE} = n_{iC}$. Vectors of individual sample sizes were generated with an skewness of +1.546, as reported by Sánchez-Meca and Marín-Martínez (1998, p. 317) in a review of meta-analytic syntheses in psychology. A total of 13 × 5 × 5 = 325 conditions were examined. For each condition, 10,000 meta-analyses were simulated, and $\hat{\tau}^2$, $\hat{\tau}^2_{res}$, and R^2 were computed with the seven alternatives above presented for each simulated data set.

The performance of the estimators for τ^2, τ^2_{res} and P^2 was compared using several criteria. Let $\hat{\theta}_i^j$ be an estimate of one of the parameters of interest obtained with any of the proposed methods in a particular condition. The bias for that estimate and condition was estimated as (Marín-Martínez & Sánchez-Meca, 2010)

$$\text{bias}\left(\hat{\theta}^j\right) = \frac{\sum_i \hat{\theta}_i^j}{10,000} - \theta, \tag{6}$$

where θ is the value of the parameter of interest (see Table 4). The percentage of bias, or relative bias, was then obtained as

$$\%\text{bias}\left(\hat{\theta}^j\right) = \frac{\text{bias}\left(\hat{\theta}^j\right)}{\theta} \times 100. \tag{7}$$

Moreover, the MSE was estimated as

$$\text{MSE}\left(\hat{\theta}^j\right) = \frac{\sum_i \left(\hat{\theta}_i^j - \theta\right)^2}{10,000}. \tag{8}$$

Finally, as described earlier, the computation of the R^2 value may require truncation in various cases. When τ^2 and τ^2_{res} are both actually positive (in which case $0 < P^2 < 1$), a

[2] From $\theta_i = \beta_1 X_i + u_i$, the total amount of heterogeneity in the true effect sizes, τ^2, can easily be computed with $\tau^2 = \beta_1^2 V(X_i) + \tau^2_{res} = \beta_1^2 + \tau^2_{res}$, as X_i and u_i are independent and normally distributed with mean zero and variances 1 and τ^2_{res}, respectively. This leads to the expression $\tau^2_{res} = \tau^2 - \beta_1^2$.

Table 4. Parameter values considered in this simulation for τ^2 and P^2 (and the resulting values for β_1^2 and τ^2_{res})

τ^2	0		0.08			0.16				0.32			
P^2	0	0	0.25	0.50	0.75	0	0.25	0.50	0.75	0	0.25	0.50	0.75
β_1^2	0	0	0.02	0.04	0.06	0	0.04	0.08	0.12	0	0.08	0.16	0.24
τ^2_{res}	0	0.08	0.06	0.04	0.02	0.16	0.12	0.08	0.04	0.32	0.24	0.16	0.08

large rate of truncated R^2 values would reflect undesirable performance of equation (3). Therefore, the proportion of R^2 values truncated to 0 or 1 was also examined for the different estimators along the simulated scenarios.

8. Results

Due to limitations of space, only part of the results will be presented in this section. The full set of results is available from the corresponding author upon request.

8.1. Total heterogeneity variance

Because any negative estimates of τ^2 were truncated to zero, all estimators showed the expected positive bias under the homogeneous scenario ($\tau^2 = 0$). On the other hand, for the conditions with $\tau^2 > 0$, Table 5 shows the percentage of bias for the total heterogeneity variance estimates provided by each method when setting the number of studies and the average within-study sample size to values that can often be found in meta-analytic reviews in several psychological fields (i.e., $k = 20$ and $\bar{N} = 50$).

The HS and ML estimators provided the most negatively biased estimates, with a deviation of around 16% from the parameter value. The SJ estimator showed the most (positively) biased results, although its performance improved as τ^2 increased. The DL and REML estimators performed similarly for small to medium amounts of heterogeneity, with a negative bias slightly over 5%, while the DL estimator yielded more biased results for large values of τ^2. The HE estimator showed the best results in terms of bias, with a positive deviation smaller than 3% and better results as the parameter value increased. Finally, the EB estimator performed reasonably well in terms of bias, with a negative deviation from the parameter value around 2%. With smaller values of k, all estimators showed a larger bias. Conversely, the estimates obtained with 40 and 80 studies were more accurate than with $k = 20$ for the different methods. Finally, higher average sample sizes also led to more accurate results for all estimators.

When comparing the estimators in terms of their relative efficiency, the SJ and HE methods provided the largest MSE values, while the HS and ML estimators showed the

Table 5. Percentage of bias for the total heterogeneity variance estimators with $k = 20$ and $\bar{N} = 50$

τ^2	HE	HS	DL	SJ	ML	REML	EB
0.08	2.81	−17.33	−6.23	45.19	−17.92	−6.96	−1.82
0.16	1.05	−16.03	−7.49	16.96	−14.27	−6.01	−2.36
0.32	0.47	−16.83	−9.83	4.69	−11.93	−5.25	−2.36

most efficient performance. The remaining estimators (DL, REML, and EB) performed similarly as k increased. All methods yielded more accurate estimates with a larger k, with MSE values clearly decreasing with 20 or more studies, and an increase in the average sample size per study also led to better results.

8.2. Residual heterogeneity variance

Trends for the different methods when estimating the residual heterogeneity variance were very similar to those detailed for τ^2. Regarding bias, the SJ estimator again showed the most biased results – the positive bias was now larger than for τ^2 – unless the parameter value was large enough ($\tau^2_{res} = 0.24$ and $\tau^2_{res} = 0.32$). Moreover, HS and ML methods provided again negatively biased estimates, with a deviation from the parameter value around 25% with 20 studies, larger than that observed for τ^2. Finally, the HE, DL, REML, and EB estimators performed similarly as for τ^2.

Figure 1 shows the MSE results for the estimators as a function of k and $\bar{N}$. The HS and ML methods performed very similarly, so their results are presented jointly, as are those for the REML and EB estimators. As found in the results for τ^2, the number of studies showed the largest influence on the efficiency of all estimators of τ^2_{res} and the MSE values especially decreased when going from 5 to 10 and from 10 to 20 studies. The average sample size also showed some influence on the efficiency of the estimates, with smaller MSE values obtained as $\bar{N}$ increased. The SJ and HE estimators showed the largest MSE values, while the HS and ML methods provided the most efficient estimates. All estimators except the SJ method performed similarly with $k = 80$.

8.3. Model predictive power

The R^2 values obtained with all estimators were quite variable, but the estimates tended to fall closer to the parameter value as k, $\bar{N}$, τ^2, and P^2 increased. As an illustration, Table 6

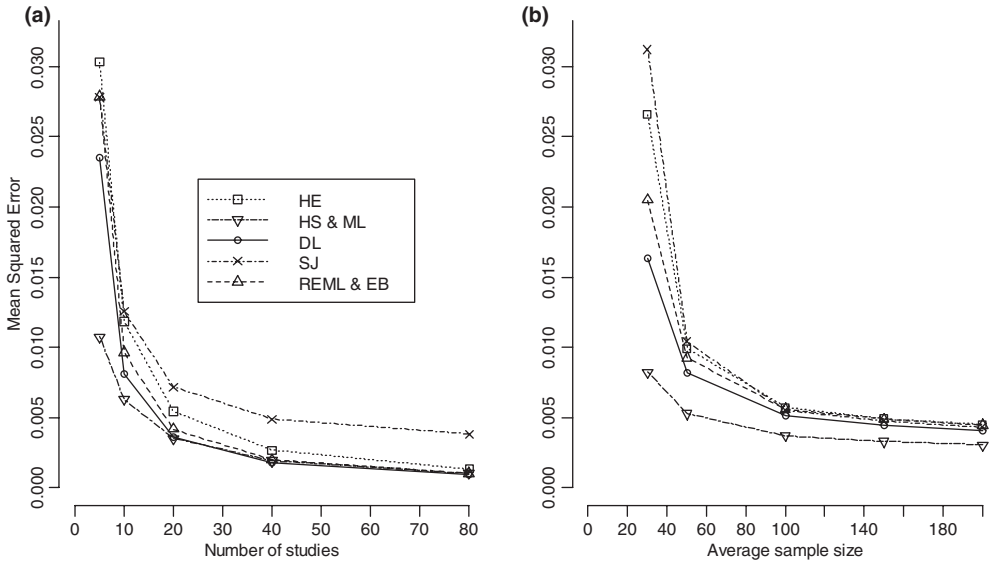

Figure 1. MSE for the residual heterogeneity variance estimators.

Table 6. Correlations between the R^2 values obtained with the different methods, for adverse conditions (lower triangle) and the optimal scenario (upper triangle)

	HE	HS	DL	SJ	ML	REML	EB
HE		.9727	.9731	.9934	.9958	.9960	.9991
HS	.7070		.9999	.9692	.9869	.9865	.9803
DL	.9368	.7201		.9701	.9871	.9868	.9807
SJ	.8227	.5720	.8196		.9907	.9915	.9935
ML	.7627	.8395	.7677	.5943		.9999	.9988
REML	.9322	.6796	.9516	.8221	.7591		.9989
EB	.9678	.6991	.9772	.8314	.7725	.9626	

presents the correlations between the estimates obtained with the different methods under two opposite scenarios. Figures below the main diagonal are correlations under adverse conditions ($k = 5, \bar{N} = 50, \tau^2 = 0.16$, and $P^2 = 0.25$), while those above the main diagonal are correlations obtained under an optimal scenario ($k = 80, \bar{N} = 100, \tau^2 = 0.32$, and $P^2 = 0.50$).

Under adverse conditions, the highest correlations were found between the DL, REML, and EB estimators, with values over .95, while most of the remaining combinations yielded values below .90 and even below .60 (e.g., the correlation between the HS and SJ estimators). Conversely, all estimators performed very similarly under the optimal scenario, with all correlations falling above .96. Table 6 shows, therefore, that the differences between estimators are especially important under the most adverse conditions, while performance for all methods tends to convergence for the optimal scenarios.

Among the different factors manipulated in this simulation, the accuracy of the P^2 estimates was mostly influenced by k. This finding is illustrated in Figure 2 using the EB estimator, which provided slightly more accurate results than the other methods, and considering scenarios with $P^2 = 0.25$.

The boxplots in Figure 2 reveal substantial variability in the P^2 estimates, especially for small values of k (e.g., less than 20 studies), represented on the *X*-axis of each chart. The picture is worrying for a typical meta-regression, as it reveals that no value between 0 and 1 (including a truncated estimate) is unlikely unless k is large enough (40 or more studies), especially with small to medium sample sizes ($\bar{N} = 50$) for the individual studies. Results with 5 studies, which are not shown in this figure, were very unstable, showing even more variability than with $k = 10$. Moreover, an increase in the average sample size per study led to more precise estimates (as can be seen when looking at Figure 1(b)), while increasing the heterogeneity variance parameter, represented with different bar shades, led to a smaller rate of truncation of the R^2 values to zero and one.

Several descriptives were computed for the R^2 values obtained with the different estimators, considering conditions with $k = 40$ and setting the other factors to realistic values for a meta-regression with one covariate ($\bar{N} = 50, \tau^2 = 0.16$, and $P^2 = 0.25$). Table 7 gathers the mean, the median, the 2.5 and 97.5 percentiles, and the rates of values truncated to zero and one for each estimator.

Regarding the comparison of the different estimators in terms of bias, the HE, DL, REML, and EB estimators performed appropriately, with their mean estimates deviating less than 0.01 from the parameter value ($P^2 = 0.25$). In contrast, the HS and ML estimators

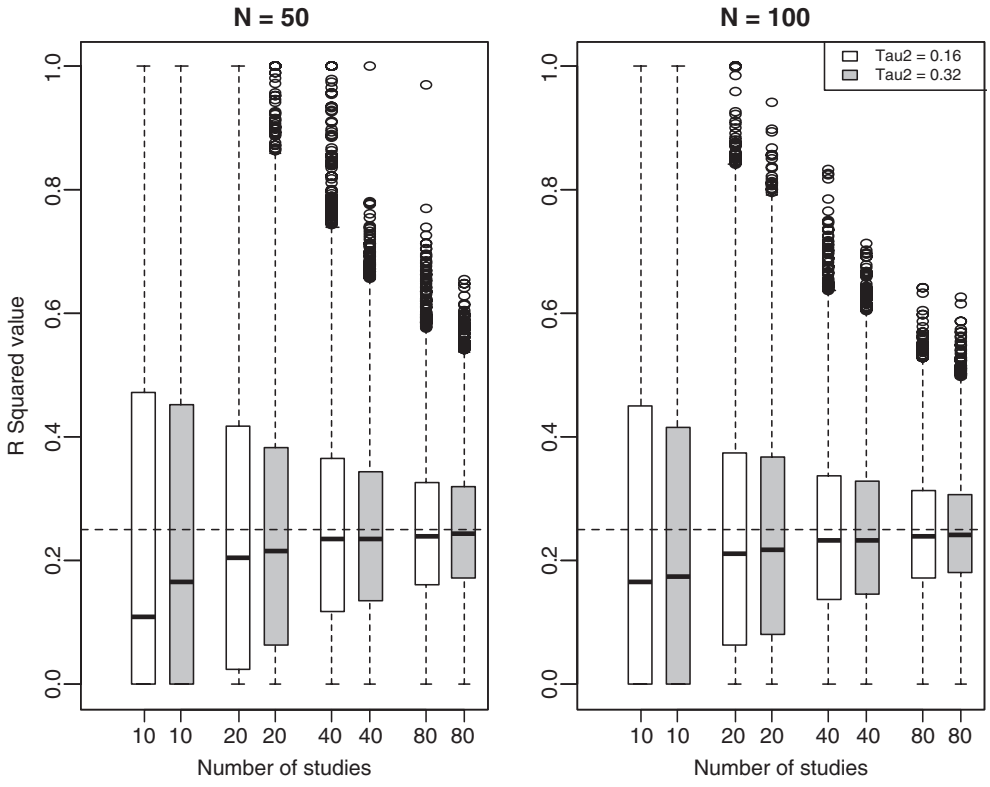

Figure 2. R^2 values using the EB estimator with $P^2 = 0.25$.

showed a positive bias, while the mean estimate for the SJ estimator showed a large negative bias.[3]

In addition to the bias that was found for the HS, ML, and SJ estimators, the remaining methods showed some problems as well. When examining the percentiles presented in Table 7, it can be seen that there was a wide variation among the individual estimates, and that 95% of the central values ranged from 0 to 0.65. Moreover, a non-negligible proportion of the estimates (over 5%) were truncated to zero, especially for the DL and REML estimators. While the rates of truncation to zero were clearly lower for the HS and ML estimators, the bias showed by these two methods advises against their use. Finally, despite the parameter value of $P^2 = 0.25$, the HE, DL, REML, and EB methods still provided some estimates that were truncated to one. On the other hand, since the SJ estimator always yields a positive value, R^2 can never reach 1 when using this estimator and hence never required truncation at the upper end of the scale, although in turn it provided the largest bias.

Table 8 presents the MSE results with $k = 40$ and $P^2 = 0.25$ for the different estimators. Only conditions with some heterogeneity among the parameter effects ($\tau^2 > 0$) were considered here.

[3] Since the negative bias for the HS and ML estimators and the positive bias for the SJ estimator tended to be larger for τ^2_{res} than for τ^2, the bias for these three methods was reversed when estimating P^2.

Table 7. R^2 values with $k = 40$, $\bar{N} = 50$, $\tau^2 = 0.16$, and $P^2 = 0.25$

Estimator	HE	HS	DL	SJ	ML	REML	EB
Mean	.2534	.2950	.2488	.1464	.3052	.2588	.2555
$P_{2.5}$	0	.0157	0	0	.0166	0	0
Median	.2311	.2752	.2281	.1332	.2843	.2390	.2351
$P_{97.5}$	.6512	.6974	.6458	.3734	.7379	.6781	.6547
Pr ($R^2 = 0$)	.0585	.0003	.0689	.0570	.0062	.0630	.0565
Pr($R^2 = 1$)	.0021	.0029	.0017	0	.0011	.0010	.0015

Notes. $P_{2.5}$ and $P_{97.5}$ are 2.5 and 97.5 percentiles, respectively. $\Pr(R^2 = i)$ is the rate of values truncated to i, for $i = 0, 1$.

Table 8. MSE values for the P^2 estimators with $k = 40$ and $P^2 = 0.25$

Estimator		HE	HS	DL	SJ	ML	REML	EB
$\bar{N} = 30$	$\tau^2 = 0.08$	.0932	.1227	.0997	.0395	.1495	.1187	.1043
	$\tau^2 = 0.16$	.0678	.0807	.0667	.0306	.0984	.0776	.0662
	$\tau^2 = 0.32$	.0377	.0417	.0390	.0234	.0482	.0415	.0363
$\bar{N} = 50$	$\tau^2 = 0.08$	.0641	.0758	.0625	.0292	.0872	.0696	.0634
	$\tau^2 = 0.16$	.0322	.0345	.0323	.0218	.0373	.0335	.0317
	$\tau^2 = 0.32$	.0218	.0225	.0231	.0174	.0231	.0228	.0220
$\bar{N} = 100$	$\tau^2 = 0.08$	.0285	.0300	.0285	.0206	.0312	.0292	.0285
	$\tau^2 = 0.16$	.0202	.0197	.0202	.0165	.0202	.0204	.0203
	$\tau^2 = 0.32$	.0172	.0161	.0171	.0149	.0168	.0173	.0172
$\bar{N} = 150$	$\tau^2 = 0.08$	.0230	.0228	.0229	.0182	.0234	.0232	.0231
	$\tau^2 = 0.16$	.0179	.0168	.0176	.0153	.0175	.0180	.0179
	$\tau^2 = 0.32$	.0164	.0148	.0159	.0146	.0159	.0165	.0164
$\bar{N} = 200$	$\tau^2 = 0.08$	.0199	.0193	.0197	.0165	.0197	.0200	.0199
	$\tau^2 = 0.16$	.0175	.0164	.0171	.0151	.0171	.0176	.0175
	$\tau^2 = 0.32$	.0159	.0142	.0153	.0145	.0153	.0159	.0159

All methods performed more efficiently as $\bar{N}$ and τ^2 increased. When comparing the different methods, the ML and HS estimators provided the largest MSE values unless the average sample size per study was 150 or 200 participants, while the SJ estimator was the most efficient method, especially under the most adverse conditions. Regarding the influence of k, weak performance was reported before for the method proposed by Raudenbush (1994) with a small number of studies (see Figure 2 and Table 7). With $k = 20$, trends were already similar to those shown in Table 8, although the MSE values were twice as large as for $k = 40$. With $k = 80$, MSEs were on average smaller than 0.04 under all of the conditions examined here, although trends for the different estimators remained the same.

9. Discussion

In this study, the performance of seven methods for the estimation of the total and residual heterogeneity variances, as well as the model predictive power, was assessed under a variety of realistic scenarios in applied research. The estimators here compared showed different performance, especially under adverse and intermediate conditions, while all

methods provided similar and accurate estimates of the parameters of interest for the most favourable conditions (e.g., large number of studies and large number of participants per study).

Regarding the results for the total heterogeneity variance, the patterns found in this simulation are comparable to the ones reported by Viechtbauer (2005). The DL, REML, and EB estimators performed reasonably well in terms of bias and efficiency, although the DL method yielded negatively biased estimates for large parameter values, as was found in previous simulations (Malzahn *et al.*, 2000; Sidik & Jonkman, 2005b, 2007; Viechtbauer, 2005). The HE estimator showed essentially unbiased results (the slight positive bias observed in Table 5 can be regarded as a consequence of truncating the negative estimates to zero) but large MSE values, while the HS and ML methods performed very efficiently but with a negative bias. Finally, the SJ method showed a large positive bias for small parameter values, as has been previously described (Sidik & Jonkman, 2005b), and the largest MSE values. The performance of the various estimators remained very similar after the inclusion of a moderator.

Regarding the estimation of the predictive power in meta-regression models with one predictor, no estimator performed accurately with less than 40 studies. Again, the HS, ML, and SJ estimators yielded the most biased estimates. The remaining estimators performed more precisely, although their estimates still showed wide variation even with a moderate to large k, including truncated values to zero and one, as shown in Table 7. Given the large MSE of the SJ estimator for τ^2 and τ^2_{res}, the SJ estimator showed surprisingly efficient performance for estimating P^2, while the HS and ML methods now provided the largest MSE values.

Out of the different factors manipulated in this simulation, our results suggest that the number of studies exerts an important influence on the accuracy of the results, and that precise estimates of the heterogeneity variances and the model predictive power can only be expected with at least 20 and 40 studies, respectively. An increase in the average sample size also improved the results for all estimators. The critical influence of k on the accuracy of the heterogeneity variance estimators has already been discussed by several authors both in the context of random-effects models (e.g., Borenstein *et al.*, 2009; Schulze, 2004) and mixed-effects models (Thompson & Higgins, 2002). The fact that results were more accurate as k and $\bar{N}$ increased is in agreement with large-sample theory, which underlies the statistical models and methods in meta-analysis (Hedges, 2009). Moreover, as shown in Figure 2 and Table 8, the P^2 estimators performed more efficiently as the total heterogeneity variance increased. An explanation of this fact is that, when estimating τ^2, a small parameter value will lead more often to negative estimates requiring truncation, and this will also lead to truncated R^2 values.

10. Conclusion

When a meta-analysis is carried out, some variability is usually found among the effect sizes from the individual studies. The part of that variability due to systematic differences among studies can be quantified by estimating the heterogeneity (or between-studies) variance, τ^2. Moreover, if the results from the studies are not homogeneous, the meta-analyst may be interested in the identification of one or more study characteristics that can explain part of the variability among the results. This goal can be addressed through meta-regression analyses, which are typically conducted under a mixed-effects model. Two parameters of interest in a mixed-effects meta-regression model are the residual

heterogeneity variance after including one or more moderators, τ^2_{res}, and the predictive power of the moderator(s) included in the model, P^2.

In the present simulation study, we found a different performance for the seven estimation methods available both for τ^2 and τ^2_{res} in the set of simulated conditions. For a small number of studies ($k < 20$) no estimator performed accurately. When the number of studies was moderate (20–40 studies), the REML and EB methods yielded the most accurate results when considering bias and efficiency criteria jointly. Finally, with 80 studies, all methods converged and showed similar (and accurate) results. Increasing the average sample size per study also led to more accurate results. These results are of interest not only for the accurate estimation of the heterogeneity variances, but also for the computation of an R^2 index in meta-analysis, which can be obtained by comparing the $\hat{\tau}^2$ and $\hat{\tau}^2_{res}$ values (Raudenbush, 1994).

The results obtained in this simulation study suggest that about 40 studies are required to get accurate estimates of P^2 in mixed-effects meta-regression models, so that a cautious interpretation of R^2 values should be advised for meta-regression models fitted with a smaller number of studies (Thompson, 1994). Among the different estimators here compared, the REML, DL, and EB methods showed the most accurate results across the different scenarios and criteria considered. Although the present study focused on standardized mean differences, it is likely that our findings can be generalized to meta-analyses with other effect size measures that are (at least approximately) normally distributed. However, conclusions from this simulation are restricted to the scenarios considered here, so that further simulation studies are needed in order to account for conditions different from the ones included in the present study.

Acknowledgement

This research was supported by a grant from the Fundación Séneca, Region of Murcia, Spain, and by the Ministerio de Economía y Competitividad and FEDER funds from the Spanish Government, Project No. PSI2012-31399.

References

Aloe, A. M., Becker, B. J., & Pigott, T. D. (2010). An alternative to R^2 for assessing linear models of effect size. *Research Synthesis Methods*, *1*, 272–283. doi:10.1002/jrsm.23

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd ed.) Hillsdale, NJ: Erlbaum.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.) (2009). *The handbook of research synthesis and meta-analysis*. (2nd ed.) New York: Russell Sage Foundation.

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods*, *14*, 165–176. doi:10.1037/a0015565

DerSimonian, R., & Laird, N. (1986). Meta-analysis of clinical trials. *Controlled Clinical Trials*, 7, 177–188. doi:10.1016/j.cct.2006.04.004

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, *136*, 103–127. doi:10.1037/a0018851

Hedges, L. V. (2009). Statistical considerations. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 37–47). New York: Russell Sage Foundation.

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

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods*, *9*, 426–445. doi:10.1037/1082-989X.9.4.426

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

Jennrich, R. I., & Sampson, P. F. (1976). Newton-Raphson and related algorithms for maximum likelihood variance component estimation. *Technometrics*, *18*, 11–17. doi:10.2307/1267911

Knapp, G., Biggerstaff, B. J., & Hartung, J. (2006). Assessing the amount of heterogeneity in random-effects meta-analysis. *Biometrical Journal*, *48*, 271–285. doi:10.1002/bimj.200510175

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, *22*, 2693–2710. doi:10.1002/sim.1482

Konstantopoulos, S., & Hedges, L. V. (2009). Analyzing effect sizes: Fixed-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 279–293). New York: Russell Sage Foundation.

Lipsey, M. W. (2009). Identifying interesting variables and analysis opportunities. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 147–158). New York: Russell Sage Foundation.

Malzahn, U., Böhning, D., & Holling, H. (2000). Nonparametric estimation of heterogeneity variance for the standardised difference used in meta-analysis. *Biometrika*, *87*, 619–632. doi:10.1093/biomet/87.3.619

Marín-Martínez, F., & Sánchez-Meca, J. (2010). Weighting by inverse variance or by sample size in random-effects meta-analysis. *Educational and Psychological Measurement*, *70*, 56–73. doi:10.1177/0013164409344534

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, *78*, 47–55. doi:10.2307/2287098

National Research Council (1992). *Combining information: Statistical issues and opportunities for research*. Washington, DC: National Academy Press.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated analysis*. Hillsdale, NJ: Erlbaum.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper, & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–321). New York: Russell Sage Foundation.

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

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods*, *11*, 306–322. doi:10.1037/1082-989X.11.3.306

Sánchez-Meca, J., & Marín-Martínez, F. (1998). Testing continuous moderators in meta-analysis: A comparison of procedures. *British Journal of Mathematical and Statistical Psychology*, *51*, 311–326. doi:10.1111/j.2044-8317.1998.tb00683.x

Sánchez-Meca, J., & Marín-Martínez, F. (2008). Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods*, *13*, 31–48. doi:10.1037/1082-989X.13.1.31

Sánchez-Meca, J., & Marín-Martínez, F. (2010). Meta-analysis. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed., Vol. 7, pp. 274–282). Oxford, UK: Elsevier.

Schmidt, F. L., Oh, I.-S., & Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology*, *62*, 97–128. doi:10.1348/000711007X255327

Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Toronto, Canada: Hogrefe & Huber.

Sidik, K., & Jonkman, J. N. (2005a). A note on variance estimation in random effects meta-regression. *Journal of Biopharmaceutical Statistics*, *15*, 823–838. doi:10.1081/BIP-200067915

Sidik, K., & Jonkman, J. N. (2005b). Simple heterogeneity variance estimation for meta-analysis. *Applied Statistics*, *54*, 367–384. doi:10.1111/j.1467-9876.2005.00489.x

Sidik, K., & Jonkman, J. N. (2007). A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine*, *26*, 1964–1981. doi:10.1002/sim.2688

Thompson, S. G. (1994). Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal*, *309*, 1351–1355. doi:10.1136/bmj.309.6965.1351

Thompson, S. G., & Higgins, J. P. T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*, *21*, 1559–1573. doi:10.1002/sim.1187

Thompson, S. G., & Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine*, *18*, 2693–2708. doi:10.1002/(SICI)1097-0258(19991030)18:20<2693::AID-SIM235>3.0.CO;2-V

Van den Noortgate, W., & Onghena, P. (2003). Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational and Psychological Measurement*, *63*, 765–790. doi:10.1177/0013164402251027

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

Viechtbauer, W. (2007a). Accounting for heterogeneity via random-effects models and moderator analyses in meta-analysis. *Zeitschrift für Psychologie/Journal of Psychology*, *215*, 104–121. doi:10.1027/0044-3409.215.2.104

Viechtbauer, W. (2007b). Hypothesis tests for population heterogeneity in meta-analysis. *British Journal of Mathematical and Statistical Psychology*, *60*, 29–60. doi:10.1348/000711005X64042

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

Viechtbauer, W., López-López, J. A., Sánchez-Meca, J., & Marín-Martínez, F. (2012). *A comparison of procedures to test for moderators in meta-regression models*. Unpublished manuscript.

Received 31 July 2012; revised version received 7 November 2012