Psychological Methods
2015, Vol. 20, No. 3, 360–374

1082-989X/15/$12.00 http://dx.doi.org/10.1037/met0000023

A Comparison of Procedures to Test for Moderators in Mixed-Effects Meta-Regression Models

Wolfgang Viechtbauer
Maastricht University

José Antonio López-López
University of Bristol

Julio Sánchez-Meca and Fulgencio Marín-Martínez
University of Murcia

Several alternative methods are available when testing for moderators in mixed-effects meta-regression models. A simulation study was carried out to compare different methods in terms of their Type I error and statistical power rates. We included the standard (Wald-type) test, the method proposed by Knapp and Hartung (2003) in 2 different versions, the Huber–White method, the likelihood ratio test, and the permutation test in the simulation study. These methods were combined with 7 estimators for the amount of residual heterogeneity in the effect sizes. Our results show that the standard method, applied in most meta-analyses up to date, does not control the Type I error rate adequately, sometimes leading to overly conservative, but usually to inflated, Type I error rates. Of the different methods evaluated, only the Knapp and Hartung method and the permutation test provide adequate control of the Type I error rate across all conditions. Due to its computational simplicity, the Knapp and Hartung method is recommended as a suitable option for most meta-analyses.

Keywords: meta-analysis, meta-regression, moderator analysis, heterogeneity estimator, standardized mean difference

Supplemental materials: http://dx.doi.org/10.1037/met0000023.supp

A meta-analysis is a form of systematic review using statistical methods to integrate the results of a set of related studies about a given topic (Cooper, Hedges, & Valentine, 2009). To accomplish this objective, studies fulfilling certain inclusion criteria are obtained and an effect size estimate is extracted from each study. An overall/average effect size combining all of the individual estimates can then be computed. However, in practice, the effect size estimates are often found to be more variable than would be expected based on sampling variability alone. This suggests that there are differences (*heterogeneity*) in the true effect sizes of the individual studies. Heterogeneity may be purely random or, at least in part, a result of systematic differences between the studies (in terms of design or sample characteristics) that are related to the size of the effect (Raudenbush, 2009). Therefore, meta-analysts often examine to what extent the heterogeneity in the effect sizes can be accounted for based on various study characteristics (*moderators*).

The process of examining the relationship between study characteristics and the effect sizes is typically called a *moderator analysis*. While simple subgrouping of the studies can be used for that purpose (Borenstein, Hedges, Higgins, & Rothstein, 2009), meta-analysts increasingly employ so-called meta-regression models to study one or multiple moderating variables, where the effect size estimates are used as the dependent and the moderators as the independent variables. In addition, a random effect is typically included in such models to account for any *residual heterogeneity* that is not accounted for by the moderators included in the model (Thompson & Sharp, 1999). Since the predictors included in the model are usually added as fixed effects, this approach then leads to a mixed-effects meta-regression model.

When fitting such meta-regression models, it is therefore necessary to estimate not only the model coefficients, but also the amount of residual heterogeneity in the effect sizes. At least seven methods have been proposed in the literature for estimating this parameter (e.g., Raudenbush, 2009; Thompson & Sharp, 1999), including the Hedges, DerSimonian–Laird, Sidik and Jonkman, maximum likelihood, restricted maximum likelihood, and empirical Bayes estimators (these estimators are described in more detail below). Once the model has been fitted, the individual model coefficients can be examined to determine the extent to which the

This article was published Online First August 11, 2014.

Wolfgang Viechtbauer, Department of Psychiatry and Neuropsychology, Maastricht University; José Antonio López-López, School of Social and Community Medicine, University of Bristol; Julio Sánchez-Meca and Fulgencio Marín-Martínez, Department of Basic Psychology and Methodology, University of Murcia.

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

Correspondence concerning this article should be addressed to Wolfgang Viechtbauer, Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Faculty of Health, Medicine, and Life Sciences, Maastricht University, PO Box 616 (VIJV1), 6200 MD Maastricht, the Netherlands. E-mail: wolfgang.viechtbauer@maastrichtuniversity.nl

moderators are related to the effect sizes. It is also customary in this context to test the model coefficients for statistical significance.

The standard (Wald-type) method for testing the significance of the model coefficients (e.g., Raudenbush, 2009) does not take into account the fact that the amount of residual heterogeneity has to be estimated based on the data at hand (Thompson & Higgins, 2002). This may lead to either inflated or overly conservative Type I error rates and possibly incorrect conclusions about the statistical significance of the moderator variables. Knapp and Hartung (2003) suggested an alternative method for testing the model coefficients that accounts for the imprecision in the estimated amount of residual heterogeneity and that yields rejection rates closer to the nominal significance level. Similar results were also found by Sidik and Jonkman (2005).

Another approach to possibly improve on the standard method uses a robust (Huber–White) estimate of the variance–covariance matrix of the model coefficients (Raudenbush, 2009; Sidik & Jonkman, 2005), which, in principle, should perform acceptably when the number of studies included in the meta-analysis is large. However, a simulation study by Sidik and Jonkman (2005) using log risk ratios as the effect size measure suggested that this method does not consistently bring the Type I error rate closer to the nominal significance level. It remains to be determined whether the method holds any promise for effect size measures more commonly used in the social and behavioral sciences (in particular, correlations and standardized mean differences). Moreover, a simple correction factor may help to improve the accuracy of this method when the number of studies is low (Hedges, Tipton, & Johnson, 2010).

In the context of random-effects models (i.e., in models without moderators), the use of likelihood-based methods has been suggested as an alternative for testing and obtaining confidence intervals (CIs) for the overall/average effect (Hardy & Thompson, 1996). The extension to mixed-effects models is straightforward, leading to likelihood ratio tests of the model coefficients. Recent findings suggest that this approach provides slightly better control of the Type I error rate when compared to standard Wald-type tests (Huizenga, Visser, & Dolan, 2011).

Finally, Follmann and Proschan (1999) and Higgins and Thompson (2004) proposed the use of permutation methods for testing the significance of the overall/average effect size and when testing moderator variables in meta-regression models. Results from simulation studies by Follmann and Proschan and Huizenga et al. (2011) suggest that permutation tests may perform close to acceptable levels. However, the resampling method examined by Huizenga et al. for testing moderator variables was based on permutations of the residuals, a slightly different approach than the simpler permutation test suggested by Higgins and Thompson. The performance of the latter has, to our knowledge, never been examined in a systematic manner.

In summary, various methods have been proposed in the meta-analytic literature to estimate the amount of residual heterogeneity and to perform statistical tests of the model coefficients in mixed-effects meta-regression models. One can easily find examples where the different methods, applied to the same meta-analytic data set, lead to different results and even conflicting conclusions. Therefore, an important objective is to compare their performance in terms of Type I error rates and statistical power in order to decide which methods are to be preferred. In the present article, the various methods described above are compared, using the standardized mean difference as the effect size measure.

In the next section, the mixed-effects meta-regression model is outlined, followed by a description of several different residual heterogeneity variance estimators proposed in the literature and different methods for conducting significance tests of the model coefficients. We then describe the methods and results from a Monte Carlo simulation study comparing the performance of the different methods. Next, we use an example to illustrate the various methods, showing how conclusions about the relevance of a moderator can be affected by the chosen method. The article then finishes with a discussion of the main results, as well as their implications for carrying out a meta-analysis.

Mixed-Effects Meta-Regression

In a meta-analysis with k independent studies, let y_i denote the observed effect size estimate in the ith study. The mixed-effects meta-regression model (e.g., Raudenbush, 2009) is then given by the expression

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots \beta_p x_{ip} + u_i + e_i, \quad (1)$$

where x_{ij} denotes the value of the jth moderator in the ith study, β_j represents the corresponding model coefficient indicating how the size of the effect changes as x_{ij} increases by one unit, and β_0 stands for the model intercept. Furthermore, u_i denotes a random effect with distribution $N(0,\tau^2)$ and e_i the within-study error with distribution $N(0,v_i)$. The v_i terms denote the within-study sampling variances of the studies and are assumed to be known. The amount of residual heterogeneity is denoted by τ^2, which indicates the variability in the true effects not accounted for by the moderators in the model.

In matrix notation, the model can be compactly written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \mathbf{e}, \quad (2)$$

where $\mathbf{y}$ denotes the $(k \times 1)$ column vector with the k effect size estimates. The first column of the $(k \times (p + 1))$ matrix $\mathbf{X}$ contains a vector of ones, corresponding to the model intercept, while the remaining columns contain the values of the p moderator variables. Finally, $\boldsymbol{\beta}$ is a $((p + 1) \times 1)$ column vector with the regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$, and $\mathbf{u}$ and $\mathbf{e}$ are $(k \times 1)$ vectors with the u_i and e_i values. We assume that the number of moderators p in the model is limited and that $\mathbf{X}$ is of full rank, so that $\mathbf{X}'\mathbf{X}$ is invertible. Therefore, any perfect (multi)collinearity among the moderators must be removed (e.g., by considering a simpler model) before Equation 2 can be fitted.

A special case of Equation 2 is the random-effects model where $\mathbf{X}$ only contains a column of ones (i.e., a model without moderators). In that case, $\beta_0 \equiv \mu$ reflects the average true effect and τ^2 the total amount of heterogeneity in the effect sizes. Note that under normality assumptions regarding $\mathbf{u}$ and $\mathbf{e}$, Equations 1 and 2 imply that $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \tau^2\mathbf{I} + \mathbf{V})$, where $\mathbf{V}$ is diagonal with elements v_i.

The regression coefficients are typically estimated using weighted least squares, by means of the equation

$$\mathbf{b} = (\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{W}}\mathbf{y}, \quad (3)$$

with $\hat{\mathbf{W}}$ denoting a $(k \times k)$ diagonal weight matrix with elements $w_i = 1/(v_i + \hat{\tau}^2)$. Note that $\hat{\tau}^2$ is an estimate of the unknown value

τ^2. Methods for estimating this parameter are detailed in the next section.

Heterogeneity Estimators in the Mixed-Effects Model

Several alternative methods have been proposed in the literature for estimating τ^2 in the random-effects model (López-López, Marín-Martínez, Sánchez-Meca, Van den Noortgate, & Viechtbauer, 2014; Sánchez-Meca & Marín-Martínez, 2008; Viechtbauer, 2005). Most of these estimators have also been extended to the mixed-effects model. In this section, we describe seven estimators for the latter case. Four of these estimators are noniterative, while three require iterative computations. All the estimators can be succinctly expressed after defining the matrix

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

where $\mathbf{W}$ is a diagonal weight matrix whose elements can change from one estimator to another. For example, for the Hedges estimator, $\mathbf{W}$ is defined as the identity matrix. When weights are included, noniterative estimators (e.g., DerSimonian–Laird estimator) make use of the inverse of the within-study sampling variances, while for the iterative estimators (e.g., maximum likelihood estimator) $\mathbf{W}$ contains the inverse variances plus an estimate of the amount of residual heterogeneity. Further details on the elements of $\mathbf{W}$ for each residual heterogeneity estimator are provided below.

Moreover, as will be seen on the basis of the set of equations to be presented in this section, the underlying logic for all methods is to estimate the residual heterogeneity based on the difference or ratio between some estimate of the total variability among the population effect sizes not accounted for by the explanatory variables included in the model and the amount of variability expected from random sampling error alone.

In particular, the total variability not accounted for by the explanatory variables is expressed as a quadratic form of the effect size estimates and (a function of) the $\mathbf{P}$ matrix. For example, with the elements of the diagonal weight matrix $\mathbf{W}$ set equal to $w_i = 1/v_i$, we obtain the residual heterogeneity statistic $Q_E = \mathbf{y}'\mathbf{P}\mathbf{y}$ (Hedges & Olkin, 1985), which is simply equal to the residual sums of squares under weighted least squares estimation (e.g., Christensen, 1996).[1] On the other hand, the amount of sampling variability in the effect size estimates is given by the v_i values, which we can collect in the diagonal matrix $\mathbf{V}$. The value of the quadratic form, relative to (some function of) $\mathbf{V}$ and/or the degrees of freedom of the model under assessment (i.e., $df = k - p - 1$), then provides an estimate of the residual amount of heterogeneity.

Hedges (HE) Estimator

Hedges (1983; see also Hedges & Olkin, 1985) proposed a method of moments estimator of τ^2 in the random-effects model based on ordinary least squares estimation. The estimate is obtained by calculating the difference between an unweighted estimate of the total variance of the effect sizes and an unweighted estimate of the average within-study variance (Sánchez-Meca & Marín-Martínez, 2008). When moderators are included in the model, the extension of the HE estimator (Raudenbush, 2009) can be written as

$$\hat{\tau}^2_{HE} = \frac{\mathbf{y}'\mathbf{P}\mathbf{y} - tr(\mathbf{P}\mathbf{V})}{k - p - 1}, \tag{5}$$

with $tr()$ denoting the trace of the matrix in between the parentheses and with $\mathbf{W}$ equal to a $(k \times k)$ identity matrix $\mathbf{I}$ for the calculation of $\mathbf{P}$ (in which case Equation 4 simplifies to $\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$).[2] It is possible that $\hat{\tau}^2_{HE}$ turns out to be negative, which is a value outside the parameter space for a variance component. In this case, the value is truncated to 0.

Hunter and Schmidt (HS) Estimator

Hunter and Schmidt (2004) proposed an estimator of τ^2 in the random-effects model, which, in essence, is given by

$$\hat{\tau}^2_{HS} = \frac{\sum w_i(y_i - \hat{\theta})^2}{\sum w_i} - \frac{\sum w_i v_i}{\sum w_i} = \frac{\sum w_i(y_i - \hat{\theta})^2 - k}{\sum w_i}, \tag{6}$$

where $\hat{\theta} = \sum w_i y_i / \sum w_i$ and $w_i = 1/v_i$ (Viechtbauer, 2005). Note that Equation 6 can be regarded as the difference between a weighted estimate of the total variance of the effect sizes and a weighted average of the within-study variances. Although no extension has been suggested yet for this estimator when one or more covariates are included in the model, a logical proposal for computing this estimator in mixed-effects models is given by

$$\hat{\tau}^2_{HS} = \frac{\mathbf{y}'\mathbf{P}\mathbf{y}}{tr(\mathbf{W})} - \frac{tr(\mathbf{W}\mathbf{V})}{tr(\mathbf{W})} = \frac{\mathbf{y}'\mathbf{P}\mathbf{y} - k}{tr(\mathbf{W})}, \tag{7}$$

with $\mathbf{P}$ again defined in Equation 4 and the diagonal elements of $\mathbf{W}$ given by $w_i = 1/v_i$ (note that $\mathbf{y}'\mathbf{P}\mathbf{y} = \sum w_i(y_i - b_0 - b_1 x_{i1} - \ldots - b_p x_{ip})^2$, where $b_0, b_1, \ldots, b_p$ are the estimates obtained with $\mathbf{b} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$ and $tr(\mathbf{W}) = \sum w_i$, which shows how Equation 7 generalizes the HS estimator to mixed-effects models). When $\hat{\tau}^2_{HS}$ turns out negative, it is truncated to 0.

DerSimonian and Laird (DL) Estimator

The estimator proposed by DerSimonian and Laird (1986) for random-effects models, probably the most widely employed in meta-analyses up to date, is also based on the method of moments (DerSimonian & Kacker, 2007) in the context of weighted least squares estimation. When including covariates in the model, the estimator is given by

$$\hat{\tau}^2_{DL} = \frac{\mathbf{y}'\mathbf{P}\mathbf{y} - (k - p - 1)}{tr(\mathbf{P})}, \tag{8}$$

[1] Note that the Q_E statistic provides the usual test for residual heterogeneity that, under the null hypothesis $\tau^2 = 0$, follows a chi-square distribution with degrees of freedom equal to $df = k - p - 1$.

[2] On the basis of the properties of quadratic forms (e.g., Christensen, 1996), it follows from $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \tau^2\mathbf{I} + \mathbf{V})$ that $E[\mathbf{y}'\mathbf{P}\mathbf{y}] = tr(\mathbf{P}(\tau^2\mathbf{I} + \mathbf{V})) + (\mathbf{X}\boldsymbol{\beta})'\mathbf{P}(\mathbf{X}\boldsymbol{\beta})$. Since $(\mathbf{X}\boldsymbol{\beta})'\mathbf{P} = \mathbf{0}$, the expectation simplifies to $tr(\mathbf{P}(\tau^2\mathbf{I} + \mathbf{V})) = \tau^2 tr(\mathbf{P}) + tr(\mathbf{P}\mathbf{V})$. With $\mathbf{W} = \mathbf{I}$, $tr(\mathbf{P}) = tr(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = tr(\mathbf{I}) - tr(\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) = k - p - 1$. Therefore, we find that $E[\mathbf{y}'\mathbf{P}\mathbf{y}] = \tau^2(k - p - 1) + tr(\mathbf{P}\mathbf{V})$. Dropping the expectation and rearranging the equation then leads to the HE estimator.

with the diagonal elements of $\mathbf{W}$ again given by $w_i = 1/v_i$ (Knapp & Hartung, 2003; Raudenbush, 2009; Sidik & Jonkman, 2005).[3] A negative value of $\hat{\tau}^2_{DL}$ is again truncated to 0.

Sidik and Jonkman (SJ) Estimator

Another alternative for estimating the residual variance component was proposed by Sidik and Jonkman (2005) and is also based on weighted least squares estimation. The SJ estimator is obtained by starting with an initial (rough) estimate of τ^2, denoted by $\hat{\tau}^2_0$ and given by

$$\hat{\tau}^2_0 = \frac{\sum (y_i - \bar{y})^2}{k}, \tag{9}$$

which is then updated with the expression

$$\hat{\tau}^2_{SJ} = \frac{\mathbf{y}'\mathbf{P}\mathbf{y}}{k - p - 1}, \tag{10}$$

with $w_i = \hat{\tau}^2_0/(v_i + \hat{\tau}^2_0)$ for the diagonal elements of $\mathbf{W}$.[4] The SJ estimator always provides a nonnegative value and therefore does not require truncation (note that $\hat{\tau}^2_{SJ} = 0$ can only happen if all of the y_i values are exactly identical to each other, which theoretically is not possible, but could happen in practice, for example, if k is small and y_i values are rounded).

Maximum Likelihood (ML) Estimator

For $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \tau^2\mathbf{I} + \mathbf{V})$, the log-likelihood function of the parameter vector $(\boldsymbol{\beta}, \tau^2)$ is given by

$$ll_{ML} = -\frac{1}{2}k\ln(2\pi) - \frac{1}{2}\ln|\tau^2\mathbf{I} + \mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{11}$$

with the diagonal elements of $\mathbf{W}$ equal to $w_i = 1/(v_i + \tau^2)$. The ML estimator of $(\boldsymbol{\beta}, \tau^2)$ is that set of values that maximizes ll_{ML} under the constraint that $\tau^2 \geq 0$. The problem of finding the ML estimates is considerably simplified after realizing that Equation 3 actually corresponds to the ML estimator of $\boldsymbol{\beta}$ for a given value of τ^2. Therefore, we can substitute Equation 3 into Equation 11, which, after some simplification, yields the profiled log-likelihood

$$ll_{ML} = -\frac{1}{2}k\ln(2\pi) - \frac{1}{2}\ln|\tau^2\mathbf{I} + \mathbf{V}| - \frac{1}{2}\mathbf{y}'\mathbf{P}\mathbf{y}. \tag{12}$$

The problem therefore simplifies to finding that value of τ^2 that maximizes Equation 12. We will denote this value by $\hat{\tau}^2_{ML}$.

Since there is no closed-form solution for obtaining $\hat{\tau}^2_{ML}$, iterative computations are required. Various procedures can be used for this purpose. We suggest here the use of the Fisher scoring algorithm, which is robust to poor starting values and usually converges quickly (Harville, 1977; Jennrich & Sampson, 1976). For this, we start with an initial estimate of $\hat{\tau}^2$, for example, the value obtained with any of the other (noniterative) estimators described above. This initial estimate is then adjusted based on a factor Δ (the inverse Fisher information of τ^2 multiplied by the first derivative of the profiled log-likelihood with respect to τ^2), yielding a new estimate $\hat{\tau}^2_{New}$. This process continues until convergence and can be expressed by

$$\hat{\tau}^2_{New} = \hat{\tau}^2_{Current} + \Delta, \tag{13}$$

where $\hat{\tau}^2_{Current}$ is the current estimate of τ^2. For ML estimation, the adjustment factor can be shown to be equal to

$$\Delta_{ML} = \frac{\mathbf{y}'\mathbf{P}\mathbf{P}\mathbf{y} - tr(\mathbf{W})}{tr(\mathbf{W}\mathbf{W})}, \tag{14}$$

with $\mathbf{P}$ defined in Equation 4 and the diagonal elements of $\mathbf{W}$ given by $w_i = 1/(v_i + \hat{\tau}^2_{Current})$. Therefore, after each step, we first update $\mathbf{W}$, then $\mathbf{P}$, and finally we can compute Δ_{ML} to obtain $\hat{\tau}^2_{New}$. The iterative process terminates when Δ_{ML} is smaller than some preset threshold (e.g., when $\Delta_{ML} < 10^{-5}$).

An additional complication arises, because Equation 13 may yield a negative value for $\hat{\tau}^2_{New}$. This problem can be easily avoided by using step halving (Jennrich & Sampson, 1976). For this, we check on each iteration whether $\hat{\tau}^2_{Current} + \Delta_{ML} < 0$, and if this is the case, we continue to multiply Δ_{ML} by 1/2 (i.e., first by 1/2, then by 1/4, then by 1/8, and so on) until Δ_{ML} becomes small enough, such that $\hat{\tau}^2_{New}$ stays nonnegative. This ensures that the final value obtained for $\hat{\tau}^2_{ML}$ is also nonnegative.

Restricted Maximum Likelihood (REML) Estimator

ML estimates of variance components tend to be negatively biased (Harville, 1977). To correct for this bias, REML estimation is usually recommended and can be easily adapted for the meta-analytic mixed-effects model (Raudenbush, 2009). In particular, the REML estimator of τ^2 is that value that maximizes the restricted log-likelihood given by

$$ll_{REML} = -\frac{1}{2}k\ \ln(2\pi) + \frac{1}{2}\ln|\mathbf{X}'\mathbf{X}| - \frac{1}{2}\ln|\tau^2\mathbf{I} + \mathbf{V}| - \frac{1}{2}\ln|\mathbf{X}'\mathbf{W}\mathbf{X}| - \frac{1}{2}\mathbf{y}'\mathbf{P}\mathbf{y}. \tag{15}$$

We will denote this value by $\hat{\tau}^2_{REML}$. For REML estimation, the Fisher scoring algorithm works as described above, with the only difference that Δ is now given by

$$\Delta_{REML} = \frac{\mathbf{y}'\mathbf{P}\mathbf{P}\mathbf{y} - tr(\mathbf{P})}{tr(\mathbf{P}\mathbf{P})}. \tag{16}$$

Again, step halving can be used to avoid negative estimates.

[3] As described in footnote 2, $E[\mathbf{y}'\mathbf{P}\mathbf{y}] = \tau^2 tr(\mathbf{P}) + tr(\mathbf{P}\mathbf{V})$. For $\mathbf{W} = \mathbf{V}^{-1}$, the $tr(\mathbf{P}\mathbf{V})$ term simplifies to $tr((\mathbf{W} - \mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W})\mathbf{V}) = tr(\mathbf{I}) - tr(\mathbf{X}'\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}) = k - p - 1$, so that $E[\mathbf{y}'\mathbf{P}\mathbf{y}] = \tau^2 tr(\mathbf{P}) + (k - p - 1)$. Dropping the expectation and rearranging the equation then leads to the DL estimator.

[4] Equation 2 implies that $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \tau^2\tilde{\mathbf{V}})$, where $\tilde{\mathbf{V}} = \mathbf{I} + \frac{1}{\tau^2}\mathbf{V}$, which corresponds exactly to the form of the linear regression model with weight matrix $\tilde{\mathbf{V}}$, where the weights are known up to the proportionality constant τ^2 (e.g., Christensen, 1996). Replacing the unknown value of τ^2 in $\tilde{\mathbf{V}}$ by the crude estimate $\hat{\tau}^2_0$, the usual weighted least squares estimate of the proportionality constant τ^2 is then equal to the SJ estimator.

Empirical Bayes (EB) Estimator

The last estimator we will consider was first proposed by Morris (1983) and was later adapted to the meta-analytic context by Berkey, Hoaglin, Mosteller, and Colditz (1995). This estimator can be derived based on EB methods (Morris, 1983) and will therefore be denoted by $\hat{\tau}^2_{EB}$. Again, there is no closed-form solution, so iterative methods must again be used. It can be shown that $\hat{\tau}^2_{EB}$ can be obtained via the numerical procedure described above, where Δ is now given by

$$\Delta_{EB} = \frac{k/(k-p-1)\mathbf{y}'\mathbf{P}\mathbf{y} - k}{tr(\mathbf{W})}. \quad (17)$$

Again, negative values of $\hat{\tau}^2_{EB}$ can be avoided by means of step halving.[5]

The EB estimator actually shares some noteworthy properties with the HS and SJ estimators. Note that if the initial estimate of τ^2 for the EB estimator is set equal to 0 (i.e., $\hat{\tau}^2_{Current} = 0$) and only a single iteration is carried out, then the equations for $\hat{\tau}^2_{EB}$ and $\hat{\tau}^2_{HS}$ would only differ by the scale factor $k/(k-p-1)$. Moreover, if one continued to iterate the SJ estimator (i.e., by setting $\hat{\tau}^2_0$ equal to $\hat{\tau}^2_{SJ}$ and then reapplying Equation 10 until convergence), one would in fact obtain $\hat{\tau}^2_{EB}$. Therefore, both the HS and SJ estimators can be seen as special cases of the EB estimator based on a single iteration.

Furthermore, the EB estimator can be shown to be identical to another estimator, going back to the work of Paule and Mandel (1982), that was recently described in the meta-analytic context by DerSimonian and Kacker (2007). In particular, for the mixed-effects model, the Paule–Mandel (PM) estimator is that value of τ^2 for which

$$\mathbf{y}'\mathbf{P}\mathbf{y} = k - p - 1, \quad (18)$$

with $\mathbf{P}$ again defined in Equation 4 and diagonal elements of $\mathbf{W}$ given by $w_i = 1/(v_i + \tau^2)$. We will denote this value by $\hat{\tau}^2_{PM}$. Since $\mathbf{y}'\mathbf{P}\mathbf{y}$ is a strictly decreasing function of τ^2, $\hat{\tau}^2_{PM}$ is set to 0 if $\mathbf{y}'\mathbf{P}\mathbf{y} < k - p - 1$ for $\tau^2 = 0$. The equivalence of the EB and PM estimators leads to some interesting properties to be described further below.[6]

Hypothesis Tests for the Model Coefficients

Once an estimate of τ^2 has been computed, the vector of model coefficients can be obtained with Equation 3. The next step in a meta-regression analysis is to determine the precision of these estimates and to test whether the relationship between moderators and effect sizes is statistically significant. Six alternative methods for testing the regression coefficients are presented below.

The first one is a Wald-type test (Raudenbush, 2009), and it is the one that is most commonly applied in practice. Accordingly, we will refer to this approach as the standard method. However, concerns have been raised by findings that this test does not adequately control the Type I error rate (Knapp & Hartung, 2003; Sidik & Jonkman, 2005). Recently, Knapp and Hartung (2003) proposed an improved method that appears to rectify some of the problems with the standard approach, especially when the number of studies is low. However, two implementations of this method can be constructed, corresponding to the second and third alternative methods we will consider. The fourth method makes use of a robust estimate of the variance–covariance matrix of the model coefficients. Another alternative considered here is the likelihood ratio test. Finally, a permutation test is described. While the latter is computationally more demanding than the other tests, it is, in principle, free of distributional assumptions.

Standard (Wald-Type) Method

If we could estimate τ^2 without error, then the variance–covariance matrix of the model coefficients computed with Equation 3 is equal to $\mathbf{\Sigma} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$, with diagonal elements of $\mathbf{W}$ given by $w_i = 1/(v_i + \tau^2)$. However, since τ^2 is unknown in practice, we cannot compute $\mathbf{\Sigma}$ directly. The standard approach is to substitute the estimate of τ^2 for the unknown variance component in $\mathbf{W}$, yielding an estimate of $\mathbf{\Sigma}$ given by the equation

$$\hat{\mathbf{\Sigma}} = (\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}, \quad (19)$$

where the diagonal elements of $\hat{\mathbf{W}}$ are equal to $w_i = 1/(v_i + \hat{\tau}^2)$. The test statistic for a particular model coefficient can then be obtained with

$$z_j = \frac{b_j}{\sqrt{Var_{\hat{\mathbf{\Sigma}}}[b_j]}}, \quad (20)$$

with b_j denoting a particular element of the $\mathbf{b}$ vector and $\sqrt{Var_{\hat{\mathbf{\Sigma}}}[b_j]}$ the square root of the corresponding diagonal element of the $\hat{\mathbf{\Sigma}}$ matrix (i.e., the estimated standard error of b_j). The value obtained by Equation 20 is then compared against the critical values of a standard normal distribution for a desired significance level (e.g., ± 1.96 for $\alpha = .05$, two-sided). Despite its widespread use, this method ignores the imprecision in the estimate of τ^2 when estimating $\mathbf{\Sigma}$. Thus, if τ^2 is estimated poorly, the actual Type I error rate of this method may deviate from the nominal significance level, leading to either an overly conservative or, usually, a too liberal rejection rate (Huizenga et al., 2011; Knapp & Hartung, 2003; Sidik & Jonkman, 2005).

Knapp and Hartung Method

The Knapp and Hartung method (2003) is based on an adjusted estimate of the variance–covariance matrix of the model coefficients that is expected to improve the Type I error rate compared to the standard method described above. The adjusted variance–covariance matrix is given by

$$\hat{\mathbf{\Sigma}}_{KH} = s^2(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}, \quad (21)$$

[5] The equation for the EM estimator given by Berkey et al. (1995, p. 398) can be written as $\hat{\tau}^2 = \hat{\tau}^2 + \frac{k/(k-p-1)\mathbf{y}'\mathbf{P}\mathbf{y} - tr(\mathbf{W}\mathbf{V})}{tr(\mathbf{W})} - \hat{\tau}^2$. We can reformulate the part after the plus sign into $\frac{k/(k-p-1)\mathbf{y}'\mathbf{P}\mathbf{y} - tr(\mathbf{W}\mathbf{V} + \hat{\tau}^2\mathbf{W})}{tr(\mathbf{W})}$. Finally, after noting that $tr(\mathbf{W}\mathbf{V} + \hat{\tau}^2\mathbf{W}) = tr(\mathbf{I}) = k$, we obtain Equation 17.

[6] The equivalence between the EB and PM estimators is apparent after noting that, upon convergence of the iterative algorithm, $\Delta_{EB} = 0$, which implies $k/(k-p-1)\mathbf{y}'\mathbf{P}\mathbf{y} - k = 0$ (cf. Equation 17). The latter is equivalent to $\mathbf{y}'\mathbf{P}\mathbf{y} = k - p - 1$, which is the same as Equation 18.

where

$$s^2 = \frac{\mathbf{y}'\mathbf{P}\mathbf{y}}{k - p - 1}, \tag{22}$$

with $\mathbf{P}$ again defined in Equation 4 and diagonal elements of $\hat{\mathbf{W}}$ given by $w_i = 1/(v_i + \hat{\tau}^2)$. The test statistic for a particular model coefficient is then computed with

$$t_j = \frac{b_j}{\sqrt{Var_{\hat{\boldsymbol{\Sigma}}_{KH}}[b_j]}}, \tag{23}$$

with b_j denoting the respective element of $\mathbf{b}$ and $\sqrt{Var_{\hat{\boldsymbol{\Sigma}}_{KH}}[b_j]}$ the square root of the corresponding diagonal element of $\hat{\boldsymbol{\Sigma}}_{KH}$. The value obtained by Equation 23 is then compared against the critical values of a t-distribution with $df = k - p - 1$ degrees of freedom.

The principle underlying the Knapp and Hartung method is as follows. If the exact value of τ^2 were known and used to compute $\mathbf{P}$ with the diagonal elements of $\mathbf{W}$ given by $w_i = 1/(v_i + \tau^2)$, then $\mathbf{y}'\mathbf{P}\mathbf{y}$ would follow a chi-square distribution with $df = k - p - 1$ degrees of freedom. This follows directly from the properties of quadratic forms (e.g., Christensen, 1996) when $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \tau^2\mathbf{I} + \mathbf{V})$. The s^2 statistic then scales $\mathbf{y}'\mathbf{P}\mathbf{y}$ by its degrees of freedom, so that Equation 23 takes on the form of a t-distributed random variable under the null hypothesis $H_0 : \beta_j = 0$; that is, t_j is the ratio of a random variable following a standard normal distribution and the square root of a random variable following a chi-square distribution scaled by its degrees of freedom (e.g., Hogg & Craig, 1995). When $\mathbf{P}$ is computed using $\hat{\mathbf{W}}$ with elements $w_i = 1/(v_i + \hat{\tau}^2)$, this derivation is only approximate, but the use of the t-distribution still helps to counteract the typically too liberal Type I error rate of the standard method.

It is worth noting that when using the EB estimator of τ^2, the adjustment factor s^2 is always automatically equal to 1 for positive values of $\hat{\tau}^2_{EB}$ (Knapp & Hartung, 2003). This result follows immediately from the equivalence of the EB and PM estimators described earlier (see Equation 18). Therefore, when using the EB estimator, the adjustment factor s^2 is essentially already incorporated into the estimate of the variance–covariance matrix of the model coefficients.

Knapp and Hartung Method With Truncation

Knapp and Hartung (2003) originally proposed that the adjustment factor s^2 should always be equal to or greater than 1. A value smaller than 1 is likely to be obtained with Equation 22 in scenarios where the effect sizes are very homogeneous, so that the total variability unaccounted for by the moderators, $Q_E = \mathbf{y}'\mathbf{P}\mathbf{y}$, is even smaller than its expected value (i.e., $df = k - p - 1$ when $\tau^2 = 0$). However, when working with small samples (i.e., small number of studies, small average number of participants per study, or both), such counterintuitive results can easily happen, since meta-analytic estimates are then generally also quite inaccurate (Hedges, 2009).

Following the recommendations provided by Knapp and Hartung (2003), the adjustment factor s^2 should be truncated to 1 when a smaller value is obtained. With this practice, the variance estimate of b_j obtained with their method would never be smaller than the one obtained with the standard method, always leading to more conservative tests than those obtained with the standard approach. However, this practice may actually be overly conservative, leading to a loss of power, thereby increasing the chance that relevant moderators may be missed. This will be examined in more detail further below.

Robust (Huber–White) Method

The robust method is based on the work of Huber (1967) and White (1980) and was first proposed in the meta-analytic literature by Sidik and Jonkman (2005). In general, the purpose of robust methods is to account for potential model misspecification. Failing to account for dependencies in the effect size estimates (e.g., due to clustering) would be one form of model misspecification (as addressed by Hedges et al., 2010). Other issues include heteroscedastic and/or autocorrelated residuals, which can be handled by using heteroscedasticity-consistent and/or heteroscedasticity-and-autocorrelation-consistent estimators, such as the Huber–White or the Andrews estimator (Andrews, 1991; Huber, 1967; White, 1980). Following Sidik and Jonkman (2005), we regard the issue here as a problem of incorrectly specifying the exact marginal variances of the effect size estimates (i.e., the $v_i + \tau^2$ values) due to the substitution of corresponding estimates. Given that these marginal variances are heteroscedastic, the Huber–White estimator could be used to obtain a consistent estimate of the variance–covariance matrix of the model coefficients.

In particular, for this method, the variance–covariance matrix of the model coefficients is estimated with

$$\hat{\boldsymbol{\Sigma}}_{HW} = (\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{W}}\hat{\mathbf{E}}^2\hat{\mathbf{W}}\mathbf{X}(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}, \tag{24}$$

where $\hat{\mathbf{E}}$ is a diagonal matrix with elements obtained from the vector $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\mathbf{b}$. The test statistic for a particular model coefficient is then given by Equation 23, except that $\sqrt{Var_{\hat{\boldsymbol{\Sigma}}_{KH}}[b_j]}$ is replaced with $\sqrt{Var_{\hat{\boldsymbol{\Sigma}}_{HW}}[b_j]}$. Again, the test statistic is compared against the critical values of a t-distribution with $df = k - p - 1$ degrees of freedom.

In their simulation study, Sidik and Jonkman (2005) found that the robust method does not consistently improve the performance of the standard method regarding the control of the Type I error rate. Hedges et al. (2010) recently proposed that a simple correction to $\sqrt{Var_{\hat{\boldsymbol{\Sigma}}_{HW}}[b_j]}$ should yield closer to acceptable performance levels regarding α. The test statistic is then given by

$$t_j = \frac{b_j}{\sqrt{k/(k - p - 1)Var_{\hat{\boldsymbol{\Sigma}}_{HW}}[b_j]}}, \tag{25}$$

which yields a more conservative test, especially when k is small. However, it remains to be determined how the robust method with this correction performs in comparison to the other approaches considered in the present article.

Likelihood Ratio Test

All of the approaches described so far are based on a test statistic that divides the model coefficient to be tested by some estimate of its standard error. An alternative approach is based on likelihood ratio testing (Huizenga et al., 2011), which can be used in the context of ML estimation. Let $ll_{ML}(\hat{\tau}^2_{ML})$ denote the value of

the profiled log-likelihood, defined by Equation 12, based on the ML estimate of τ^2. Next, let $ll_{ML}(\hat{\tau}^2_{ML,\beta_j=0})$ denote the value of the log-likelihood under the null hypothesis $H_0: \beta_j = 0$. Obtaining this value requires that we reestimate τ^2 after removing the column corresponding to β_j from the $\mathbf{X}$ matrix. The likelihood ratio test statistic is then obtained with

$$LRT = -2(ll_{ML}(\hat{\tau}^2_{ML,\beta_j=0}) - ll_{ML}(\hat{\tau}^2_{ML})), \tag{26}$$

which is compared against the critical value of a chi-square distribution with 1 degree of freedom (i.e., 3.84 for $\alpha = .05$). On the basis of Huizenga et al. (2011), we expect the likelihood ratio test to provide slightly better control of the Type I error rate compared to the standard method.

Permutation Test

Finally, the use of permutation tests has been suggested as another alternative in the meta-analytic context (Follmann & Proschan, 1999; Higgins & Thompson, 2004). To carry out the test for a particular model coefficient, we first obtain z_j, the test statistic based on the standard approach, given by Equation 20. Then, for each of the $k!$ possible permutations of the rows of the $\mathbf{X}$ matrix, the model is refitted and the value of the test statistic is recomputed (note that each permutation requires that τ^2, $\boldsymbol{\beta}$, and $\boldsymbol{\Sigma}$ are reestimated). Let z^m_j denote the value of the test statistic for the mth permutation. By permuting the rows of the $\mathbf{X}$ matrix, any relationship between the effect sizes and the moderator values is now purely a result of chance, so that the z^m_j values reflect the sampling distribution of the test statistic under the null hypothesis. Therefore, the (two-sided) p value for the permutation test is equal to 2 times the proportion of cases where the test statistic under the permuted data is as extreme or more extreme than under the actually observed data (i.e., $2 \times \sum_{m=1}^{k!} I(z^m_j \geq z_j)/k!$ when z_j is positive and $2 \times \sum_{m=1}^{k!} I(z^m_j \leq z_j)/k!$ when z_j is negative, where $I()$ is the indicator function that is equal to 1 if the condition in the parentheses is true and 0 otherwise).

Note that k must be at least as large as 5 before it is actually possible to obtain a p value below $\alpha = .05$ (i.e., for $4! = 24$ permutations, the p value can never be smaller than $2 \times 1/24 = .0833$, while for $5! = 120$, the p value can be as small as .0167). On the other hand, as k increases, $k!$ quickly grows so large that it may not be possible in practice to obtain the full set of permuted test statistics. In that case, one can approximate the exact permutation-based p value by going through a certain number of random permutations of the rows of the $\mathbf{X}$ matrix. Using a sufficiently large number of such random permutations ensures that the resulting p value is stable.

The permutation approach may be especially appropriate when the data cannot be regarded as a random sample from a given population (Manly, 1997). In the context of a meta-analysis, the sample refers to the set of included studies, which are assumed to be a random selection from a larger (hypothetical) population of studies (Hedges & Vevea, 1998). This conceptualization is often questionable in practice, which makes permutation tests especially appealing for meta-analyses. Moreover, this method is, in principle, free of distributional assumptions. However, the use of a nonparametric approach may be less efficient than parametric methods, potentially resulting in lower power. This is examined in more detail below.

Simulation Study

In summary, we have described seven estimators of τ^2 (i.e., the HE, HS, DL, SJ, ML, REML, and EB estimators) and six methods for conducting hypothesis tests for the model coefficients in the context of mixed-effects meta-regression models (i.e., the standard method, the Knapp and Hartung method once without and once with truncation, the robust [Huber–White] method, the likelihood ratio test, and the permutation test). The likelihood ratio test is only applicable when using ML estimation. Moreover, the permutation test is computationally very demanding when it is combined with an iterative estimator of τ^2 (i.e., the ML, REML, and EB estimators). We will therefore only consider the permutation test when using one of the four noniterative estimators of τ^2. Therefore, combining the various τ^2 estimators with the various testing methods yields in principle 33 ways of testing the statistical significance of model coefficients in mixed-effects meta-regression models. To compare the performance of these methods, we conducted a Monte Carlo simulation study using standardized mean differences as the effect size measure.

In particular, assume that each study included in a meta-analysis compared subjects in an experimental (E) group with those in a control (C) group with respect to some quantitative outcome. Assuming that the scores of the subjects in the respective groups are normally distributed with true means μ^E_i and μ^C_i and common standard deviation σ_i, then

$$\theta_i = \frac{\mu^E_i - \mu^C_i}{\sigma_i} \tag{27}$$

denotes the true standardized mean difference in the ith study and an unbiased estimate of θ_i can be obtained with

$$y_i = \left(1 - \frac{3}{4(n^E_i + n^C_i) - 9}\right) d_i, \tag{28}$$

where $d_i = (\bar{x}^E_i - \bar{x}^C_i)/s_i$, $\bar{x}^E_i$ and $\bar{x}^C_i$ denote the observed means of the n^E_i and n^C_i subjects in the respective groups, and s_i the observed (pooled) standard deviation (Hedges & Olkin, 1985). The sampling variance of y_i can then be estimated with

$$v_i = \frac{1}{n^E_i} + \frac{1}{n^C_i} + \frac{y^2_i}{2(n^E_i + n^C_i)}. \tag{29}$$

For the simulation study, we assumed that a single moderator influences the size of the true effects, such that

$$\theta_i = \beta_0 + \beta_1 x_i + u_i. \tag{30}$$

For each iteration of the simulation, the values of the moderator were randomly generated based on a standard normal distribution and the u_i values from $N(0, \tau^2)$. We considered three values for τ^2, namely 0, 0.08, and 0.32, corresponding to the absence, a medium amount, and a large amount of residual heterogeneity in the true effects. Without loss of generality, we set β_0 equal to 0. For β_1, three conditions were examined, namely $\beta_1 = 0$, $\beta_1 = 0.2$, and $\beta_1 = 0.5$, the first yielding information on the Type I error rate of the various tests, the latter providing information about the power of the tests when the null hypothesis is in fact false. Note that for

each combination of the three τ^2 values and the three β_1 values, the model predictive power could be defined as $P^2 = \beta_1^2/(\beta_1^2 + \tau^2)$ (see Borenstein et al., 2009; López-López et al., 2014; Raudenbush, 1994).[7]

The conditions manipulated in the present study were intended to represent scenarios commonly found by meta-analysts and similar to those used in previous simulation studies (e.g., Huizenga et al., 2011; Knapp & Hartung, 2003; Sidik & Jonkman, 2005). For k, we considered five values, namely 5, 10, 20, 40, and 80, corresponding to a small to large number of studies for the meta-analysis. After simulating k values of θ_i based on Equation 30, we then generated the corresponding observed effect size estimates with $d_i = Z_i/\sqrt{X_i/m_i}$, where $Z_i \sim N(\theta_i, 1/n_i^E + 1/n_i^C)$, $X_i \sim \chi^2_{m_i}$, and $m_i = n_i^E + n_i^C - 2$. Unbiased estimates of θ_i were then obtained by applying Equation 28. The corresponding sampling variances were then computed with Equation 29.

We also manipulated the sample sizes of the individual studies, assuming $n_i = n_i^E = n_i^C$ and setting n_i equal to either (6, 8, 9, 10, 42), (16, 18, 19, 20, 52), or (41, 43, 44, 45, 77), corresponding to average sample sizes of 30, 50, and 100 subjects for the studies (these sample size distributions were obtained based on a review of published meta-analyses; for more details, see Sánchez-Meca & Marín-Martínez, 1998). For the $k = 10$, $k = 20$, $k = 40$, and $k = 80$ conditions, the sample size vectors were repeated 2, 4, 8, and 16 times, respectively.

Thus, a total of $5(k) \times 3(n_i) \times 3(\beta_1) \times 3(\tau^2) = 135$ conditions were examined. For each of these conditions, 10,000 meta-analyses were simulated. After generating the data within a particular iteration of a particular condition, we fitted the meta-regression model using the various residual heterogeneity estimators and then tested the model coefficient β_1 for statistical significance with the various procedures described earlier, using $\alpha = .05$ as the nominal significance level. For $k = 5$, an exact permutation test was carried out. For larger values of k, obtaining the exact permutation-based p values was not feasible. Therefore, we then used 5,000 random permutations for the test. The rejection rates of the various procedures were recorded for each condition. The simulation was conducted with R (R Core Team, 2013), using the metafor package to fit the meta-regression models (Viechtbauer, 2010).

Results

In this section, we describe and compare the performance of the different methods under the simulated conditions. In general, only the results for the standard method appeared to be influenced to some extent by the residual heterogeneity estimator used. However, even for the standard method, the overall trends were similar regardless of how τ^2 was estimated. Therefore, for brevity, we only present the results for the DL and ML estimators (the full set of results are provided as part of the supplemental materials). We highlight these findings, since the DL estimator is the most commonly used estimator in practice, while the ML estimator allows us to examine the performance of the likelihood ratio test in comparison to the other methods. This section is divided into two parts, corresponding to the Type I error rate and the statistical power of the tests, respectively.

Type I Error

Setting $\beta_1 = 0$ allowed us to compare the methods in terms of their Type I error rates. Note that by setting $\alpha = .05$, values around .05 for the empirical Type I error rate indicate that the Type I error rate is adequately controlled. Figure 1 shows our findings for the different methods when using the DL estimator. Since values for the Knapp and Hartung method and the permutation test were essentially indistinguishable, results for both tests were averaged. Also, no results for the likelihood ratio test are given here, since it is only applicable when using ML estimation. Finally, since the average within-study sample size had relatively little influence on the Type I error rate of the different methods, we averaged the rates over this factor.

Results were very different depending on the method used to test the moderator. The rejection rates of the standard method generally fell above the nominal significance level, except when $\tau^2 = 0$, in which case the Type I error rates were slightly conservative. As the number of studies increased, the rejection rates converged to the nominal significance level, although convergence appears to be slow when $\tau^2 = 0$.

On the other hand, both the Knapp and Hartung method and the permutation test performed very close to the nominal significance level regardless of the simulated scenario. In contrast, the truncated Knapp and Hartung method provided overly conservative results, especially when the number of studies was small and when there was no residual heterogeneity among the true effects. Finally, the Huber–White method showed empirical rejection rates above the nominal significance level. Interestingly, the method does not appear to be sensitive to the amount of residual heterogeneity. Again, the Type I error rates converged to the nominal significance level as the number of studies increased.

Figure 2 presents the results for the different statistical tests when using the ML estimator. Performance of the Huber–White method was similar to that when using the DL estimator and therefore was not included in this figure. Also, the permutation test is not included in these results because, as stated before, this method is computationally overly demanding when combined with an iterative estimator of τ^2.

The general trend in the performance of the methods was similar when using the DL and the ML estimator. The standard method showed rejection rates clearly above the nominal significance level, especially with a small number of studies and a large amount of residual heterogeneity among the true effects, while the Knapp and Hartung method adequately controlled the Type I error rate irrespective of the simulated scenario. On the other hand, the rejection rate of the truncated Knapp and Hartung method again fell below the nominal significance level, getting closer to .05 as the number of studies and the amount of residual heterogeneity

[7] With $\theta_i = \beta_1 x_i + u_i$, the total amount of heterogeneity in the true effect sizes is equal to $Var(\theta_i) = \beta_1^2 Var(x_i) + Var(u_i) = \beta_1^2 + \tau^2$, as x_i and u_i are independent and normally distributed with mean 0 and variances 1 and τ^2, respectively. The denominator in the formula proposed by Raudenbush (1994) for the model predictive power is the total heterogeneity, $\beta_1^2 + \tau^2$, while the numerator in this equation represents the part of the heterogeneity explained by the predictor, namely β_1^2.

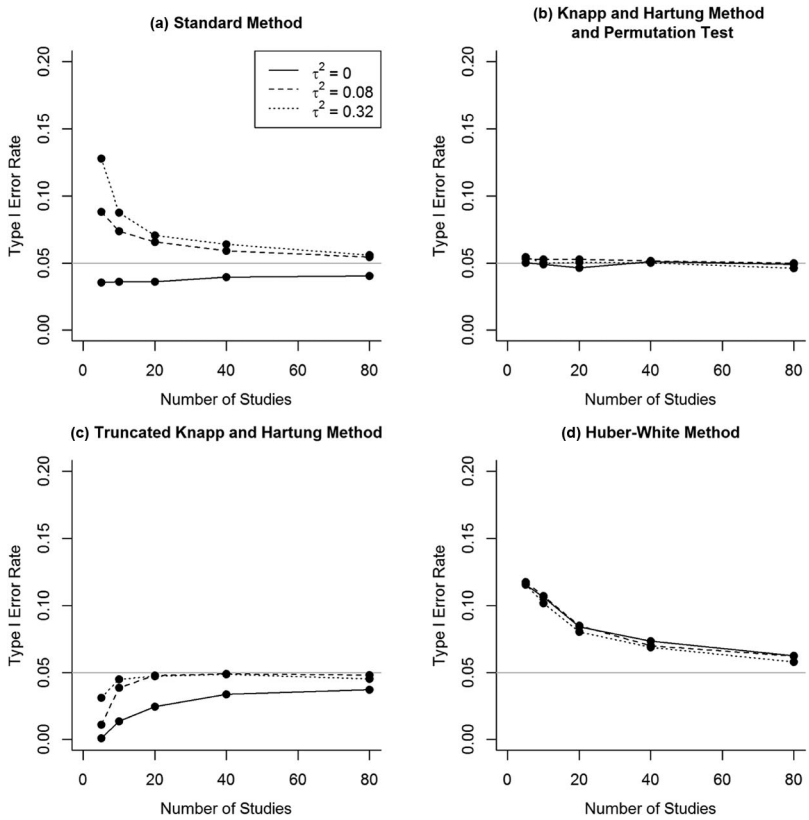

Figure 1. Empirical Type I error rates of the methods when using the DerSimonian and Laird estimator (likelihood ratio test not applicable here).

increased. Finally, results for the likelihood ratio test were similar to those of the standard method, but slightly closer to the nominal significance level when $\tau^2 > 0$.

Statistical Power

Statistical power reflects the probability of a method rejecting the null hypothesis when it is in fact false (i.e., $\beta_1 \neq 0$ in our simulation study). Generally, power rates equal to or greater than 0.8 are considered satisfactory in the psychological science (Cohen, 1988). In order to assess the statistical power of the different procedures for testing the significance of regression coefficients, conditions with $\beta_1 = 0.2$ are considered here.

Figure 3 presents our findings for the various methods when using the DL estimator. Again, the Knapp and Hartung method and the permutation test showed very similar results, so that values for both methods were averaged and are presented jointly. The likelihood ratio test is again not applicable here. Also, while the power of the various methods increased as the average within-study sample size increased, in general this factor had only a relatively minor influence on the power rates. We therefore again averaged the rates over this factor.

Although differences were not very pronounced, the standard and Huber–White methods systematically showed the highest rejection rates, with the truncated Knapp and Hartung method providing the lowest rates. Note, however, that differences in the Type I error rates of the various methods obfuscate such direct comparisons between the power rates (i.e., the lower or higher power of a method may in fact just be an artifact of an overly conservative or inflated Type I error rate to begin with).

The influence of the different conditions manipulated in the simulation was similar for all of the methods. As expected, the number of studies showed a strong positive relationship with the power of the tests. However, at least 40 studies were required for the different methods to provide power rates close to the desired value of 0.8, as long as the amount of residual heterogeneity was not large. In the presence of substantial amounts of residual heterogeneity, up to 80 studies would be needed to achieve power rates close to .80. Therefore, as expected, the amount of residual heterogeneity showed a negative relationship with power, with larger residual τ^2 values corresponding to smaller rejection rates.

Finally, power rates for the methods when using the ML estimator are presented in Figure 4. Again, results for the Huber–White method were not included, since the trends for this method were similar to the ones already described in combination with the DL estimator.

Figure 4 shows that the highest power rates were obtained with the standard and likelihood ratio tests, while the truncated Knapp and Hartung method yielded again the lowest rejection rates. Similar to the DL estimator, all methods showed higher power rates as the number of studies increased. Also, power for all methods decreased as the amount of residual heterogeneity among

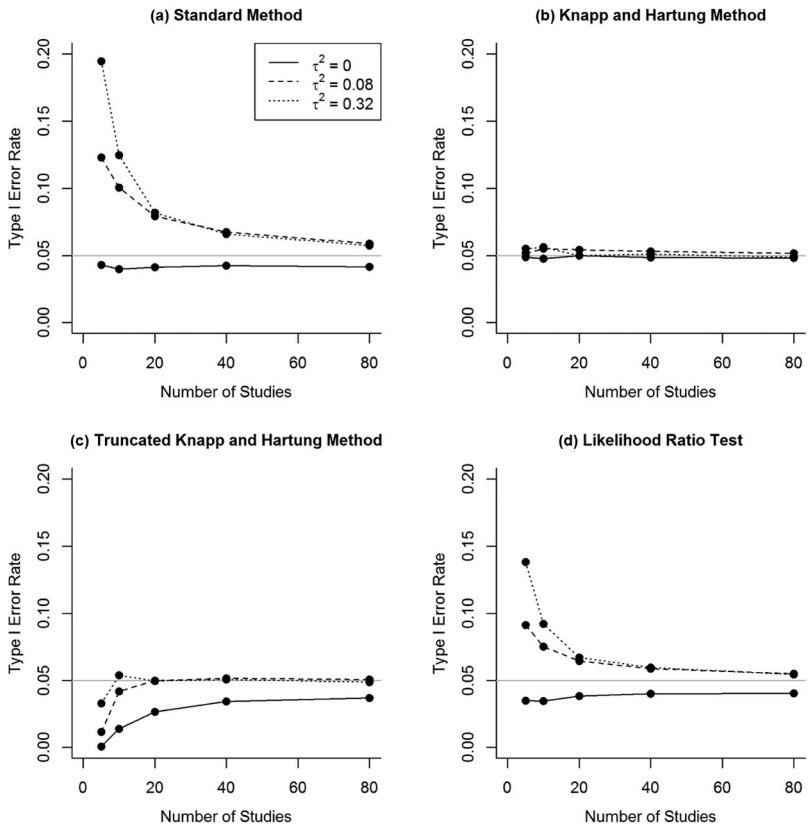

Figure 2. Empirical Type I error rates of the methods when using the maximum likelihood estimator (results for Huber–White method omitted).

the true effects increased, with the rejection rate of the truncated Knapp and Hartung method gradually converging to that of the untruncated version of the test.

Results for $\beta_1 = 0.5$ are not presented here. With such a large slope value, all methods provided rejection rates close to or over .80 with 20 or more studies. With smaller values of k, trends for the different methods were very similar to the ones described above for $\beta_1 = 0.2$.

Illustrative Example

We now consider an example to illustrate the various methods. For this purpose, we use data from a meta-analysis on the effectiveness of school-based writing-to-learn interventions on academic achievement (Bangert-Drowns, Hurley, & Wilkinson, 2004). In each of the studies included in this meta-analysis, an experimental group (i.e., a group of students that received instruction with increased emphasis on writing tasks) was compared against a control group (i.e., a group of students that received conventional instruction) with respect to some content-related measure of academic achievement (e.g., final grade, an exam/quiz/test score). As in the simulation study above, the effect size measure used for this meta-analysis was the standardized mean difference.

The y_i and corresponding v_i values for 46 studies included in this meta-analysis are given in Table 1. Positive values for y_i indicate that the students receiving the intervention performed, on average, better than those in the control group condition. However, there is quite a bit of variability in the observed effects, which may be related to differences in how the studies were conducted. The treatment length (in weeks) is also reported for each study, which may be a potential moderator of the treatment effectiveness. We will now examine this hypothesis in more detail. The analyses described below were conducted with R, using the metafor package. The corresponding code to replicate these analyses is provided in the supplemental materials.

First, we fitted the meta-regression model $y_i = \beta_0 + \beta_1 length_{i1} + u_i + e_i$ to these data, in turn using each of the seven residual heterogeneity estimators described previously, and then applied the standard (Wald-type) test of $H_0 : \beta_1 = 0$. The results are given in Table 2. The estimated values of τ^2 ranged from $\hat{\tau}^2 = 0.0373$ for the HS estimator to $\hat{\tau}^2 = 0.0832$ for the SJ estimator. The differences in $\hat{\tau}^2$ result in different $\hat{\mathbf{W}}$ matrices that lead to slightly different estimates of the model coefficient β_1 and more pronounced differences in the corresponding estimated standard errors of b_1. As a result, the null hypothesis is rejected at $\alpha = .05$ (two-sided) when using the HS, DL, ML, REML, and EB estimators, but not when using the HE and SJ estimators. Therefore, as this example demonstrates, the choice of τ^2 estimator may have an impact on the conclusions.

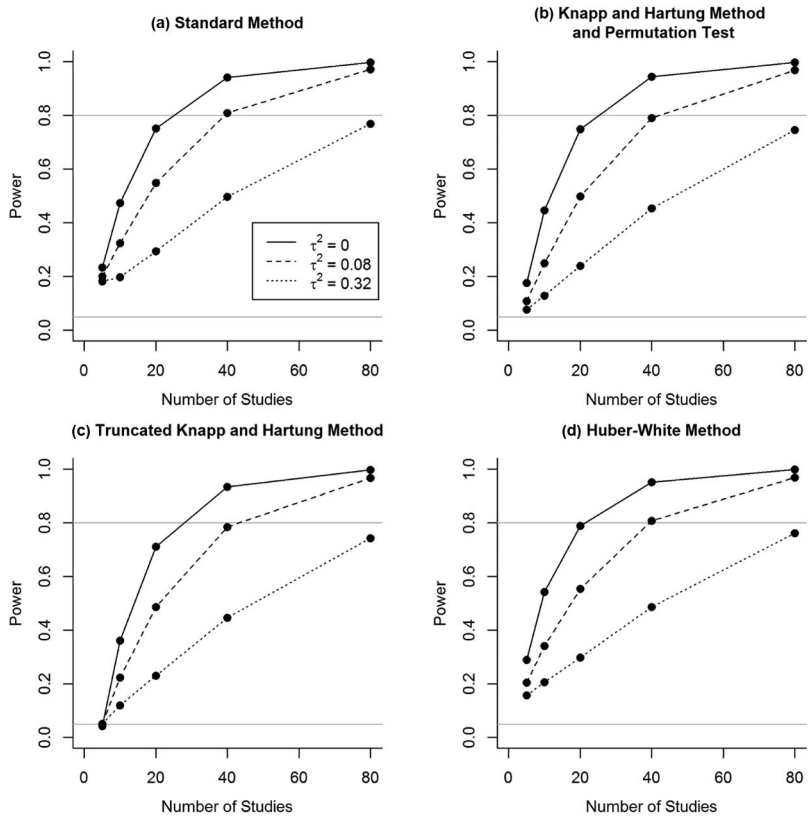

Figure 3. Statistical power rates of the methods when using the DerSimonian and Laird estimator (likelihood ratio test not applicable here).

Instead of using the standard testing method, one could combine each of the seven estimators of τ^2 with the Knapp and Hartung method (without or with truncation), the robust (Huber–White) method, and the permutation test. When using ML estimation, one final possibility is to use the likelihood ratio test. Instead of illustrating all possibilities, we provide the results of the different testing procedures when combined with ML estimation. As shown in Table 3, the Knapp and Hartung method leads to a larger standard error of b_1 when compared with the standard method (hence the adjustment factor s^2 must have been larger than 1 and no distinction can be made between the truncated and untruncated versions of the method). The resulting p value (which is now computed based on a t-distribution with 44 degrees of freedom) is not significant. Similarly, the permutation test (based on 100,000 random permutations) yields a nonsignificant p value. In comparison, the remaining methods lead to the rejection of H_0. Again, the example demonstrates how the choice of method can lead to conflicting conclusions.

Discussion

Several different methods are available for analyzing the association between one or more covariates and the effect sizes. In this article, we compared a variety of different methods in the context of mixed-effects meta-regression models. Specifically, seven residual heterogeneity variance estimators and six methods for testing the statistical significance of the regression coefficients were compared in a Monte Carlo simulation study with standardized mean differences as the effect size measure.

Two comparative criteria were considered for assessing the adequacy of each method across conditions similar to those typically found in psychological research. On the one hand, empirical Type I error rates were examined in order to assess which methods adequately control the rejection rate when a covariate is unrelated to the size of the effects. On the other hand, statistical power rates were obtained, to check which methods are more likely to detect a real moderator variable. Except for the standard method, the results were not found to be affected by the residual heterogeneity estimator used. However, some notable differences were observed depending on the method employed for testing the regression coefficients.

Some authors have criticized that the standard method does not take into account the uncertainty due to the variance estimation process, which in turn increases the risk of reaching statistically significant results that might be inappropriate (e.g., Thompson & Higgins, 2002). When examining the empirical Type I error rates from our simulation study, results for the standard method were in fact not satisfactory, with rates clearly above the nominal significance level in most situations, especially when some residual heterogeneity was present in the true effects and the number of studies was low. The higher statistical power of the standard method (in comparison with the Knapp and Hartung method and the permutation test) is therefore an artifact of the method rejecting

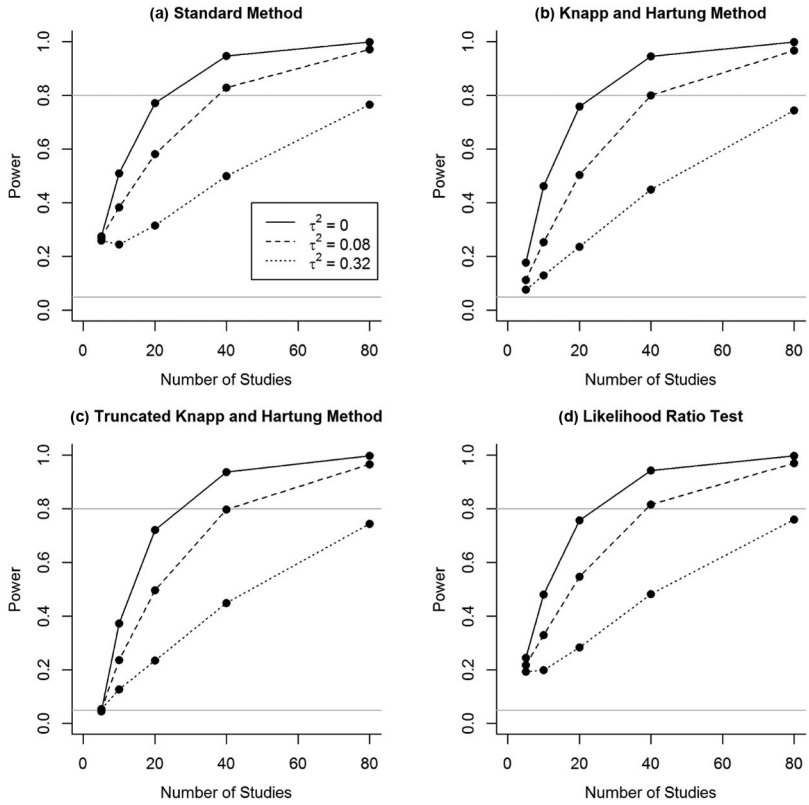

Figure 4. Statistical power rates of the methods when using the maximum likelihood estimator (results for Huber–White method omitted).

the null hypothesis more often than the desired α value when the null hypothesis is in fact true. However, a test with deficient control of the Type I error rate should be avoided for routine use. Therefore, these results lead us to encourage meta-analysts to consider alternative methods, particularly when the number of studies in a research synthesis is small.

Due to the problems related to the standard method, some authors have suggested various alternatives for testing the regression coefficients. Among those is the Knapp and Hartung method, which incorporates an adjustment factor into the standard formula for estimating the variance–covariance matrix of the regression coefficients and whose statistical test is based on the *t* instead of the normal distribution. When this test was first proposed (Knapp & Hartung, 2003), the authors suggested truncating the adjustment factor to 1 if a smaller value was obtained. With this practice, the variance estimates of the regression coefficients would always be equal to or greater than the ones obtained with the standard method, so that the test would always yield a more conservative outcome (note that even when the adjustment factor is 1, due to the use of the *t*-distribution, the resulting *p* value will still be more conservative than the one obtained by means of the standard method).

However, the untruncated Knapp and Hartung method provided adequate control of the Type I error rate, while truncating this method led to overly conservative results, as seen in Figures 1 and 2. Moreover, when comparing the methods in terms of their power in this simulation study, Figures 3 and 4 show that the truncated Knapp and Hartung method provided systematically lower rejection rates than all of the remaining methods under assessment. Therefore, results of the present study suggest better performance of the Knapp and Hartung method without the truncation of its adjustment factor. This is of particular concern, given that some software macros for meta-analysis (e.g., those that can be found in Stata) have implemented the Knapp and Hartung method only in combination with the truncation.

The trends described in the last paragraph for both versions of the Knapp and Hartung method, illustrated in Figures 1–4 for the DL and ML estimators, were observed as well when combining these methods with the EB estimator, despite the fact that the adjustment factor s^2 is then always equal to 1 for positive values of $\hat{\tau}^2_{EB}$, as pointed out before. Our results therefore indicate that the truncation proposed by Knapp and Hartung (2003) will make a difference especially in situations where the residual heterogeneity estimate is likely to require truncation as well (Borenstein et al., 2009), that is, when the amount of residual heterogeneity is (or is close to) 0.

The performance of the Huber–White and likelihood ratio tests was also assessed in the present study. As found in previous Monte Carlo simulations (Huizenga et al., 2011; Sidik & Jonkman, 2005), our results showed empirical Type I error rates above the nominal significance level for both tests (except in the absence of residual heterogeneity, in which case the likelihood ratio test was slightly

conservative). Therefore, these methods cannot be recommended for routine use, at least in their present form.

Finally, the performance of a permutation test was also analyzed. This method provided results very similar to those of the (untruncated) Knapp and Hartung method. The Knapp and Hartung method is, however, simpler to compute than the permutation test (the latter requiring intensive computations), so that it seems a reasonable choice for most situations. Note, however, that the true effects were generated in our simulation study as if one selects a random sample of studies from a superpopulation of studies (with normally distributed true effects). This corresponds to the usual conceptualization of the random/mixed-effects model in meta-analysis (Hedges & Vevea, 1998) and therefore also underlies the Knapp and Hartung method for testing the regression coefficients. In that sense, the Knapp and Hartung method is a suitable option as long as the set of studies can be reasonably assumed to be a random sample from a broader population of studies. On the other hand, if no random sampling of studies can be assumed, then the permutation test constitutes a more appropriate method (Manly, 1997).

Table 1
Results From 46 Studies on the Effectiveness of Writing-to-Learn Interventions

Study	N_i	y_i	v_i	Length
1	60	0.65	0.070	15
2	34	−0.75	0.126	10
3	95	−0.21	0.042	2
4	209	−0.04	0.019	9
5	182	0.23	0.022	14
6	462	0.03	0.009	1
7	38	0.26	0.106	4
8	542	0.06	0.007	15
9	99	0.06	0.040	4
10	77	0.12	0.052	9
11	40	0.77	0.107	15
12	190	0.00	0.021	15
13	113	0.52	0.037	8
14	50	0.54	0.083	4
15	47	0.20	0.086	14
16	44	0.20	0.091	15
17	24	−0.16	0.167	4
18	78	0.42	0.052	10
19	46	0.60	0.091	10
20	64	0.51	0.065	3
21	57	0.58	0.073	24
22	68	0.54	0.061	19
23	40	0.09	0.100	4
24	68	0.37	0.060	12
25	48	−0.01	0.083	1
26	107	−0.13	0.037	1
27	58	0.18	0.069	1
28	225	0.27	0.018	1
29	446	−0.02	0.009	14
30	77	0.33	0.053	20
31	243	0.59	0.017	10
32	39	0.84	0.112	7
33	67	−0.32	0.060	11
34	177	−0.12	0.023	1
35	20	−0.44	0.205	6
36	120	−0.07	0.033	15
37	16	0.70	0.265	15
38	105	0.49	0.039	2
39	195	0.20	0.021	4
40	62	0.58	0.067	24
41	289	0.15	0.014	11
42	25	0.63	0.168	15
43	250	0.04	0.016	8
44	51	1.46	0.099	15
45	46	0.04	0.087	15
46	56	0.25	0.072	15

Note. Data originally from Bangert-Drowns et al. (2004). Two studies with missing information on treatment length omitted. N_i denotes the total sample size of the study. We assumed $n_i^E = n_i^C = N_i/2$ for the computation of v_i.

Table 2
Results for the Meta-Regression Model Using the Seven Estimators of Residual Heterogeneity Combined With the Standard (Wald-Type) Test of the Model Coefficient b_1

Estimator	$\hat{\tau}^2$	b_1	$SE[b_1]$	$b_1/SE[b_1]$	p
HE	0.0645	0.016	0.0081	1.949	.051
HS	0.0373	0.015	0.0070	2.092	.036
DL	0.0424	0.015	0.0072	2.065	.039
SJ	0.0832	0.016	0.0087	1.860	.063
ML	0.0393	0.015	0.0071	2.081	.037
REML	0.0441	0.015	0.0073	2.056	.040
EB	0.0541	0.015	0.0077	2.002	.045

Note. HE = Hedges; HS = Hunter and Schmidt; DL = DerSimonian and Laird; SJ = Sidik and Jonkman; ML = maximum likelihood; REML = restricted maximum likelihood; EB = empirical Bayes.

The statistical power of all methods was lower than .80 when including 20 studies in the meta-regression analysis and when the slope parameter only had a small to moderate value (i.e., $\beta_1 = 0.2$ in our study). Moreover, all methods provided lower power rates as the residual heterogeneity among effect size parameters increased. An explanation for this fact is that, all else equal, larger τ^2 values will lead to a decrease in the predictive power of a model.[8]

In summary, results of our simulation study suggest that out of the different alternatives considered in the present study, the Knapp and Hartung method is a suitable option for most situations due to its satisfactory performance and computational simplicity. The present simulation study was conducted with standardized mean differences, but its results can be expected to apply to other effect size measures with (asymptotically) normal sampling distributions. However, it should be noted that the results of our simulation study are limited to the manipulated conditions. Although the values for the parameters and factors were chosen to represent typical conditions found in practice, additional simulation studies are needed to assess the performance of the methods under more adverse conditions, such as nonnormal random errors and/or true effects, multiple moderators with multicollinearity, categorical moderators with unbalanced designs, or results affected by publication bias.

It would have been of interest to examine the empirical coverage probability of CIs for the model coefficients. However, while the

[8] Specifically, for a slope parameter of $\beta_1 = 0.2$, values of τ^2 equal to 0, .08, and .32 correspond to $P^2 = 1$, $P^2 = .33$, and $P^2 = .11$, respectively, if the model predictive power is computed with the formula proposed by Raudenbush (1994). On the other hand, for $\beta_1 = 0.5$, the corresponding values are $P^2 = 1$, $P^2 = .76$, and $P^2 = .44$. This illustrates how the increase in τ^2 will generally lead to a decrease in the power of the statistical tests.

Table 3
Results for the Meta-Regression Model When Using Either the Standard, Knapp and Hartung, Robust (Huber–White), Likelihood Ratio, or Permutation Test of the Model Coefficient b_1 Using Maximum Likelihood Estimation

Testing method	b_1	$SE[b_1]$	$b_1/SE[b_1]$	p
Standard method	0.015	0.0071	2.081	.037
Knapp and Hartung method	0.015	0.0076	1.942	.059
Robust (Huber–White) method	0.015	0.0059	2.502	.016
Likelihood ratio test				.038
Permutation test				.052

Note. Permutation test based on 100,000 random permutations.

standard, Knapp and Hartung, and Huber–White methods are easily inverted to provide CIs (e.g., $b_j \pm 1.96\sqrt{\widehat{Var}[b_j]}$ would provide an approximate 95% CI for β_j based on the standard method), doing the same for the likelihood ratio and permutation tests would have required additional iterative methods. Due to computational constraints, we therefore opted to focus on the Type I error rate and power of the various tests. Nevertheless, we fully agree with one of the reviewers that the binary decision of a null hypothesis significance test (i.e., reject/do not reject) is often of limited value and that CIs are typically preferred. However, under $H_0: \beta_j = 0$, there is a one-to-one correspondence between the empirical Type I error rate at $\alpha = .05$ (two-sided) and the empirical coverage of the corresponding 95% CI (i.e., one minus the Type I error rate is then the coverage rate). Therefore, the simulation study does in fact provide us with information on how the various procedures compare with respect to their coverage, at least for the case where the null hypothesis holds (e.g., the standard method then yields CIs that are typically too narrow, leading to coverage rates below 95%, while the untruncated Knapp and Hartung method yields CIs with coverage probability approximately at the nominal significance level).

Finally, it is worth noting that the way moderators are tested in meta-analyses is receiving increasing attention in the literature, and several new methods have recently been developed for addressing this issue. Huizenga et al. (2011) proposed the use of a Bartlett-corrected likelihood ratio test that might improve the performance of the uncorrected likelihood ratio test regarding its Type I error rates. Guolo (2012) also recently proposed a new likelihood-based test for meta-regression models. Finally, Friedrich and Knapp (2013) presented a new method that can outperform the Knapp and Hartung method in terms of coverage probability under certain conditions. These proposals were not considered for the present comparison of methods, although it should be very interesting to evaluate their performance in future simulation studies.

References

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica, 59,* 817–858. doi:10.2307/2938229

Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of Educational Research, 74,* 29–58. doi:10.3102/00346543074001029

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

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, England: Wiley. doi:10.1002/9780470743386

Christensen, R. (1996). *Plane answers to complex questions: The theory of linear models* (2nd ed.). New York, NY: Springer. doi:10.1007/978-1-4757-2477-6

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, NY: Russell Sage Foundation.

DerSimonian, R., & Kacker, R. (2007). Random-effects model for meta-analysis of clinical trials: An update. *Contemporary Clinical Trials, 28,* 105–114. doi:10.1016/j.cct.2006.04.004

DerSimonian, R., & Laird, N. (1986). Meta-analysis of clinical trials. *Clinical Controlled Trials, 7,* 177–188. doi:10.1016/0197-2456(86)90046-2

Follmann, D. A., & Proschan, M. A. (1999). Valid inference in random effects meta-analysis. *Biometrics, 55,* 732–737. doi:10.1111/j.0006-341X.1999.00732.x

Friedrich, T., & Knapp, G. (2013). Generalised interval estimation in the random effects meta regression model. *Computational Statistics & Data Analysis, 64,* 165–179. doi:10.1016/j.csda.2013.03.011

Guolo, A. (2012). Higher-order likelihood inference in meta-analysis and meta-regression. *Statistics in Medicine, 31,* 313–327. doi:10.1002/sim.4451

Hardy, R. J., & Thompson, S. G. (1996). A likelihood approach to meta-analysis with random effects. *Statistics in Medicine, 15,* 619–629. doi:10.1002/(SICI)1097-0258(19960330)15:6<619::AID-SIM188>3.0.CO;2-A

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association, 72,* 320–338. doi:10.1080/01621459.1977.10480998

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin, 93,* 388–395. doi:10.1037/0033-2909.93.2.388

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, NY: Russell Sage Foundation.

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

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1,* 39–65. doi:10.1002/jrsm.5

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

Higgins, J. P. T., & Thompson, S. G. (2004). Controlling the risk of spurious findings from meta-regression. *Statistics in Medicine, 23,* 1663–1682. doi:10.1002/sim.1752

Hogg, R. V., & Craig, A. T. (1995). *Introduction to mathematical statistics* (5th ed.). London, England: Prentice-Hall.

Huber, P. (1967). The behavior of maximum-likelihood estimates under nonstandard conditions. In L. M. LeCam & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 221–233). Berkeley: University of California Press.

Huizenga, H. M., Visser, I., & Dolan, C. V. (2011). Testing overall and moderator effects in random effects meta-regression. *British Journal of Mathematical and Statistical Psychology, 64,* 1–19. doi:10.1348/000711010X522687

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research synthesis* (2nd ed.). Thousand Oaks, CA: Sage.

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., & 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

López-López, J. A., Marín-Martínez, F., Sánchez-Meca, J., Van den Noortgate, W., & Viechtbauer, W. (2014). Estimation of the predictive power of the model in mixed-effects meta-regression models: A simulation study. *British Journal of Mathematical and Statistical Psychology, 67,* 30–48. doi:10.1111/bmsp.12002

Manly, B. F. J. (1997). *Randomization, bootstrap and Monte Carlo methods in biology*. London, England: Chapman & Hall.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association, 78,* 47–55. doi:10.1080/01621459.1983.10477920

Paule, R. C., & Mandel, J. (1982). Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards, 87,* 377–385. doi:10.6028/jres.087.022

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

R Core Team. (2013). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

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

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

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

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

White, H. (1980). A heteroskedasticity-consistent covariance matrix and a direct test for heteroscedasticity. *Econometrica, 48,* 817–838. doi:10.2307/1912934

Received January 29, 2013
Revision received February 2, 2014
Accepted May 18, 2014 ■