OPEN ACCESS

University of BRISTOL

Lopez-Lopez, J. A., Van den Noortgate, W., Tanner-Smith, E., Wilson, S., & Lipsey, M. (2017). Assessing meta-regression methods for examining moderator relationships with dependent effect sizes: A Monte Carlo simulation. *Research Synthesis Methods*, *8*(4), 435-450. https://doi.org/10.1002/jrsm.1245

Peer reviewed version

License (if available):
CC BY-NC

Link to published version (if available):
10.1002/jrsm.1245

Link to publication record in Explore Bristol Research
PDF-document

© 2017 Crown copyright. Research Synthesis Methods. This is the author accepted manuscript (AAM). The final published version (version of record) is available online via Wiley at http://onlinelibrary.wiley.com/doi/10.1002/jrsm.1245/full. Please refer to any applicable terms of use of the publisher. This article is published with the permission of the Controller of HMSO and the Queen's Printer for Scotland.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/

Assessing meta-regression methods for examining moderator relationships with dependent effect sizes: A Monte Carlo simulation

José Antonio López-López

University of Bristol, Bristol, UK

Wim Van den Noortgate

University of Leuven, Leuven, Belgium

Emily E. Tanner-Smith, Sandra Jo Wilson, and Mark W. Lipsey

Peabody Research Institute, Vanderbilt University, Nashville, TN, USA

Running Head: Meta-regression with dependent effect sizes

Author Footnotes

Corresponding author:

José Antonio López-López, Ph.D.

Senior Research Associate (Statistician)

E-mail: ja.lopez-lopez@bristol.ac.uk

Phone: + 44 (0)117 9287343

Address for correspondence:

School of Social and Community Medicine

University of Bristol

Canynge Hall, 39 Whatley Road

Bristol BS8 2PS, UK

Abstract

Dependent effect sizes are ubiquitous in meta-analysis. Using Monte Carlo simulation, we compared the performance of two methods for meta-regression with dependent effect sizes—robust variance estimation (RVE) and three-level modeling—with the standard meta-analytic method for independent effect sizes. We further compared bias-reduced linearization and jackknife estimators as small-sample adjustments for RVE, and Wald-type and likelihood ratio tests for three-level models. The bias in the slope estimates, width of the confidence intervals around those estimates and empirical Type I error and statistical power rates of the hypothesis tests from these different methods, were compared for mixed-effects meta-regression analysis with one moderator either at the study or at the effect size level. All methods yielded nearly unbiased slope estimates under most scenarios, but as expected, the standard method ignoring dependency provided inflated Type I error rates when testing the significance of the moderators. RVE methods yielded the best results in terms of Type I error rate, but also the widest confidence intervals and the lowest power rates, especially when using the jackknife adjustments. Three-level models showed a promising performance with a moderate to large number of studies, especially with the likelihood ratio test, and yielded narrower confidence intervals around the slope and higher power rates than those obtained with the RVE approach. All methods performed better when the moderator was at the effect size level, the number of studies was moderate to large, and the between-studies variance was small. Our results can help meta-analysts deal with dependency in their data.

Key-words: meta-analysis, meta-regression, dependency, robust variance estimation, three-level model.

Assessing meta-regression methods for examining moderator relationships with dependent effect sizes: A Monte Carlo simulation

Introduction

Heterogeneous effect sizes are common in meta-analyses of intervention studies and identifying moderator variables that may account for some of that variability is often an important objective of a meta-analysis (Hedges and Olkin, 1985; Lipsey and Wilson, 2001; Sánchez-Meca and Marín-Martínez, 2010). For example, a meta-analyst may be interested in the relationships between the effect sizes and such moderator variables as the type of intervention, characteristics of the participant samples, methodological procedures such as randomization or the operationalization of dependent variables, and characteristics of the research context such as where an intervention was delivered (Lipsey, 2009). Indeed, many meta-analysts are moving away from asking only about average effect sizes ("does the intervention work?") to also exploring moderators of effect sizes ("for whom and under what conditions does the intervention work best?").

The most appropriate meta-analytic model for examining moderator relationships is a carefully conducted and interpreted meta-regression (Baker et al., 2009; Thompson and Higgins, 2002). Like multiple regression analysis with primary data, standard meta-regression analysis assumes that, after controlling for the predictor effects, residuals within a given analysis are statistically independent (e.g., Stevens and Taylor, 2009). But dependency among effect size estimates can occur in many ways and is quite common (Ahn et al., 2012; Becker, 2000; Hedges et al., 2010; Van den Noortgate et al., 2013, 2015). An especially common type of dependency arises when multiple effect sizes are extracted from the same participant sample on similar outcome constructs (i.e., multiple effect sizes are clustered within studies). For instance, a meta-analysis looking at the effectiveness of cognitive-behavioral interventions for depressed adults might find that some studies report depression scores on multiple scales, so that multiple effect sizes can be extracted from the same study. If the variance between clusters of effect sizes is not accounted for in the analysis, the standard errors of the regression coefficients may be underestimated, leading to statistical significance tests that are spuriously liberal.

Some methods for handling this kind of dependency require knowing the covariance structure of the outcome variables on which the multiple effect sizes are based in each study (Gleser and Olkin, 2009; Tipton, 2013). However, that information is rarely reported in primary studies or otherwise available. To satisfy the assumption of independent effect sizes when dependencies exist, most meta-analysts historically have created one effect size per cluster by averaging effect sizes within each cluster or applying some rule for selecting one effect size from the cluster (Hedges and Olkin, 1985; Marín-Martínez and Sánchez-Meca, 1999; Rosenthal and Rubin, 1986). Other meta-analysts, however, have ignored the dependencies and incorrectly analyzed the whole set of effect sizes using standard methods (as noted in Gleser and Olkin, 2009; see also Jackson et al., 2011).

Because these strategies involve a loss of information or an increased risk of misleading findings, an important advance has been the development of new statistical methods for dealing with dependent effect sizes. Two such methods—robust variance estimation (RVE) and multilevel modeling—are sufficiently well developed and accessible to offer attractive options for researchers undertaking a meta-analysis of intervention studies that report effects on multiple outcomes of interest. Neither the RVE nor the multilevel approach requires knowledge of within-study correlations, which is an important advantage that allows these methods to be implemented widely in meta-regression applications.

The rationales for RVE and multilevel approaches are different, and these approaches estimate different parameters. Namely, the RVE method allows use of a straightforward mixed-effects meta-regression models and estimates only one variance parameter, the between-studies variance (see Hedges et al., 2010), but makes adjustments in the standard errors to better represent the interdependence of the clustered effect sizes. On the other hand, multilevel models are often used to analyze clustered data (e.g., effect sizes nested in studies) by decomposing the variance in the dependent variable into between and within clusters variance components, which need to be estimated separately (e.g., Cheung, 2014; Konstantopoulos, 2011). If the variance between clusters is larger than zero, the dependent variable values within clusters are more similar than those from different clusters. By modeling both variance components, these within-cluster dependencies are thus accounted for.

Given these differences, a meta-analyst planning to perform meta-regression on a database with dependent effect sizes might well ask whether these two approaches can be expected to yield the same results and, if not, which provides more accurate results under what circumstances. A particular complication in such applications is the multilevel nature of potential moderator variables, some of which occur at the study level (e.g., sample characteristics) and some of which occur at the within-study (or effect size) level (e.g., measurement characteristics for the multiple effect sizes within a study). Assessing the relative performance of the RVE vs. multilevel modeling approaches for moderator analysis under these circumstances requires comparison of their respective results across a range of realistic meta-analytic scenarios. Of particular importance for practical application are any differences in the accuracy with which these methods estimate the regression coefficients for study level and effect size level moderators, and the validity of the associated statistical significance tests.

Objectives and hypotheses of this study

The study reported here uses Monte Carlo simulation techniques to examine three alternative methods for conducting meta-regression with dependent effect sizes derived from intervention studies. We focus solely on the common situation in which dependency arises from correlated error terms due to individual effect sizes clustered within studies and consider both moderators that vary across studies and those that vary across effect sizes within studies. Our simulations include a wide range of scenarios by manipulating the number of studies, the number of outcomes extracted from each study, the overlap among outcomes within the same study, and the degree of heterogeneity in the effects across studies, with the aim to provide some guidance about choice of statistical method for a range of realistic conditions.

The first analysis strategy we examine is the use of standard mixed-effects meta-regression models that ignore the dependency structure of the effect sizes clustered within studies. While this approach is not appropriate in the face of such dependencies, we include it, first, to assess the extent of the errors it produces and, second, to provide something of a baseline against which to compare the performance of the other two approaches. We expect this analysis strategy to yield standard error estimates that are too small (Becker, 2000; Van den Noortgate et al., 2013,

2015) with the statistical tests of the moderator relationships, therefore, showing unacceptably inflated Type I error rates.

The other two analysis strategies examined here are the RVE and the multilevel model approaches. For the multilevel approach, we focus on three-level models. Both analysis approaches attempt to account for the dependency structure of the multiple within-study effect sizes included in the meta-regression. Consequently, we expect both methods to outperform the standard method in terms of accuracy of the statistical tests as dependency increases. Remaining questions for any meta-analyst dealing with dependent effect sizes are whether these approaches also perform well for estimating and testing moderators at the effect size or study level, and which of these approaches performs best. To our knowledge, this is the first simulation study comparing both RVE and three-level model approaches. Further details of each approach are provided in the next section.

The criteria we used to examine the results of the different methods focused on the considerations likely to be most important to meta-analysts using these methods. An initial concern, of course, is the accuracy with which the regression coefficients for the moderator variables are estimated. The expectation is that all the methods will estimate the coefficients without bias, while the different ways of handling the statistical dependencies will affect the standard errors for those coefficients. Misestimation of the standard errors would produce erroneous conclusions from statistical significance tests, so we examined the empirical Type I error rates when the true regression coefficient was zero and compared them with the nominal alpha = .05 rate stipulated in the significance tests. Additionally, we examined the statistical power rates when the true regression coefficient was different from zero. We further examined the width of the confidence intervals for the estimates of the regression coefficients. While related to the Type I error and statistical power rates, the width of the confidence interval more directly reflects the precision of the estimate.

Mixed-effects meta-regression models

The present study is focused on standardized mean difference effect sizes with experimental and control groups compared in terms of their mean scores on a continuous dependent variable representing an intervention effect of interest. Assuming a common population standard

deviation under both conditions, an unbiased estimator of the standardized mean difference in the *i*th study, d_i, can be obtained with the expression (Hedges and Olkin, 1985)

$$d_i = \left(1 - \frac{3}{4(n_{iE} + n_{iC}) - 9}\right) \frac{\overline{Y}_{iE} - \overline{Y}_{iC}}{S_i}, \quad (1)$$

where n_{iE} and n_{iC} are the sample sizes for the experimental and control groups, $\overline{Y}_{iE}$ and $\overline{Y}_{iC}$ represent their means, and S_i is the pooled standard deviation, computed as

$$S_i = \sqrt{\frac{(n_{iE} - 1)S_{iE}^2 + (n_{iC} - 1)S_{iC}^2}{n_{iE} + n_{iC} - 2}}, \quad (2)$$

with S_{iE}^2 and S_{iC}^2 being the variances for the scores of the respective groups. An estimate of the within-study variance for the *i*th study, v_i, is then obtained with

$$v_i = \frac{n_{iE} + n_{iC}}{n_{iE} n_{iC}} + \frac{d_i^2}{2(n_{iE} + n_{iC})}. \quad (3)$$

In the remainder of this section, we present different alternatives for fitting mixed-effects meta-regression models. First, we briefly outline the standard meta-analytic method that assumes independent effect sizes. Next we describe the RVE and three-level hierarchical models for meta-regression with dependent effect sizes. For all these methods, the model is presented first, followed by its estimators and statistical tests.

Standard meta-analytic method

Although alternatives are available in the literature (e.g., Hunter and Schmidt, 2004), the meta-regression approach considered here is that proposed by Hedges and Olkin (1985) because it has been the most widely employed when dealing with standardized mean difference effect sizes. According to this approach, in a meta-analytic database with *r* rows (with *r* being the total number of effect sizes), let $\mathbf{T}$ be an (*r* x 1) vector of effect sizes, and $\mathbf{X}$ an [*r* x (*p* + 1)] design matrix with a column of ones, and a column for each of *p* moderator variables. Then, a mixed-effects meta-regression model is defined with the expression

$$\mathbf{T} = \mathbf{Xb} + \mathbf{u} + \mathbf{e}, \tag{4}$$

Where $\mathbf{T}$ is the (r x 1) vector with the observed effect sizes, $\mathbf{b}$ is a [(p + 1) x 1] vector containing the population regression coefficients $\{\beta_0, \beta_1, ..., \beta_p\}$, $\mathbf{u}$ is an (r x 1) vector of random study-specific effects with distribution N(0, τ^2), and $\mathbf{e}$ is an (r x 1) vector of within-study errors with distribution N(0, v_i), with v_i assumed to be known and defined in Equation 3. The parameter τ^2 therefore refers to the amount of residual between-studies variance, which is the variability in the true effects not accounted for by the moderators included in the model. A weighted least squares (WLS) estimator of $\mathbf{b}$ can be computed with the formula

$$\hat{\mathbf{b}}_{STD} = \left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{W}}\mathbf{T}, \tag{5}$$

where $\hat{\mathbf{W}}$ is an (r x r) diagonal weighting matrix. In a mixed-effects model, the weights that maximize the precision are

$$w_i = 1/\left(v_i + \tau^2\right). \tag{6}$$

The value of τ^2 needs to be estimated, and the most widely employed estimator in random effects meta-regression models is the method of moments (DerSimonian and Laird, 1986), given by the expression (Raudenbush, 2009)

$$\hat{\tau}^2_{STD} = \frac{Q_E^{STD} - (r - p - 1)}{tr(\mathbf{M})}, \tag{7}$$

where tr denotes the trace of a matrix and $\mathbf{M}$ is obtained with

$$\mathbf{M} = \mathbf{W} - \mathbf{WX}(\mathbf{X}'\mathbf{WX})^{-1}\mathbf{X}'\mathbf{W}, \tag{8}$$

with $\{1/v_i\}$ elements for $\mathbf{W}$. Moreover, the residual heterogeneity test statistic, Q_E^{STD}, is computed as

$$Q_E^{STD} = \mathbf{T}'\mathbf{MT}. \tag{9}$$

The variance-covariance matrix for the model regression coefficient estimates can be estimated with the expression

$$\hat{\boldsymbol{\Sigma}}_{STD} = (\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}. \quad (10)$$

Then, a 100(1 - α)% confidence interval assuming a standard normal distribution can be calculated with

$$\hat{\beta}_j^{STD} \pm z_{1-\alpha/2}\sqrt{V(\hat{\beta}_j^{STD})}, \quad (11)$$

where $\hat{\beta}_j^{STD}$ is the element (j + 1) of the $\hat{\mathbf{b}}_{STD}$ vector, computed with Equation 5, $z_{1-\alpha/2}$ is the 100(1 - α/2) percentile of the standard normal distribution, α is the significance level, and $V(\hat{\beta}_j^{STD})$ is the diagonal element (j + 1) of the $\hat{\boldsymbol{\Sigma}}_{STD}$ matrix, defined in Equation 10. Finally, the statistical test for the effect of the jth moderator variable can be obtained with the Wald-type formula

$$z_j = \frac{\hat{\beta}_j^{STD}}{\sqrt{V(\hat{\beta}_j^{STD})}}. \quad (12)$$

Robust variance estimation (RVE)

In the RVE framework (Hedges, Tipton, and Johnson, 2010; Tipton, 2013), Equation 4 is modified to account for the fact that data are grouped in k clusters (i.e., studies). Thus, the model is defined with the formula

$$\begin{pmatrix} \mathbf{T_1} \\ \mathbf{T_2} \\ \cdots \\ \mathbf{T_c} \\ \cdots \\ \mathbf{T_k} \end{pmatrix} = \begin{pmatrix} \mathbf{X_1} \\ \mathbf{X_2} \\ \cdots \\ \mathbf{X_c} \\ \cdots \\ \mathbf{X_k} \end{pmatrix} \mathbf{b} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \cdots \\ \boldsymbol{\varepsilon}_c \\ \cdots \\ \boldsymbol{\varepsilon}_k \end{pmatrix}, \quad (13)$$

where $\mathbf{T_c}$ is an (a_c x 1) vector, a_c being the number of effect sizes in cluster c (any of the k clusters); $\mathbf{X}_c$ is the [a_c x (p + 1)] design matrix for data in cluster c, $\boldsymbol{\varepsilon}_c$ is the (a_c x 1) vector of residuals from cluster c. The estimator of the model regression coefficients, $\mathbf{b}$, is expressed now as

$$\hat{\mathbf{b}}_{\mathbf{RVE}} = \hat{\mathbf{U}}^{-1}\sum_{\mathbf{c=1}}^{k}\mathbf{X}_c'\hat{\mathbf{W}}_\mathbf{c}\mathbf{T_c}\,, \tag{14}$$

and $\hat{\mathbf{U}}$ is given by

$$\hat{\mathbf{U}} = \sum_{\mathbf{c=1}}^{k}\mathbf{X}_c'\hat{\mathbf{W}}_\mathbf{c}\mathbf{X_c}\,. \tag{15}$$

Weights are the non-zero elements for each diagonal matrix $\hat{\mathbf{W}}_\mathbf{c}$. They are defined as $\hat{w}_{ic} = 1/\left(v_{.c} + \hat{\tau}^2_{RVE}\right)$, where $v_{.c} = \sum v_{ic}/a_c$ is the average sampling variance within the cth cluster, and $\hat{\tau}^2_{RVE}$ is estimated with the expression

$$\hat{\tau}^2_{RVE} = \frac{Q_E^{RVE} - k + tr\left(\mathbf{U^{-1}U}\right) + (\hat{\rho})tr\left[\mathbf{U^{-1}}\left(\sum_{\mathbf{c=1}}^{k}\mathbf{X}_c'\mathbf{W_cX_c} - \mathbf{X}_c'\mathbf{X_c}\right)\right]}{\sum_{\mathbf{c=1}}^{k}\mathbf{W_c} - tr\left[\mathbf{U^{-1}}\left(\sum_{\mathbf{c=1}}^{k}\mathbf{X}_c'\mathbf{W_cW_cX_c}\right)\right]}, \tag{16}$$

where the weights employed for the $\mathbf{U}$ and $\mathbf{W_c}$ matrices are initial weights using the inverse of the within-study variance, that is, $w_{ic} = 1/v_{.c}$. Moreover, Q_E^{RVE} is now defined as (Hedges et al., 2010)

$$Q_E^{RVE} = \sum_{\mathbf{c=1}}^{\mathbf{k}}\mathbf{T_c'W_cT_c} - \left(\sum_{\mathbf{c=1}}^{\mathbf{k}}\mathbf{T_c'W_cX_c}\right)\mathbf{U^{-1}}\sum_{\mathbf{c=1}}^{\mathbf{k}}\mathbf{X_c'W_cT_c}\,, \tag{17}$$

with $\{\, w_{ic} = 1/v_{.c} \,\}$ elements for both $\mathbf{U}$ and $\mathbf{W_c}$. (Note that Q_E^{RVE} is an extension of Q_E^{STD}, so it could also be defined as $Q_E^{RVE} = \sum_{\mathbf{c=1}}^{k}\mathbf{T_c'M_cT_c}$. However, in practice this would imply solving k

matrices and, because one or more of those are likely to be singular, this solution is not efficient). Moreover, $\hat{\rho}$ is a scalar quantifying the correlation between effect sizes in the same cluster (assuming a common correlation between all pairs of effect sizes). Although some methods for computing $\hat{\rho}$ have recently been proposed (Ahn et al., 2012), Hedges and colleagues (2010) found that this is not a major issue in the calculation of $\hat{\tau}^2_{RVE}$ because the value of $\hat{\rho}$ has little effect on the resulting estimate.

The variance-covariance matrix for the model regression coefficients then can be obtained with the expression

$$\mathbf{\hat{V}_{RVE}} = \mathbf{\hat{U}^{-1}}\left(\sum_{\mathbf{c=1}}^{k}\mathbf{X_c'\hat{W}_c\hat{A}_c\hat{e}_c\hat{e}_c'\hat{A}_c\hat{W}_cX_c}\right)\mathbf{\hat{U}^{-1}}, \tag{18}$$

with $\hat{w}_{ic} = 1/\left(v_{.c} + \hat{\tau}^2_{RVE}\right)$ weights for both $\mathbf{\hat{U}}$ and $\mathbf{W_c}$, $\mathbf{\hat{e}_c} = \mathbf{T_c} - \mathbf{X_c\hat{b}_{RVE}}$, and $\mathbf{\hat{A}_c}$ stands for an adjustment factor to be discussed below in this section. A 100(1 - α)% confidence interval assuming a t-distribution can be obtained with

$$\hat{\beta}_j^{RVE} \pm {}_{1-\alpha/2}t_{df}\sqrt{V\left(\hat{\beta}_j^{RVE}\right)}. \tag{19}$$

Furthermore, the statistical test is given by

$$T = \frac{\hat{\beta}_j^{RVE}}{\sqrt{V\left(\hat{\beta}_j^{RVE}\right)}}, \tag{20}$$

which is compared with critical values of the t-distribution with $k - p - 1$ degrees of freedom. However, this test has been found to provide inflated Type I error rates in various scenarios (Sidik and Jonkman, 2005; Viechtbauer et al., 2015) and various adjustments have been proposed (Cribari-Neto and Da Silva, 2011; MacKinnon and White, 1985). Tipton (2015) studied the performance of several such adjustments and found that those that performed best corrected both the standard errors (obtained from Equation 18) and the degrees of freedom for the t-test (Equation 20). In particular, she found that the bias reduced linearization estimator (MBBS) that was proposed by Bell and McCaffrey (2002) and extended to weighted least

squares by McCaffrey and colleagues (2001) and the jackknife estimator (JKS) provide accurate rejection rates across a wide range of scenarios. We therefore included both in our comparison of different approaches for handling dependent effect sizes in meta-regression.

RVE meta-regression can be implemented in R (using the *robumeta* package) or using macros developed for SPSS and Stata (see Fisher and Tipton, 2014; Tanner-Smith and Tipton, 2014).

Three-level model

The multilevel meta-analytic approach builds on the inherently multilevel structure of meta-analytic data, which has participants clustered within studies (e.g., Raudenbush and Bryk, 2002; Van den Noortgate and Onghena, 2003). Whereas traditional random effects models can be considered as two-level models (participants within studies), three-level models have been proposed as a way to deal with dependent effect sizes (Beretvas and Pastor, 2003; Van den Noortgate et al., 2013). These models include an intermediate level to represent the clustering of effect sizes within studies.

In three-level hierarchical models the *i*th effect size in the *c*th cluster (e.g., study), T_{ic}, is equal to the population effect size value, θ_{ic}, for the respective outcome and cluster plus a random deviation due to studying a sample of participants rather than the whole population:

$$T_{ic} = \theta_{ic} + e_{ic}, \tag{21}$$

where e_{ic} is the error term at Level 1, with distribution $N\left(0, v_{ic}\right)$. When the three-level approach is applied to a meta-analytic database, as in the two previous approaches, the usual practice is to constrain the sampling variances of the effect size estimates (the level-1 variances) to their estimated value (Equation 3). The population effect for outcome *i* in cluster *c* can vary, both randomly and as a function of the characteristics of the outcomes:

$$\theta_{ic} = \beta_{0c} + \beta_{1c} X_{1ic} + ... + \beta_{Pc} X_{Pic} + h_{ic}, \tag{22}$$

where $X_{1ic}, ..., X_{Pic}$ denote each of the *P* effect size level moderators, and $h_{ic} \sim N\left(0, \sigma_h^2\right)$, with σ_h^2 being the residual variance between outcomes from the same study. At Level 3, model

coefficients from Equation 22, $\beta_{0c}, \beta_{1c}, ..., \beta_{Pc}$, are allowed to vary among different clusters. For example, the predicted values for the intercepts of the outcomes within the *c*th cluster, β_{0c}, vary as a function of

$$\beta_{0c} = \gamma_{00} + \gamma_{01} Z_{1c} + ... + \gamma_{0P'} Z_{P'c} + l_{0c}, \tag{23}$$

where $\gamma_{00}, ..., \gamma_{0P}$ are the level 3 regression coefficients and $Z_{1c}, ..., Z_{P'c}$ denote each of the *P'* study level moderators. The same rationale would apply for the remaining model coefficients, $\beta_{1c}, ..., \beta_{P'c}$ but for this study, we assumed that the regression coefficients $\beta_{1c}, ..., \beta_{P'c}$ are the same over all clusters, that is $\beta_{pc} = \gamma_{p0}$ with p = 1, 2, ..., P. Finally, σ_l^2 denotes the variance between clusters in the intercept.

Parameter estimation in three-level models requires iterative computation using either Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) algorithms. The estimates of the model coefficients and their standard errors can be used to calculate a 100(1 – α)% confidence interval assuming a standard normal distribution, as in Equation 11, although a 100(1 – α)% likelihood-based confidence interval can also be obtained (Cheung, 2014). A more detailed description of the estimation process can be found elsewhere (Konstantopoulos, 2011; Raudenbush and Bryk, 2002). For the present study, we employed ML algorithms because they are more appropriate than REML algorithms when the aim is to compare different models that also differ in the fixed part, and each effect size is weighted by its inverse total variance.

Several alternatives are available to the researcher for testing the statistical significance of the regression model coefficients in three-level models. In this study, we examined the results from two widely implemented tests in multilevel modelling, namely a *z*-based and a likelihood-based strategy (Cheung, 2014; Raudenbush and Bryk, 2002; Snijders and Bosker, 1999). The former is based on a Wald-type *z*-test, as defined in Equation 12, whereas the latter entails implementing a likelihood ratio test, which compares the change in the deviance of two nested models and is computed with the expression (Raudenbush and Bryk, 2002):

$$\chi^2 = -2 \ln \frac{L_0}{L_1}, \tag{24}$$

where L_0 is the likelihood of the null model (not including the *j*th moderator) and L_1 is the likelihood of the model including the *j*th moderator. The result is compared against the critical value of a Chi-square distribution with one degree of freedom (i.e., 3.84 for $\alpha = 0.05$). The three-level model has been found to outperform the standard method ignoring dependency in the estimation of standard errors when the meta-analytic database includes dependent effect sizes (Van den Noortgate et al., 2013; 2015). What has not yet been investigated is its performance for meta-regression or how that performance compares with that of the RVE approach when both are applied to the same data.

For general application, three-level models for dependent effect sizes clustered within studies can be implemented in R, using the *metaSEM* (see Cheung, 2014 for some example code) or the *metafor* packages, and also in SAS (Proc Mixed, see Van den Noortgate et al., 2015).

An illustrative example

To illustrate the potential variation in results and conclusions a meta-analyst might draw from using the meta-regression methods outlined above, we applied these methods to a meta-analysis of intervention programs for preventing antisocial behavior (Wilson et al., 2003; Wilson and Lipsey, 2007). The analytic dataset used in this example contained 870 standardized mean difference effect sizes (ranging from -4.34 to 6.28 with a mean of 0.26) extracted from 316 studies with, therefore, a mean of 2.8 effect sizes per study.

We fitted several simple meta-regression models for two moderators selected from the database, one at the study level and the other at the effect size level. The study-level moderator represented the existence of implementation problems in the study (0= No, 1= Yes); the effect size level moderator was the degree of content overlap (alignment) between the outcome measure and the intervention components (0=Low, 1=High).

All analyses were conducted in the R statistical environment. For the standard method, the *metafor* package was used (Viechtbauer, 2010), running the analysis as if the effect sizes were independent and using weights as defined in Equation 6 and the DerSimonian and Laird method to estimate τ^2. For the RVE method, the analyses used the MBBS corrections implemented in the *robumeta* package (Fisher and Tipton, 2014), as well as the JKS corrections as implemented in

Tipton (2015). For this method, we set a value of .50 for the correlation between effect sizes, $\hat{\rho}$. (We conducted additional analyses using a value of .80 for this correlation and the results were almost identical). Lastly, analyses for the three-level models were computed as described in the previous section using the *metaSEM* package (Cheung, 2015), which makes use of *OpenMx* (Boker et al., 2011). For the three-level approach, maximum likelihood estimation algorithms were employed weighting each effect size by its inverse sampling variance estimate.

The estimates of the regression coefficients, the widths of the confidence intervals for those estimates and the p-values of the statistical tests are presented in Table 1. Note that two different results are presented for the RVE method, due to the different approaches for small-sample corrections considered in this study (MBBS and JKS). Likewise, for the three-level approach we compared the performance of confidence intervals with Wald-type z-tests and likelihood-based confidence intervals with likelihood ratio tests.

TABLE 1 HERE

As shown in Table 1, there were remarkable differences in the results obtained for the different methods. For the study level moderator, the estimates of the regression coefficient were 0.106 for the standard method ignoring dependency, 0.224 for the RVE method and 0.257 for the three-level method. The confidence interval for the estimate was substantially narrower for the standard method, whereas the widest intervals were obtained with the RVE method using the JKS corrections. Only the statistical test of the standard method reached statistical significance, whereas the tests of the three-level model yielded marginally significant results and the tests of the RVE method showed p-values above 0.10. For the effect size level moderator, the estimates of the coefficients were 0.109 for the standard method, 0.201 for the RVE method and 0.155 for the three-level method. Here also the narrowest confidence interval was obtained with the standard method, followed by the three-level method. The widest interval was again obtained when using the JKS corrections for the RVE method. Statistical significance for this moderator was reached when using the standard method and both of the three-level method variations, but not for the RVE method.

This illustration with data from an actual meta-analysis revealed clear differences in the coefficient estimates, but most importantly in the conclusions that would be drawn from these

three methods both with regard to the confidence intervals of the estimates of the regression coefficients and their statistical significances. We turn now to a description of the simulations we conducted to better understand the performance of the three methods under different conditions.

Simulation studies

The simulations were conducted to compare the alternative methods described above for fitting meta-regression models with multiple standardized mean difference effect sizes clustered within studies. Two separate simulations were undertaken in which the sole difference was the use of either a study-level or an effect size-level moderator in the meta-regression models. The simulations were programmed in the R statistical environment, using random number generators to produce the matrices of raw scores within each study.

We manipulated several factors in our simulations. First, we used values of k = {5, 10, 20, 40, 80} for the number of studies. Second, the number of clustered effect sizes per study was manipulated to simulate a range and distribution representative of what is generally found in intervention studies. The average number of effect sizes per study was set to values of $\bar{a}$ = {2, 4, 8}, so that scenarios with low, medium, and high average cluster sizes were present in the simulations. Variation across studies around those means was generated by drawing random values from Chi-square distributions (with 2, 4 and 8 degrees of freedom, respectively) and rounding them to the nearest integer. Because of the positive asymmetry in the Chi-square distribution, for example, a mean value of 8 effect sizes per study averages over a few studies providing one or two effect sizes, a majority of studies providing between 3 and 12 effect sizes, and the few remaining studies contributing between 13 and approximately 30 effect sizes.

A two-group design was defined for all of the simulated studies, where the first (experimental, E) group received an intervention and the other (control, C) group was not treated. The participant sample size varied across studies but, within each study, $n_E = n_C$ was assumed. In a review of several meta-analyses from journals focused on behavioral sciences, Sánchez-Meca and Marín-Martínez (1998) found that participant sample size distributions are usually not symmetric, and they reported an average asymmetry value of +1.46. Therefore, for the present simulations, participant sample sizes were generated from an asymmetric (log-normal) distribution with a mean of 25 participants per group and an asymmetry level of +1.46, resulting in an approximate

range of between 5 and 100 participants per study. In order to explore the influence of sample size, we also generated scenarios with a mean of 500 participants per group and present results as supplementary figures.

For each study, raw data were generated from a multivariate standard normal distribution, referring to the scores that could have been observed for a number of participants on multiple outcomes. The strength of the correlation among outcomes was also manipulated in our simulations. Namely, values for the intercorrelations among the dependent variables on which the effect sizes were based within each study were obtained from uniform distributions with a range of values of either [.10, .50] or [.50, .90], with the aim to reflect conditions of low to moderate and moderate to large amounts of dependency among effect sizes within the same study. In order to simulate an intervention effect, a population effect size was added to the scores of half of the participants (those belonging to the experimental group). In this way, the expected standardized mean difference between the experimental and control group is equal to the chosen population effect size. The population effect size for the *c*th study, δ_c, was calculated using a regression model including the moderator variable and a random study effect

$$\delta_c = \beta_0 + \beta_1 X_c + \eta_c, \quad (25)$$

where β_0 is the intercept of the regression model, set equal to 0.5, β_1 is the model slope, another manipulated factor, which was set to values of 0 or 0.2 along the simulated conditions; X_c represents a column vector with the moderator values for the *c*th study, and η_c is the error term for the *c*th study, with distribution N~(0, τ^2). Likewise, when the simulated moderator was at the effect size level, the regression equation to generate final scores for the experimental group was

$$\delta_c = \beta_0 + \beta_1 X_{ic} + \eta_c, \quad (26)$$

where X_{ic} is now a column vector with the moderator values for the *i*th outcome from the *c*th study, and the rest remains as in Equation 25. In order to simulate realistic scenarios, a conditional statement was added so that effect sizes stemming from highly correlated variables within the same study were more likely to have the same moderator value. Both the study and

effect size level moderators were continuous variables randomly generated from a standard normal distribution.

The residual between-studies variance (e.g., unexplained variance at the study level after the effect moderators have been included in the model) was also manipulated, with values $\tau^2 = \{0, 0.08, 0.32\}$. A value of 0 means that there are no differences in the effect sizes between studies once the moderator effects have been accounted for, as assumed by the fixed-effect model (e.g., Borenstein et al., 2010), whereas the two remaining values reflect conditions with moderate and large heterogeneity, respectively[1].

For each combination of the 3 values of *k,* 3 values of $\bar{a}$, 2 ranges of correlations among outcomes, 2 values of β_1, and 3 values of τ^2 1,000 meta-analytic data sets were generated, leading to 108,000 simulated meta-analyses for each of the two simulations (one with moderator at the study level, one with moderator at the effect size level). Each meta-analytic database was analyzed using the same approaches described in the example.

The accuracy of the slope estimates produced by each method was assessed by the bias for each condition. A comparison between the observed bias and the true value of the slope provides information on relative bias (e.g. percentage of bias), which can be useful to assess the magnitude of bias and the potential implications of the results for applied meta-analyses using the methods examined in this study. Bias was computed with

$$BIAS\left(\hat{\beta}_1\right) = MEAN\left(\hat{\beta}_1\right) - \beta_1 . \quad (27)$$

To assess the performance of the statistical significance tests for both study-level and effect size level moderators, we used the empirical Type I error rate (when $\beta_1 = 0$). Because the statistical tests in the simulation studies were computed assuming a 5% significance level, rejection rates close to 0.05 indicate a good performance for the statistical method when the true value for the model coefficient is zero. To assess the adequacy of the confidence intervals for the regression coefficients estimated by each meta-regression method, we considered the width of the confidence intervals obtained with each method. Once the Type I error rate is accurate (0.05 for 95 % confidence intervals), narrower confidence intervals represent more precise estimates of

the respective coefficients and higher power, and thus better performance for the meta-regression method being assessed. Last, we examined the statistical power rates of the different methods when $\beta_1 = 0.2$, considering rates above 0.8 as reflecting adequate power (Cohen, 1988).

Results

In this section, we compare the performance of the different meta-regression methods based on the criteria described above. First, we examine the accuracy of the estimates of the regression coefficients produced by each method through examining the bias in each condition. Next, we assess the performance of the statistical tests as shown by the empirical Type I error rates. Then we assess the accuracy of the confidence intervals for the regression coefficient estimates by looking at the interval width. Last, we examine the statistical power of the statistical tests. Given the large number of simulation conditions and results generated, only a subset is presented here, although additional figures (for conditions with an average of 500 participants per group) and the full data set of results are provided as Web Appendices. Because different factors were manipulated in the simulations, we present exhaustive tables containing results for each method in each condition for bias, whereas to report Type I error, confidence width results and statistical power we use graphs showing average values on the performance criteria for one factor at a time.

Bias in the slope estimates

Bias results for the different methods and simulated conditions at the study and effect size level are presented in Table 2 and Table 3, respectively. Here we only present bias results for the scenarios with $\beta_1 = 0.2$ $\tau^2 = 0.08$, although the remaining conditions yielded very similar results.

TABLE 2 HERE

Table 2 presents the bias of the slope estimates for the moderator at the study level. Most estimates were negatively biased on average, although bias only exceeded 5% in a few instances that are marked in bold. All of those instances correspond to conditions with five studies. Most of these values were also found when the range of correlations among outcomes in the same study was between 0.5 - 0.9, and when the slope was estimated using the standard method

ignoring dependency. Nonetheless, all methods provided accurate slope estimates across most scenarios.

TABLE 3 HERE

Regarding bias at the effect size level, values in Table 3 suggest that all three methods provided reasonably accurate estimates of the model slope across conditions. Once again, most estimates were negatively unbiased, although the percentage of bias was always below 5%.

Empirical Type I error rates

The empirical Type I error rates for the range of values on each of the factors of interest when testing a moderator at the study or at the effect size level, are presented in Figure 1 and Figure 2, respectively.

FIGURE 1 HERE

At the study level, average Type I error rates (Figure 1) for the standard method were greater than 0.2 in most conditions, and error rates for this method increased as the average number of outcomes per study (Figure 1A), the correlation among outcomes (Figure 1B) and the residual between-studies variance (Figure 1D) also increased. Rates for the three-level models were also too high unless the number of studies was at least 20, with the likelihood ratio test performing somewhat better than the Wald-type (z) test. For the RVE method, the jackknife estimator provided somewhat conservative results across all examined conditions (mainly in the .03 range), whereas the MBBS estimator consistently performed close to the nominal .05 significance level, albeit slightly conservative. Results for the methods accounting for dependency were relatively consistent across the different values of the different factors except for the number of studies (Figure 1C). With up to 20 studies, all the Type I error rates were somewhat more extreme in the direction of their general bias, fairly typical of their general bias with 40 studies, and less biased with 80 studies. Results with an average of 500 participants per group led to the same conclusions (Suppl. Figure 1).

FIGURE 2 HERE

Concerning results at the effect size level, the Type I error rates (Figure 2) showed similar trends as those found at the study level, with the standard method yielding rates over .1 for most conditions and leading to the wrong statistical conclusion even more often as the average number of outcomes per study (Figure 2A), the strength of the correlation among outcomes in the same study (Figure 2B) and the residual between-studies variance (Figure 2D) increased. Rates for the three-level models were slightly over .05 in the conditions with small databases (either small number of studies or small number of outcomes per study). Last, results for the RVE method were consistently close to the nominal level when the MBBS estimator was implemented with at least 20 studies, whereas rates were again far below .05 with the jackknife estimator unless the number of studies was 80 (Figure 2C). We observed the same findings in the scenarios with an average of 500 participants per group (Suppl. Figure 2).

Confidence interval width

We now discuss how the different methods performed in terms of interval estimation of the slope coefficients across the simulated conditions. We only present conditions with $\beta_1 = 0.2$, although we found the same trends when the parameter value was $\beta_1 = 0$. Provided that the point estimation of the slope parameter was reasonably accurate for most methods (see bias section), an average width over 0.4 would suggest that the confidence intervals regularly included the null value ($\beta_1 = 0$), which would result in poor statistical power rates.

FIGURE 3 HERE

Figure 3 shows the mean width of the confidence intervals computed for the regression coefficient estimates across the various conditions for a study-level moderator with each of the meta-regression methods. The standard method ignoring dependency, as expected, yielded the narrowest intervals, followed by the three-level methods. The RVE methods (the jackknife estimator in particular) always resulted in the widest intervals. All methods showed narrower intervals in conditions where the number of studies was large and the between-studies heterogeneity was small (see Figures 3C and 3D). These figures show that both RVE methods yielded confidence intervals with an average width far above 0.4 for the conditions with less than 40 studies. In particular, the jackknife estimator yielded intervals with an average width of 8.7

with 5 studies. We observed the same trends in the scenarios with an average of 500 participants per group, although all methods yielded narrower intervals (Suppl. Figure 3).

FIGURE 4 HERE

For models with a moderator at the effect size level (Figure 4), all the meta-regression methods provided narrower confidence intervals than for a moderator at the study level, and the average width across the different conditions was always below 0.4. Remarkably, the three-level methods provided the narrowest intervals across most conditions, followed by the method ignoring dependency. The RVE methods (especially the jackknife estimator) consistently yielded the widest intervals. The narrowest intervals for all methods were obtained with larger numbers of studies (Figure 4C), more outcomes per study (Figure 4A), and smaller between-studies heterogeneity (Figure 4D). Furthermore, all methods yielded narrower intervals – although with the same trends – in the scenarios with an average of 500 participants per group (Suppl. Figure 4).

Statistical power

Last, we discuss the statistical power rates of the hypothesis tests. Rates for the different methods at the study and at the effect size level are displayed in Figure 5 and Figure 6, respectively. Note that we are only interested in power when the Type I error rate is adequately controlled. Therefore, in this section we focus on the methods accounting for dependency.

FIGURE 5 HERE

Regarding meta-regression models with a moderator at the study level, three-level methods consistently provided higher power rates than RVE methods, with the *z* test yielding higher rates than the likelihood ratio test across all scenarios, and the jackknife estimator always obtaining the lowest rejection rates. Statistical power was substantially higher for all methods in scenarios with a large number of studies (Figure 5C) and small between-studies heterogeneity (Figure 5D). Nonetheless, power rates for most methods were only around the desirable value of 0.8 with at least 40 studies, with the jackknife estimator showing adequate power rates only with 80 studies. These trends were also observed in the scenarios with an average of 500 participants per group, although the power rates were higher for all methods (Suppl. Figure 5).

FIGURE 6 HERE

Concerning models with a moderator at the effect size level, all methods yielded power rates higher than those observed at the study level. The three-level methods even outperformed the standard method in most scenarios, with the *z*-test reaching power rates over 0.8 with as few as 10 studies. The RVE methods showed again the lowest power values, only exceeding the threshold of 0.8 with 40 and 80 studies. Again, all methods showed higher power rates in conditions with a large number of studies (Figure 6C) and small heterogeneity between studies (Figure 6D), with the jackknife estimator being the most conservative method across all scenarios. The ranking remained the same, although with improved power rates for all methods, in the simulated scenarios with an average of 500 participants per group (Suppl. Figure 6).

Discussion

Dependency among effect sizes is a common situation in meta-analysis. The present study provides the first direct comparison of standard meta-analytic methods, which assume independent effect sizes, and RVE and three-level models that both account for dependency structures. We used Monte Carlo simulation to assess the accuracy of the estimation algorithms and statistical tests when fitting mixed-effects meta-regression models with dependency among effect sizes. Results did not yield any large difference in the estimates of the slope coefficients in the model, although some noticeable discrepancies were observed among the inferential results.

When examining the bias of the slope estimates, results for all estimation methods under most conditions showed a slight negative bias, although in most cases the percentage of bias was below 5% and hence all methods can be regarded as yielding nearly unbiased slope estimates from meta-regression models. This pattern was found both when the moderator was at the study and at the effect size level. For a few conditions, bias of the moderator effect at the study level slightly exceeded 5%, especially when the number of studies was small.

Some important differences were found among the methods when testing the statistical significance of the regression model slopes. Average empirical Type I error rates indicated a great number of overly liberal statistical conclusions, with the standard method always performing much too liberally – as would be expected because this method does not account for

the dependency structure among the effect sizes. The RVE methods showed an appropriate control of Type I error, with the MBBS estimator providing rates closer to the nominal significance level and the jackknife estimator yielding more conservative results, as pointed out by Tipton (2015). At the study level, the three-level model provided rates over the nominal with a small number of studies but yielded appropriate rejection rates as *k* increased, with the likelihood ratio test showing more accurate results than the Wald-type *z*-test. Results for three-level methods were more accurate when the moderator was at the effect size level, in particular with the likelihood ratio test, which performed close to nominally across all simulated scenarios with at least 10 studies.

Regarding confidence intervals around the slope estimates, wider intervals were consistently yielded by RVE methods than for the other approaches at both the study and the effect size levels. In scenarios with small number of studies or large heterogeneity variance at the study level, the average width yielded by the RVE approaches suggests that intervals obtained with these methods regularly included the null value, which would substantially limit their ability to detect a true relationship in such scenarios. The jackknife adjustment provided the widest intervals, which again suggests that this method is more conservative than the others and has lower power. An examination of the statistical power rates confirmed these shortcomings for RVE methods. At the study level, the narrowest confidence intervals were obtained with the standard method, although this method also showed highly inflated Type I error rates which discourage its use in this context. Remarkably, three-level methods yielded the narrowest intervals for most conditions at the effect size level. In general, narrower intervals were yielded by all methods when the number of studies was large and the between-studies variance was small. Moreover, all methods yielded narrower intervals when the moderator was at the effect size level than at the study level.

Limitations and usefulness of this study

This study was conducted with the aim of helping meta-analytic researchers deal with dependency in their data, namely when more than one effect size is available in several studies. Traditionally, meta-analytic researchers have applied the standard methods proposed by Hedges and Olkin (1985) for independent effect sizes by averaging all outcomes from the same cluster or

choosing only one of them, which may lead to a loss of relevant information (Becker, 2000). Other meta-analysts ignore dependency by analyzing a whole dataset as if the effect sizes were independent. In light of new methods that allow modeling of this dependency, we explored the performance of some of them under a wide range of conditions. Our data suggest that all methods can be expected to provide nearly unbiased estimates of the coefficients from a meta-regression model, but the method for testing the significance of the model moderators may have an important influence on the results.

Out of the different methods, the standard method ignoring dependency showed highly inflated Type I error rates across all simulated scenarios, suggesting that researchers should avoid using this method when they intend to implement meta-regression models with inferential purposes on a meta-analytic database with dependent effect sizes. The RVE approach showed an appropriate control of the Type I error rate, especially when correcting the residuals of the variance estimator and the degrees of freedom using the MBBS estimator. However, the confidence intervals around the slope estimate yielded by these methods were very wide in some scenarios, especially for study-level moderators, and the statistical power was lower than desirable unless the number of studies was at least 40. Note that some other correction factors have been suggested for this statistical test (see Cribari-Neto and Da Silva, 2011), so it will be important to assess their performance in future simulation studies. Moreover, the underlying regression model in the current study was constant, and the slope values were kept fixed, so future studies should examine the performance of the RVE method under different regression models and a wide range of slope values.

Regarding three-level models, the estimation algorithms using the maximum likelihood criterion can also be obtained using SAS. For moderators at the study level, the likelihood ratio test provided empirical Type I error rates closer to the nominal level than the *z*-test. However, both methods showed inflated rejection rates when the number of studies was small to moderate, and hence it would be interesting to explore the performance of other statistical tests for this approach in the future. Conversely, results at the effect size level suggested an appropriate performance for these methods with at least 10 studies, and particularly for the likelihood ratio test.

It is not possible to recommend a single approach across all scenarios, as both RVE and three-level methods have their merits. On the one hand, three-level models offer an interesting performance in terms of statistical power, although they need may yield too many false positives unless the number of studies is large enough, namely 20 studies (study level moderator) and 10 studies (effect size level moderator) for the likelihood ratio test. On the other hand, the RVE method with the MBBS estimator consistently controls the Type I error rate, although our results suggest that it might be underpowered with less than 40 studies. Nonetheless, our conclusions are limited to a single value of 0.2 for the slope, so that future studies should explore how the statistical power of these methods varies according to different slope values.

As a result of the algorithms employed to simulate our data, conditions with a higher clustering level also had a larger number of effect sizes. However, the influence of the number of studies was controlled in all analyses. This allowed us to assess whether the effect of an increment in the mean number of effect sizes per study was due to a higher level of dependency, or simply to a greater number of effect sizes. Results suggest that, *ceteris paribus*, the greater the dependency, the poorer the performance for the standard meta-analytic method ignoring dependency.

In summary, our results suggest that fitting mixed-effects meta-regression models when some amount of dependency among the effect sizes is present requires some method accounting for those dependency structures. The RVE method provided the best results in terms of control of the Type I error rate, in particular applying small-sample corrections with the bias reduced linearization estimator (MBBS, see Tanner-Smith and Tipton, 2014 for a tutorial on RVE using Stata and SPSS). Three-level models also showed a promising performance, especially with the likelihood ratio test, and yielded narrower confidence intervals around the slope than those obtained with the RVE approach, suggesting a gain in statistical power. Finally, our study suggests that more accurate results can be expected when the moderator included in the meta-regression model is at the effect size level, and when the meta-analytic database includes a moderate to large number of studies with small variability among effects from different studies.

Acknowledgements

This work was carried out using the computational facilities of the Advanced Computing Research Centre, University of Bristol - http://www.bris.ac.uk/acrc/. The authors are also grateful to Beth Tipton for providing the R code to implement the JKS corrections for the RVE approach.

Footnote

1. A widely accepted index to assess the degree of heterogeneity is I^2, which can be defined as $I^2 = 100\% \frac{\tau^2}{\tau^2 + v_i}$. In our simulations both $\overline{n}_i = 25$ and $d_i = 0.5$ were kept constant, so the expected value of v_i was 0.0825. Hence values of 0.08 and 0.32 for τ^2 would lead to expected I^2 values of 49% and 80%, which can be regarded as reflecting moderate and large heterogeneity, respectively (Deeks et al., 2008).

References

Ahn S, Myers ND, Jin Y 2012. Use of the estimated intraclass correlation for correcting differences in effect size by level. *Behavior Research Methods, 44:* 490-502.

Baker WL, White CM, Cappelleri JC, Kluger J, Coleman CI, Health Outcomes, Policy, and Economics (HOPE) Collaborative Group 2009. Understanding heterogeneity in meta-analysis: the role of meta-regression. *International Journal of Clinical Practice, 63:* 1426-1434.

Becker BJ 2000. Multivariate meta-analysis. In HEA Tinsley, SD Brown (Eds.). *Handbook of applied multivariate statistics and mathematical modeling* (pp. 499-525). San Diego, CA: Academic Press.

Bell RM, McCaffrey DF 2002. Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology, 28:* 169-181.

Beretvas SN, Pastor DA 2003. Using mixed-effects models in reliability generalization studies. *Educational and Psychological Measurement, 63*: 75-95.

Boker S, Neale M, Maes H, Wilde M, Spiegel M, Brick T, Spies J, Estabrook R, Kenny S, Bates T, Mehta P, Fox J 2011. OpenMx: An open source extended strucutral equation modeling framework. *Psychometrika, 76:* 306-317.

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR 2010. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1:* 97-111.

Cheung MWL 2015. metaSEM: An R package for meta-analysis using structural equation modeling. *Frontiers in Psychology, 5:* 1521.

Cheung MWL 2014. Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods, 19:* 211-229.

Cohen J 1988. *Statistical power analysis for the Behavioral Sciences* (2nd ed). Hillsdale, NJ: Eribaum.

Cribari-Neto F, Da Silva WB 2011. A new heteroskedasticity-consistent covariance matrix estimator for the linear regression model. *Advances in Statistical Analysis, 95:* 129-146.

Deeks JJ, Higgins JPT, Altman DG 2008. Analysing data and undertaking meta-analyses. In JPT Higgins, S Green (Eds.), *Cochrane handbook of systematic reviews of interventions* (pp. 243-296). Chichester: John Wiley & Sons.

DerSimonian R, Laird N 1986. Meta-analysis of clinical trials. *Clinical Controlled Trials*, *7*: 177-188.

Fisher Z, Tipton E 2014. "robumeta: Robust variance meta-regression. R package version 1.6". Available at: http://cran.r-project.org/web/packages/robumeta/index.html (Accessed 1 September 2016)

Gleser LJ, Olkin I 2009. Stochastically dependent effect sizes: Random-effects models. In H Cooper, LV Hedges, JC Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 357-376). New York: Russell Sage Foundation.

Hedges LV, Olkin I 1985. *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges LV, Tipton E, Johnson MC 2010. Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1:* 39-65.

Hedges LV, Vevea JL 1998. Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3:* 486-504.

Huizenga HM, Visser I, Dolan CV 2011. Testing overall and moderator effects in random effects meta-regression. *British Journal of Mathematical and Statistical Psychology, 64*: 1-19.

Hunter JE, Schmidt FL 2004. *Methods of meta-analysis: Correcting errors and bias in research findings* (2nd ed.). Newbury Park, CA: Sage.

Jackson D, Riley R, White IR 2011. Multivariate meta-analysis: Potential and promise. *Statistics in Medicine, 30:* 2481-2498.

Konstantopoulos S 2011. Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods, 2:* 61-76.

Lipsey MW 2009. Identifying interesting variables and analysis opportunities. In H Cooper, LV Hedges, JC Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 147-158). New York: Russell Sage Foundation.

Lipsey MW, Wilson DB 2001. *Practical meta-analysis*. Thousand Oaks, CA: Sage.

MacKinnon JG, White H 1985. Some heteroskedastity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics, 29:* 305-325.

Marín-Martínez F, Sánchez-Meca J 1999. Averaging dependent effect sizes in meta-analysis: A cautionary note about procedures. *The Spanish Journal of Psychology, 2:* 32-38.

McCaffrey DF, Bell RM, Botts CH 2001. Generalizations of biased reduced linearization. *Proceedings of the Annual Meeting of the American Statistical Association, August 5-9, 2001.*

Raudenbush SW 2009. Analyzing effect sizes: Random-effects models. In H Cooper, LV Hedges, JC Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 295-315). New York: Russell Sage Foundation.

Raudenbush SW, Bryk AS 2002. *Hierarchical linear models: Applications and data analysis methods*. (2nd ed.) London: Sage Publications.

Rosenthal R, Rubin DB 1986. Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin, 99:* 400-406.

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.

Sánchez-Meca J, Marín-Martínez F 2010. Meta-analysis. In P Peterson, E Baker, B McGaw (Eds.), *International Encyclopedia of Education* (3rd ed.), volume 7, pp. 274-282. Oxford: Elsevier.

Schmidt FL, Oh I-S, Hayes TL 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.

Sidik K, Jonkman JN 2005. A note on variance estimation in random effects meta-regression. *Journal of Biopharmaceutical Statistics*, *15:* 823-838.

Snijders T, Bosker R 1999. *Multilevel modeling: An introduction to basic and advanced multilevel modeling.* London: Sage Publications.

Stevens JR, Taylor AM 2009. Hierarchical dependence in meta-analysis. *Journal of Educational and Behavioral Statistics, 34:* 46-73.

Tanner-Smith EE, Tipton E 2014. Robust variance estimation with dependent effect sizes: practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods, 5:* 13-30.

Thompson SG, Higgins JPT 2002. How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine, 21:* 1559-1573.

Tipton E 2013. Robust variance estimation in meta-regression with binary dependent effects. *Research Synthesis Methods, 4:* 169-187.

Tipton E 2015. Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods, 20*: 375-393.

Van den Noortgate W, López-López JA, Marín-Martínez F, Sánchez-Meca J 2013. Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods, 45:* 576-594.

Van den Noortgate W, López-López JA, Marín-Martínez F, Sánchez-Meca J. 2015. Meta-analysis of multiple outcomes: A multilevel approach. *Behavior Research Methods, 47*: 1274-1294.

Van den Noortgate W, Onghena P 2003. Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational and Psychological Measurement, 63*: 765-790.

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

Viechtbauer W, López-López JA, Sánchez-Meca J, Marín-Martínez F 2015. A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychological Methods, 20:* 360-374.

Wilson SJ, Lipsey MW 2007. School-based interventions for aggressive and disruptive behavior: Update of a meta-analysis. *American Journal of Preventive Medicine*, *33*: S130-S143.

Wilson SJ, Lipsey MW, Derzon JH 2003. The effects of school-based intervention programs on aggressive and disruptive behavior: A meta-analysis. *Journal of Consulting and Clinical Psychology*, 2003, *71*: 136-149.

Table 1. Results from an Illustrative Example: A Meta-analysis of Intervention Programs for Reducing Aggressive Behavior among School-aged Youth

	Study level moderator			Effect size level moderator		
	$\hat{\beta}$	CI width	p-value	$\hat{\beta}$	CI width	p-value
Standard method	0.106	0.179	.020	0.109	0.177	.016
RVE method (MBBS)	0.224	0.549	.106	0.201	0.530	.134
RVE method (JKS)	0.224	0.573	.121	0.201	0.550	.148
Three-level method (z)	0.257	0.537	.060	0.155	0.233	.009
Three-level method (LB)	0.257	0.552	.062	0.155	0.234	.009

$\hat{\beta}$: slope estimate of the moderator variable; CI Width: width of the 95% confidence interval for the slope estimate of the moderator variable; p-value: p-value of the significance test for the slope estimate of the moderator variable.

Table 2. Mean bias for the slope estimates at the study level ($\beta_1 = 0.2$ and $\tau^2 = 0.08$)

		Low within-study correlation			High within-study correlation		
k	$\bar{a}$	STD	RVE	3LV	STD	RVE	3LV
5	2	-0.0069	-0.0093	-0.0085	**-0.0132**	**-0.0146**	**-0.0139**
	4	-0.0005	-0.0013	-0.0029	**0.0119**	0.0091	0.0089
	8	**-0.0107**	-0.0081	**-0.0103**	**-0.0301**	**-0.0246**	**-0.0258**
10	2	0.0062	0.0064	0.0061	0.0019	-0.0002	0.0001
	4	0.0049	0.0057	0.0034	-0.0058	-0.0033	-0.0036
	8	-0.0026	0.0005	-0.0019	-0.0055	-0.0032	-0.0028
20	2	-0.0077	-0.0063	-0.0071	-0.0048	-0.0035	-0.0030
	4	-0.0082	-0.0074	-0.0080	-0.0034	-0.0024	-0.0013
	8	-0.0039	-0.0007	-0.0018	-0.0084	-0.0074	-0.0062
40	2	-0.0035	-0.0028	-0.0033	-0.0032	-0.0021	-0.0014
	4	-0.0055	-0.0038	-0.0046	-0.0034	-0.0031	-0.0020
	8	-0.0041	-0.0024	-0.0037	-0.0063	-0.0049	-0.0040
80	2	-0.0053	-0.0041	-0.0046	-0.0036	-0.0044	-0.0037
	4	-0.0053	-0.0033	-0.0043	-0.0063	-0.0047	-0.0039
	8	-0.0051	-0.0035	-0.0045	-0.0053	-0.0040	-0.0027

k: number of studies; $\bar{a}$: average number of outcomes per study; STD: standard method ignoring dependency; RVE: robust variance estimation approach; 3LV: three-level model.

Table 3. Mean bias for the slope estimates at the effect size level ($\beta_1 = 0.2$ and $\tau^2 = 0.08$)

		Low within-study correlation			High within-study correlation		
k	$\bar{a}$	STD	RVE	3LV	STD	RVE	3LV
5	2	0.0047	0.0058	0.0037	0.0034	0.0050	0.0018
	4	-0.0032	0.0001	-0.0016	0.0022	0.0038	-0.0010
	8	-0.0042	0.0002	-0.0049	-0.0037	0.0007	-0.0026
10	2	-0.0026	-0.0032	-0.0013	-0.0032	-0.0016	-0.0018
	4	-0.0052	-0.0014	-0.0041	-0.0033	-0.0003	-0.0012
	8	-0.0025	0.0006	-0.0026	-0.0037	-0.0001	-0.0037
20	2	-0.0025	-0.0010	-0.0025	-0.0059	-0.0047	-0.0034
	4	-0.0031	-0.0005	-0.0032	-0.0004	0.0004	-0.0028
	8	-0.0050	-0.0018	-0.0033	-0.0073	-0.0035	-0.0037
40	2	-0.0049	-0.0032	-0.0040	-0.0056	-0.0037	-0.0038
	4	-0.0036	-0.0020	-0.0021	-0.0051	-0.0017	-0.0029
	8	-0.0052	-0.0017	-0.0034	-0.0081	-0.0047	-0.0047
80	2	-0.0047	-0.0030	-0.0036	-0.0032	-0.0026	-0.0030
	4	-0.0053	-0.0025	-0.0038	-0.0058	-0.0032	-0.0032
	8	-0.0046	-0.0012	-0.0030	-0.0074	-0.0035	-0.0043

k: number of studies; $\bar{a}$: average number of outcomes per study; STD: standard method ignoring dependency; RVE: robust variance estimation approach; 3LV: three-level model.