Behavior Research Methods (2021) 53:702–717
https://doi.org/10.3758/s13428-020-01459-4

Estimating outcome-specific effects in meta-analyses of multiple outcomes: A simulation study

Belén Fernández-Castilla[1,2] · Ariel M. Aloe[3] · Lies Declercq[1,2] · Laleh Jamshidi[1,2] · S. Natasha Beretvas[4] · Patrick Onghena[1] · Wim Van den Noortgate[1,2]

Published online: 17 August 2020

Abstract

In meta-analysis, primary studies often include multiple, dependent effect sizes. Several methods address this dependency, such as the multivariate approach, three-level models, and the robust variance estimation (RVE) method. As for today, most simulation studies that explore the performance of these methods have focused on the estimation of the overall effect size. However, researchers are sometimes interested in obtaining separate effect size estimates for different types of outcomes. A recent simulation study (Park & Beretvas, 2019) has compared the performance of the three-level approach and the RVE method in estimating outcome-specific effects when several effect sizes are reported for different types of outcomes within studies. The goal of this paper is to extend that study by incorporating additional simulation conditions and by exploring the performance of additional models, such as the multivariate model, a three-level model that specifies different study-effects for each type of outcome, a three-level model that specifies a common study-effect for all outcomes, and separate three-level models for each type of outcome. Additionally, we also tested whether the *a posteriori* application of the RV correction improves the standard error estimates and the 95% confidence intervals. Results show that the application of separate three-level models for each type of outcome is the only approach that consistently gives adequate standard error estimates. Also, the *a posteriori* application of the RV correction results in correct 95% confidence intervals in all models, even if they are misspecified, meaning that Type I error rate is adequate when the RV correction is implemented.

Keywords Meta-analysis · Multilevel · Multivariate · Robust variance estimation method · Outcome-effects

Meta-analysis is defined as the set of statistical tools that allow the combination of evidence from different studies to get a more detailed and general conclusion (Glass, 1976). In meta-analysis, effect sizes from studies that address the same research question are pooled together. There are in general two types of statistical models to combine effect sizes: the fixed-effect model and the random-effects model. The use of a fixed-effects model without moderators implies the assumption that there is a unique underlying overall effect, whereas the use of a random-effects model accounts for the possibility that each study represents its own population effect (and typically it is assumed that these follow a normal distribution). A random-effects model is expressed as:

$$d_j = \gamma_0 + e_j + u_j, \tag{1}$$

where d_j is the effect size reported in study j. This model assumes that each effect size deviates from the overall effect, γ_0, due to differences in samples, and due to the differences across studies. Specifically, there are two random effects, e_j and u_j. The commonly used maximum likelihood (ML) or restricted maximum likelihood (REML) estimation procedures assume that these random effects are normally distributed with mean 0 and variance $\sigma^2_{e_j}$ and σ^2_u, respectively. The variance $\sigma^2_{e_j}$ is the sampling variance of the observed effect size, which is typically estimated in advance and therefore assumed to be known in the meta-analysis itself. The term σ^2_u refers to the between-studies variance, or in other words, to the heterogeneity of effect sizes due to differences between

✉ Belén Fernández-Castilla
belen.fernandezcastilla@kuleuven.be; bfcastilla@gmail.com

[1] Faculty of Psychology and Educational Sciences, KU Leuven, University of Leuven, Etienne Sabbelaan 51, 8500 Kortrijk, Belgium

[2] ITEC, an Imec research group at KU Leuven, University of Leuven, Leuven, Belgium

[3] University of Iowa, Iowa City, IA, USA

[4] University of Texas at Austin, Austin, TX, USA

studies. If σ_u^2 equals 0, then the model of Eq. (1) reduces to a fixed-effect model. Note that this model is equivalent to a two-level model, where effect sizes vary due to between-study differences (level 2) and due to differences between random samples (level 1) (Raudenbush & Bryk, 1985). Nevertheless, the model above is not adequate if several effect sizes are reported within primary studies.

In applied research, it is quite common that, within a single study, researchers use several effect sizes or outcomes to measure a common construct, or that researchers compare several treatment groups with a common control group, or that they measure a target group in several follow-ups. Effect sizes reported in the same study are likely to be more similar, especially if they are based on the same sample. Ignoring this dependency by, for instance, fitting the random-effects model of Eq. (1), might lead to biased standard error estimates, that eventually may lead to inflated Type I error rates (Becker, 2000). Several methods exist for dealing with dependent effect sizes in meta-analyses, namely the multivariate approach (Kalaian & Raudenbush, 1996; Raudenbush, Becker, & Kalaian, 1988), the application of three- (or more) level models (Cheung, 2014; Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013, 2015), or the use of the robust variance estimation method for a random-effects model (RVE; Hedges, Tipton, & Johnson, 2010).

In this study, we describe and compare the application of these methods in situations where within studies, multiple types of outcomes may have been studied and for each type of outcome multiple effect sizes might have been observed, for instance because the same type of outcome has been measured with two or more different instruments, or because the sample is repeatedly measured in several follow-ups.

For instance, Spruit, Assink, Van Vugt, Van der Put, and Stams (2016) performed a meta-analysis on the relationship between physical activity interventions and internalizing behaviors. In this meta-analysis, the authors were interested in getting the overall effects of different types of internalizing behaviors, namely depression, anxiety, or other, and in some studies, several effect sizes referred to the same type of outcome. A similar example is found in the meta-analysis of Lebuda, Zabelina, and Karwowski (2016), where the relationship between mindfulness and creativity was tested. After getting an overall estimate of this relationship, authors also wanted to compare the overall estimates of different aspects of creativity, namely fluency, flexibility, originality, insight problem-solving skill, and composite divergent thinking. A last example is found in the meta-analysis of Owen et al. (2016), where the relationship between physical activity and school engagement was evaluated. Authors were also interested in getting separate overall estimates for different types of engagement: behavioral, cognitive, and emotional engagement.

The examples mentioned above are only a small sample of meta-analyses in which the interest is in estimating outcome-specific effects. However, many simulation studies that explore the performance of methods that deal with dependent effect sizes have focused on the estimation of an overall effect size (i.e., Hedges, et al., 2010; Lee, 2014; Moeyaert et al., 2017; Van den Noortgate et al., 2013, 2015), rather than in the estimation of outcome-specific effects. Therefore, the aim of this study to test which method for treating dependent effect sizes (i.e., multivariate approach, multilevel techniques, or RVE) is better for estimating outcome-specific pooled effect sizes and their standard errors. In the following sections, we describe different models that can be used to apply each of these methods to estimate outcome-specific pooled effects.

Model 1: Multivariate model

If primary studies include effect sizes referring to, let us say, three different outcomes, a multivariate two-level meta-regression model could be applied. At the first level, the model describes variation between effect sizes within studies:

$$d_{ij} = \beta_{1j}{*}Type\_1_{ij} + \beta_{2j}{*}Type\_2_{ij} + \beta_{3j}{*}Type\_3_{ij} + e_{ij}, \quad (2)$$

where d_{ij} refers to the observed effect size i reported in study j. These effect sizes are regressed on dummy indicators for three types of outcomes, Type_1, Type _2, and Type _3. Their weights, β_{1j}, β_{2j}, β_{3j}, refer to the population effect sizes for study j for these three types, respectively. Note that this model assumes that effect sizes belonging to the same type of outcome have the same population effect. The estimation procedure usually used in this model, namely ML or REML, assumes that the vector of residuals **e** within study j follows a multivariate normal distribution with mean 0 and with the following I x I variance-covariance matrix (**V**), being I the total number of effect sizes within a study:

$$\begin{bmatrix} e_{1j} \\ e_{2j} \\ \vdots \\ e_{Ij} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{e_1} & & & \\ \sigma_{e_1e_2} & \sigma^2_{e_2} & & \\ \vdots & \vdots & \ddots & \\ \sigma_{e_1e_I} & \sigma_{e_2e_I} & \dots & \sigma^2_{e_I} \end{bmatrix} \right).$$

At Level 2, the study-specific population effects can be allowed to randomly vary across studies:

$$\begin{cases} \beta_{1j} = \gamma_{10} + u_{1j} \\ \beta_{2j} = \gamma_{20} + u_{2j} \\ \beta_{3j} = \gamma_{30} + u_{3j} \end{cases} \quad (3)$$

where γ_{10}, γ_{20}, and γ_{30} are the outcome-specific mean effects. The study-specific random effects u_{1j}, u_{2j}, and u_{3j} are assumed to follow a multivariate distribution:

$$\begin{bmatrix} u_{1j} \\ u_{2j} \\ u_{3j} \end{bmatrix} \sim MVN \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u_1} & & \\ \sigma_{u_1u_2} & \sigma^2_{u_2} & \\ \sigma_{u_1u_3} & \sigma_{u_2u_3} & \sigma^2_{u_3} \end{bmatrix} \right),$$

where the variances $\sigma^2_{u_1}$, $\sigma^2_{u_2}$, and $\sigma^2_{u_3}$ are the between-studies variances of the population effect sizes for outcomes of Type_1 (γ_{10}) , Type_2 (γ_{20}), and Type_3 (γ_{30}), respectively. Note that the model allows that the between-study variance depends on the outcome type. For instance, in the study of Spruit et al. (2016), it could be the case that effect sizes for 'anxiety' outcome varied more across studies than the effect sizes for 'depression' outcome (but effects of both types of outcomes are still likely positively correlated).

The multivariate approach has the advantage of yielding separate pooled effect size estimates for each type of outcome, enabling the statistical comparison among them. However, a disadvantage of using this approach is that, in the same way as in univariate meta-analysis the sampling variances of the observed effect sizes should be estimated before doing the meta-analysis, a multivariate meta-analysis assumes that the sampling variance-covariance matrix can be estimated in advance. Unfortunately, primary studies often do not report enough information to estimate the covariances. For instance, if a given construct has been measured using different outcomes (within studies), then the correlation between these outcome variables would be necessary to calculate the covariance among effect sizes.

Model 2: Three-level model with one random study effect

A second approach to account for dependent effect sizes consists in the application of three-level models (Cheung, 2014; Van den Noortgate et al., 2013, 2015). Additional random effects can be added to the model of Eq. (1) to address dependency among effect sizes within studies. One possible model specification for the scenario that is being considered throughout this study (i.e., existence of multiple effect sizes within multiple types of outcomes), is the following:

$$d_{ij} = \gamma_{10} * Type\_1_{ij} + \gamma_{20} * Type\_2_{ij} + \gamma_{30} * Type\_3_{ij} + e_{ij} + r_{ij} + u_{0j}. \quad (4)$$

The variance of the three random effects, e_{ij}, r_{ij}, and u_{0j} refer, respectively, to the sampling variance of the observed effect sizes (Level 1; estimated in advance), the within-study variance between true effect sizes (Level 2), and the between-studies variance once the effect of the three dummy variables has been taken into account (Level 3).

Model 3: Three-level model with separate random study effect

In Model 2, there is only one random effect at the study level. This assumes that the effect of the study on the expected effect sizes is the same for all types of outcomes in the study, and that the between-studies variance (σ^2_u) is assumed to be the same for the three types of outcomes. However, we can also specify a three-level model with a random study effect for each type of outcome (as in Model 1):

$$d_{ij} = \left[(\gamma_{10} + u_{1j}) * Type\_1_{ij}\right] + \left[(\gamma_{20} + u_{2j}) * Type\_2_{ij}\right] + \left[(\gamma_{30} + u_{3j}) * Type\_3_{ij}\right] + e_{ij} + r_{ij}. \quad (5)$$

Now, at Level 3, there are three random effects: u_{1j}, u_{2j}, and u_{3j}, that are assumed normally distributed with mean 0 and variances $\sigma^2_{u_1}$, $\sigma^2_{u_2}$, and $\sigma^2_{u_3}$ if the estimation procedure used is ML or REML. Each of these variances refers to the between-studies variances of the three outcome-specific pooled effect sizes.

Model 4: Separate three-level models

Another strategy to carry out a meta-analysis that estimates pooled-specific outcome effects is to separately carry out a three-level meta-analysis for each different type of outcome. At this point it is important to recall that we are considering the scenario in which there are several effect sizes per type of outcome within studies, so even if each type of outcome's effect size is separately synthesized, there will be still several effect sizes within studies, and hence the inclusion of a random outcome effect is appropriate. For each type of outcome separately, the following three-level model is used:

$$d_{ij} = \gamma_0 + e_{ij} + r_{ij} + u_{0j}; \quad (6)$$

A main difference with respect to the previous three-level models is that in this case, the pooled outcome-effect estimates cannot be statistically compared.

Unlike the multivariate approach, meta-analytic three-level models (wrongly) assume that the sampling covariances among effect sizes within studies are zero. However, simulation studies have shown that meta-analytic three-level models are robust to this misspecification of the correlation structure due to the incorporation of an additional random effect, r_{ij} (Van den Noortgate et al., 2013, 2015). Three-level models are especially advantageous when information for calculating covariances among effect sizes is not available. On the other hand, the main downside is that the covariance among each pair of effect sizes is implicitly assumed to be the same (Van den Noortgate et al., 2013, 2015), and this might not be a very realistic assumption.

Model 5: Two-level model (with the RVE method)

As mentioned earlier, the use of ordinary two-level models (without a random outcome effect) may result in biased standard errors and therefore in flawed inferences. The RVE method (Hedges et al., 2010; Tipton, 2013, 2015) corrects the standard error(s) of the fixed effect(s) estimates (i.e., pooled effect size and moderator effects) using sandwich variance estimators. The general meta-regression can be written as:

$$\mathbf{d} = \mathbf{X}\gamma + \boldsymbol{\varepsilon}, \tag{7}$$

where $\mathbf{d}$ is a vector of stacked observed effect sizes across studies, $\mathbf{X}$ is a design matrix whose number of rows equal the total number of effect sizes, and whose number of columns equal the number of covariates tested in the meta-regression. When no covariates are tested, then $\mathbf{X}$ equals a column vector of ones. The term γ refers to the vector of regression coefficients that have to be estimated, and finally $\boldsymbol{\varepsilon}$ is a vector of residuals. To get an estimate of γ, this equation is used:

$$\widehat{\gamma} = (\mathbf{XWX})^{-1}(\mathbf{XWd}), \tag{8}$$

where $\mathbf{W}$ is a diagonal matrix with the weights assigned to each effect size along the diagonal.

The meta-regression model of Eq. (1) is the same for both a fixed- and a random-effects model. The difference is in the weights assigned to the effect sizes under each of these models. Under a fixed-effects model, the weights of the effect sizes belonging to study j equal $\frac{1}{k_j(\overline{v}_{.j})}$, where k refers to the number of effect sizes within study j, and $\overline{v}_{.j}$ is the average sampling variance of the k effect sizes in study j. Under a random-effects model, the weights equal $\frac{1}{k_j(\overline{v}_{.j}+\widehat{\tau}^2)}$, where $\widehat{\tau}^2$ is an estimate of the between-study variance.

In order to estimate the variances of the estimated pooled effects ($\widehat{\gamma}$), Hedges et al. (2010) proposed to use the cross-products of the within-study residuals as a rough estimate of the covariance matrix for observed effect size estimates within studies. Their study shows that with a large number of studies, this rough estimation of the covariance matrix leads to unbiased standard error estimates.

$$\mathbf{V}\left(\widehat{\gamma}\right) = \left(\sum_{j=1}^{k}\mathbf{X}_j'\mathbf{W}_j\mathbf{X}_j\right)^{-1}\left(\sum_{j=1}^{k}\mathbf{X}_j'\mathbf{W}_j\boldsymbol{\varepsilon}_j\boldsymbol{\varepsilon}_j'\mathbf{W}_j\mathbf{X}_j\right)\left(\sum_{j=1}^{k}\mathbf{X}_j'\mathbf{W}_j\mathbf{X}_j\right)^{-1} \tag{9}$$

The term $\boldsymbol{\varepsilon}_j$ refers to a vector of the study residuals in study j, and $\boldsymbol{\varepsilon}_j = \mathbf{d}_j - \mathbf{X}_j\widehat{\boldsymbol{\gamma}}$. Note that $\boldsymbol{\varepsilon}_j$ will have a different value under a fixed- and under a random-effects model because the estimate of $\widehat{\gamma}$ will be different due to the use of different weights for its estimation. For calculating the between-study variance, we refer to the original paper, but an important aspect of its calculation is that the value of the correlation between effect sizes within the same study is needed. However, Hedges et al. (2010) showed that the value of the correlation selected has little effect on the parameter and standard error estimates. Therefore, this approach can be applied even if the researcher does not have information about the correlation between the effect size estimates. Another advantage is that RVE does not make strict assumptions about the distribution of the data.

Current study

In the present study, we will study which of these five models will result in the best outcome-specific effect estimates, their standard errors and their 95% confidence intervals (CIs). In addition, we will explore whether the performance when using the first four models can be further improved by using an *a posteriori* robust variance correction to the standard errors, as proposed recently (Pustejovsky, Tipton, & Aloe, 2018; Tipton, Pustejovsky, & Ahmadi, 2019). This *a posteriori* robust variance correction consists in applying the reduced-linearization correction proposed by Bell and McCaffrey (2002) to the observed variance-covariance matrix of a set of regression coefficients. More technical information about this adjustment can be found in Pustejovsky & Tipton (2017).

Park and Beretvas (2019) have recently performed a simulation study in which they have compared the performance of a meta-analytic three-level meta-regression (that specified different study-effects for each type of outcome) with the performance of a random-effects model with RVE method regarding the parameter recovery of the outcome-specific effects, standard errors, and between-studies variances. They concluded that a random-effects model using RVE method accurately estimated outcome-specific pooled effect sizes and their standard errors. However, this method sometimes underestimated the between-studies variance, especially when the true between-studies variance and the correlations between study residuals were large. On the other hand, the three-level model did not converge under some circumstances, especially when the number of primary studies was 20, the between-studies variance was small, and when datasets were unbalanced. Both methods performed equally when the number of primary studies was 50.

The present simulation study aims to extend the study of Park and Beretvas (2019) in three important ways. First, Park and Beretvas (2019) only tested the performance of the multilevel approach (estimating separate between-studies variances for each outcome-specific effect) and of the random-effects model using RVE method, whereas in this study we will additionally explore the performance of the multivariate approach and of alternative specifications of the three-level

model that are more widely used in practice (e.g., a three-level model that assumes a common study-effect for all effect sizes or separate three-level models for each type of outcome). Additionally, this study will shed light on the value of correcting standard errors *a posteriori* (Pustejovsky, Tipton, & Aloe, 2018; Tipton, Pustejovsky, & Ahmadi, 2019)

A second difference between the study of Park and Beretvas (2019) and this study is the way data are generated. They generated data from a three-level model, and we will generate data from a multivariate two-level model, with a sample-level and a between-study level (Kalaian & Raudenbush, 1996; Raudenbush, et al., 1988). Previous simulation studies have followed this strategy of generating multivariate meta-analytic data and then analyzing them using three-level models (e.g., Van den Noortgate et al., 2013, 2015), because the multivariate two-level model is generally assumed to be the actual correct model in meta-analysis (i.e., correlated effect sizes are nested within studies).

Finally, a third difference is that Park and Beretvas (2019) generated between one and 15 effect sizes per study (five effect sizes per type of outcome), but it was equally likely that a study included eight effect sizes than that a study reported 12 effect sizes. However, a recent systematic review (Fernández-Castilla et al., 2020) has shown that an important percentage of primary studies in the field of behavioral and social sciences (42.27%) include only one effect size, whereas 18% include five effects or more. In the present simulation, we will base the data generation on this more authentic pattern that was found. Furthermore, we will add another simulation factor condition in which the between-studies variances are different for the three types of outcomes (i.e., effect sizes belonging to one type of outcome vary more across studies than effect sizes belonging to another type of outcome). If the true between-studies variance of each level of the moderator (i.e., of each type of outcome) is different, but a common variance is assumed (Model 2 and Model 5), we expect to get biased standard error estimates.

In this study, we will mainly focus on the overall type-of-outcomes estimates and on their standard errors, because applied researchers are often mostly interested in testing and comparing outcome effects. We will only summarize general results on the variance components; more information can be found in supplementary material.

Method

Data generation First, Cohen's *d* were directly simulated from the multivariate two-level model described above (Model 1), with a separate dependent variable for each effect size. We have chosen to draw effect sizes directly to make our results more easily generalizable to other effect size measures with approximately normal sampling distributions (such as Fisher's z transformed correlations and log odds ratios). A drawback is, however, that this is less good mimicking real meta-analyses in which effect sizes are calculated on raw data. At the first level (Eq. 2), observed effect sizes are regressed on dummy indicators for three types of outcomes, Type_1, Type _2, and Type _3, using a model without intercept. Effect sizes belonging to the same type of outcome had the same population effect. The sampling variances-covariances were calculated with the formulas of Gleser and Olkin (1994) for Cohen's *d* effects[1].

Initially, 30 effect sizes per study were generated ($I = 30$). For each study, the first ten effect sizes referred to the type of outcome one (Type_1), the next ten effect sizes referred to the type of outcome two (Type_2), and the last ten effect sizes referred to the type of outcome three (Type_3). Afterwards, most of these effect sizes were removed in order to generate two types of balanced/unbalanced scenarios and to still keep the number of effect sizes per study fixed. These two balanced/unbalanced scenarios are explained below.

In practice, it is unlikely that effect sizes are equally distributed over types of outcome. For instance, in the meta-analysis of Spruit et al., (2016), depression outcome effect sizes were more numerous ($i = 38$) than effect sizes for anxiety ($i = 26$) or for the other outcome category ($i = 12$). We accounted for this scenario in the 'balanced- versus 'unbalanced-type of outcome' condition. In the 'balanced type of outcome condition', there were the same number of effect sizes per type of outcome: 67% of the effect sizes referring to outcome Type_1, 67% of the effect sizes referring to outcome Type_2, and 67% of the effect sizes referring to outcome Type_3 were randomly removed. In the 'unbalanced type of outcome' condition, 40% of the effect sizes referring to outcome Type_1, 70% of the effect sizes referring to outcome Type_2, and 90% of the effect sizes referring to outcome Type_3 were deleted. In both the balanced and unbalanced conditions, the expected number of effect sizes within studies after the deletion was ten.

Furthermore, it is also unlikely that all studies report the same number of effect sizes. Therefore, after generating the '(un)balanced type of outcome' condition, we generated an additional condition called 'balanced-' or 'unbalanced effect sizes' across studies condition. A recent systematic review about characteristics of meta-analyses of multiple outcomes in different research fields (Fernández-Castilla et al., 2020) has found that in the field of behavioral and social sciences, 42.27% of primary studies report only one effect size, 22.87% of studies report two effect sizes, 7.88% of studies provide three effect sizes, 8% of studies report four effects, and 18% report five or more effect sizes. Therefore, in the 'unbalanced

[1] $\sigma_{d_{ij}\,d_{ij'}} = \frac{n_c+n_e}{n_c} n_e + \delta_{ij} \frac{\delta_{ij'}\rho^2_{ij\,ij'}}{2\,(n_c+n_e)}$. n_c and n_e refers to the sample size of the experimental and control group. δ_{ij} and $\delta_{ij'}$ refer to the population effect of effect size i within study j, and $\rho_{ij\,ij'}$ refers to the correlation between outcome variables.

effect sizes' condition, we replicated these percentages. On the other hand, in the 'balanced effect sizes' condition, we randomly chose two or three effect sizes per study. The expected number of effect sizes in the meta-analysis was the same in both balanced and unbalanced condition: 56 effect sizes when the number of studies generated was 20, and 120 effect sizes when the number of studies generated was 40.

Data analysis After generating these data, effect sizes were analyzed according to the five models described above. For the multivariate approach (using Model 1), the variance-covariance matrix used to fit the multivariate approach was calculated using the population correlation between outcome variables and the population effects that were used during the data generation, meaning that it was assumed that the correlation between outcome variables was accurately known for each study. Of course, these correlations are not known in practice and should be estimated (which is exactly the main drawback of the multivariate approach). By proceeding in this way, however, we can use the multivariate approach as a benchmark, because the model used is the correct model and because no bias is induced by using estimated correlation coefficients. For the RVE approach (using Model 5), we used the small sample adjustment proposed by Tipton (2015), i.e., using the Satterthwaite approximation for the degrees of freedom for the *t*-statistic used to build CIs are obtained using (the default option in the software we used).

The estimation procedure used for Model 1 to 4 was the restricted maximum-likelihood (REML). As mentioned earlier, we first applied Model 1 to 4 using the REML estimation method, and then we applied the robust variance correction. As a result, we obtained nine different sets of estimates from five models: four sets of estimates (from Model 1 to Model 4) in which the standard errors of the fixed effects were not corrected, four sets of estimates (from Model 1 to Model 4) in which the standard errors of the fixed effects were afterwards corrected using RVE method, and then one set of estimates where only RVE was applied (Model 5).

Conditions

Several characteristics of the datasets were systematically varied besides the ones already described (i.e., the (un)balanced type of outcome condition and the (un)balanced effect sizes condition). According to the systematic review of Fernández-Castilla et al. (2020), the median number of studies (*k*) included in meta-analyses in the field of behavioral and social sciences is 39, and the first quartile is 23. Therefore, we chose two conditions for the number of studies: 20 and 40.

The same systematic review indicates that the median sample size of primary studies (*n*) in the same field is 107, and the first quartile is 50. Because the sample size is typically unbalanced across primary studies, for each primary study we extracted a number from a lognormal distribution $n \sim LnN$ (3.91, 0.7) or *LnN* (4.61, 0.7). The mean sample size in each meta-analysis was therefore 50 (when the *LnN* mean was 3.91) or 100 (when the *LnN* mean was 4.61), mimicking the results of the systematic review. Because the covariance between effect sizes might not be accurately estimated if the sample size is smaller than 200, we included another condition in which the expected average sample size was 280 (this number corresponds to the 3rd quartile of the same systematic review). The distribution from which the sample size value was extracted was *LnN* (5.63, 0.7).

Three values for the correlation between outcome variables within primary studies were selected. In one condition, the variables that represented the same outcome correlated .4 and the variables representing different outcomes correlated .2. In a second condition, both correlations were higher: the correlation between variables belonging to the same outcome was .6, and the correlation between variables belonging to different outcomes was .4. In a third condition, the variables were uncorrelated ($\rho = 0$). This last condition was included to get a better insight in how results depend on the correlation, but it is unlikely that outcomes that are supposed to belong to the same category are uncorrelated.

Regarding the between-studies variances ($\sigma^2_{u_1}, \sigma^2_{u_2}, \sigma^2_{u_3}$), three different conditions were simulated. In the first one, the three variances had the same value, 0.072. This value corresponds to the median between-studies variance obtained in the systematic review of Fernández-Castilla et al. (2020) for Hedges' *g* in the field of behavioral and social sciences. In a second condition, the variance of the first outcome ($\sigma^2_{u_1}$) had a value of 0.017, corresponding to the first quartile found in the aforementioned systematic review, the between-studies variance of the second outcome ($\sigma^2_{u_2}$) was set to 0.072, and the between-studies variance of the third outcome ($\sigma^2_{u_3}$) equaled 0.153, that corresponds to the third quartile. In a third condition, the between-studies variances of the first, second, and third outcome were set to 0.153, 0.072, and 0.017, respectively. The correlation selected to calculate the covariances between study residuals ($\sigma_{u_1}\, u_2$, $\sigma_{u_1}\, u_3$, and $\sigma_{u_2}\, u_3$) were 0, .2, or .4, which are intermediate values from the ones used by Park and Beretvas (2019). Finally, the mean outcome effects were chosen to be $\gamma_{10} = 0.20$, $\gamma_{20} = 0.40$, and $\gamma_{30} = 0.60$, that are the values also selected by Park and Beretvas (2019).

All these factors resulted in 2 (balanced/unbalanced type of outcome condition) * 2 (balanced/unbalanced effect sizes condition) * 2 (number of studies) * 3 (mean sample size) * 3 (correlation between variables) * 3 (between-studies variance values) * 3 (correlation between study residuals) = 648 conditions. For each condition, 1,000 datasets were generated. The simulated factor conditions are summarized in Table 1.

R software was used to generate and analyze the data (R Core Team, 2012). To fit the multivariate and multilevel meta-

analytic models (Model 1 to 4), package *metafor* (Viechtbauer, 2010) was used, specifically the *rma.mv* function. To fit the random-effects model with RVE as estimation method (Model 5), the function *robu* from package *robumeta* was used (Fisher, Tipton & Zhipeng, 2017). To correct the fixed parameters with RVE method, the package *clubSandwich* (Pustejovsky, 2018) was applied, specifically the *coef_test* function. The R code used to fit the five different models plus a generated dataset to run the code can be found in the following link: https://osf.io/ywum6/

Evaluation of the approaches

The fixed effect estimates were summarized across the 1000 iterations. The bias of fixed effect estimates was calculated by subtracting the real value from the average estimated value. By dividing this bias by the true value (and multiplying it by 100), we obtained the relative bias (RB), expressed as a percentage. Following the cutoffs proposed by Hoogland and Boomsma (1998), RBs between 5% and 10% were considered as acceptable, whereas RBs above 10% were considered as unacceptable. Furthermore, the accuracy of the fixed-effect estimates was evaluated by calculating the mean square error (MSE).

Bias of the standard errors of the fixed effect estimates was approximated by comparing the median of the standard error estimates of a certain condition with the standard deviation of the estimates in that condition. The standard deviation of the fixed-effects estimates can be considered as an accurate approximation of the true standard error because we simulated a large number of datasets (i.e., 1000). We decided to use the median standard error and not the mean because of the known skewed distribution of standard errors. We also looked at the coverage percentages of the 95% confidence intervals (CI) in each condition.

The non-converge rates were calculated for each model separately. To get this percentage, the number of datasets where the model did not converge was divided by the total number of datasets (i.e., 648,000), and then this number was multiplied by 100.

Last, in order to find out which conditions were related to bias, we performed analyses of variance (ANOVAs) where the dependent variable was the fixed effect or standard error bias, and the independent variables were the simulation factor conditions. Simulation factors with an eta squared (η^2) larger than .14 were considered influential. This cutoff is based on Cohen's rule of thumb (Cohen, 1988), where .14 can be considered as a large effect. Only main effects were tested, with the exception of the interaction between the factor condition (un)balanced-type of outcome and the pattern of between-studies variances for each outcome (i.e., whether between-studies variances were the same or different across outcomes). The purpose of including this interaction is to test the extent to which the standard errors are affected when the most reported outcome (i.e., outcome 1 in the unbalanced-type of outcome condition) is the one with a largest between-studies variance (i.e., condition where $\sigma^2_{u_1}$ = 0.153, $\sigma^2_{u_2}$ = 0.072, and $\sigma^2_{u_3}$ = 0.017) or the one with the lowest between-studies variance (i.e., condition where $\sigma^2_{u_1}$ = 0.017, $\sigma^2_{u_2}$ = 0.072, and $\sigma^2_{u_3}$ = 0.153).

Results

An important first result is that outcome-specific pooled estimates were unbiased in all models and across conditions, so they are not further discussed. In contrast, the MSEs, standard errors estimates, and coverage proportion of 95% CIs differed across conditions and models. The MSE is smallest using the multivariate model (Model 1), followed by the two-level model with RVE (Model 5), the use of separate three-level models for each type of outcome (Model 4), three-level models with separate random study-effects (Model 3) and finally three-level models with one random study effect (Model 2) (see Table2). In the following section, we first discuss the non-convergence rates, and then the standard error estimates and coverage proportion of 95% CIs of the outcome-specific effects.

Non-convergence rates

The multivariate model (Model 1) and the three-level model that specified different study-effects for each type of outcome (Model 3) did not converge in 0.12% of the analyses. For Model 1, the conditions most associated with the non-convergence rates were small mean sample sizes and balanced effect sizes within types of outcomes. For Model 3, the conditions most associated to the non-convergence rates were when the number of primary studies was small and effect sizes within types of outcomes were unbalanced. The three-level model that specified a common study-effect for all effect sizes (Model 2) did not converge in 0.005% of the conditions. When three different three-level models were applied for each type of outcome separately (Model 4), there was no convergence in 0.10%, 0.08% and 0.34 % for the analyses of the three types, respectively. The factor levels related to the non-convergence rates of these models were small mean sample sizes, small number of primary studies, and balanced effect sizes within types of outcomes. Finally, the random-effects model (Model 5) estimated using the RVE method always converged.

Table 1 Summary of the simulated factor conditions

Factor (number of conditions)	Simulation conditions
Number of studies (2)	k = 20, 40
Mean sample size (3)	n = 50, 100, 280
Within-study correlation among outcome variables (2)	ρ = 0 ρ = .4 if ESs belong to the same outcome and .2 if they belong to different outcomes; ρ = .6 if ESs belong to the same outcome and .4 if they belong to different outcomes
Between-studies variances (3)	$\sigma^2_{u_1} = \sigma^2_{u_2} = \sigma^2_{u_3} = 0.072$; $\sigma^2_{u_1} = 0.017, \sigma^2_{u_2} = 0.072, \sigma^2_{u_3} = 0.153$; $\sigma^2_{u_1} = 0.153, \sigma^2_{u_2} = 0.072, \sigma^2_{u_3} = 0.017$
Correlation between study random effects (3)	$\rho_{u_1\ u_2} = \rho_{u_1\ u_3} = \rho_{u_2\ u_3} = 0, 0.2, 0.4$
(Un)balanced type of outcome (2)	Unbalanced: The probability of reporting outcome Type_1, Type_2 and Type_ 3 was .60, .30 and .10, respectively. Balanced: The probability of reporting outcome Type_1, Type_2 and Type_3 was .33, .33 and .33, respectively.
(Un)balanced ESs (2)	Unbalanced: Studies reported different number of ESs. Balanced: Studies reported the same number of ESs.
Mean effects (1)	$\gamma_{10} = 0.20$, $\gamma_{20} = 0.40$, $\gamma_{30} = 0.60$

Notes. ESs = effect sizes

Standard errors and coverage proportion of the 95% CIs of the fixed effects estimates

Before analyzing in detail the standard error estimates yielded by each model, we first calculated the MSE and mean relative standard error bias (RSEB) for each type of outcome and model (see Table 2). Figure 1 shows the mean standard error bias of type of outcome 1 for all models disaggregated by each simulation factor condition[2].

The multivariate approach resulted in the lowest MSEs, followed by the random-effects model with RVE method. Model 2 (three-level model with one random study effect) led to less precise fixed effect estimates. The lowest RSEBs were obtained when using Model 4 (separate three-level models; RSEBs = – 1.62%, – 3.20%, and – 5.30% for $\widehat{\gamma}_{10}$, $\widehat{\gamma}_{20}$ and $\widehat{\gamma}_{30}$, respectively). In contrast, the largest RSEBs were obtained when Model 2 was applied (RSEBs = – 15.83%, – 20.21%, and – 19.30% for $\widehat{\gamma}_{10}$, $\widehat{\gamma}_{20}$, and $\widehat{\gamma}_{30}$, respectively). On average, the robust variance correction led to slightly more underestimated standard errors in all models, except for Model 2 (three-level model with one random study effect), where the robust variance correction substantially improved estimation of the standard errors. Interestingly, despite average standard errors were still underestimated by most models even if the robust variance correction was applied, the coverage proportion of the 95% CIs were still adequate. In contrast, the coverage proportion of the 95% CIs obtained in the models where RVE correction was not applied were too low, aligning with the results observed for the RSEB.

Figure 1 shows that Model 4 (separate three-level models) led to smaller standard error bias in all conditions. On the other hand, the three-level model that specified a common study-effect for all effect sizes showed the largest standard error biases in all conditions. In these graphs, we can also observe two strong interactions between the model fitted and the simulation factor levels. First, the robust variance correction led to better standard error estimates of the three-level model that specified a common study-effect for all outcomes, but this correction did not seem to improve estimates for the other models. Second, the same three-level model performed much worse than the other models in estimating the standard error for the first outcome type when its variance was very large, and also when all between-studies variances were equal to 0.072. In the following sections, the performance of each method is described in detail.

Model 1: Multivariate model

The simulation factors that had a larger influence in the standard error bias were the number of studies (η^2 = .344, .705, .360 for SE[$\widehat{\gamma}_1$], SE[$\widehat{\gamma}_2$], and SE[$\widehat{\gamma}_3$], respectively), whether the moderator variable 'type of outcome' was (un)balance across studies (η^2 = .170, .015, .273), and whether effect sizes were (un)balanced across studies (η^2= .106, .139, .040). Table 3 shows the RSEBs under this combination of conditions. When the number of primary studies was 20, all standard errors were underestimated, except the standard error of $\widehat{\gamma}_1$, that was properly estimated when the probability of reporting the type of outcome 1 was higher than the one of

[2] Figure 1 only includes the standard error bias of outcome Type_1 for simplicity. The graphs for the standard error bias of Type_2 and Type_3 are available at https://osf.io/ywum6/. Please note that in the graph for the standard error bias of Type_3, the scale of the *y*-axis is slightly larger.

Table 2 Mean squared error of fixed effects, relative standard error bias by model, and percentage of conditions in which each model led to unacceptable relative standard error bias

	Mean square error x 1000			Mean % relative bias			Coverage proportion of the 96% CIs		
	$\widehat{\gamma}_1$	$\widehat{\gamma}_2$	$\widehat{\gamma}_3$	SE($\widehat{\gamma}_1$)	SE($\widehat{\gamma}_2$)	SE($\widehat{\gamma}_3$)	$\widehat{\gamma}_1$	$\widehat{\gamma}_2$	$\widehat{\gamma}_3$
Model 1	.014	.040	.040	– 2.89	– 9.02	– 11.87	.93	.92	.90
Model 1+ RVc	.014	.040	.040	– 4.22	– 9.60	– 17.44	.94	.93	.92
Model 2	.039	.113	.114	– 15.84	– 20.21	– 19.30	.88	.87	.86
Model 2 + RVc	.039	.113	.114	– 5.04	– 7.86	– 8.35	.94	.94	.94
Model 3	.032	.103	.103	– 3.28	– 8.86	– 12.36	.93	.91	.90
Model 3 + RVc	.032	.103	.103	– 4.52	– 9.62	– 17.47	.94	.93	.92
Model 4	.030	.101	.101	– 1.62	– 3.19	– 5.27	.93	.93	.92
Model 4 + RVc	.030	.101	.101	– 1.98	– 3.20	– 5.77	.95	.95	.95
Model 5	.026	.074	.074	– 4.66	– 6.23	– 10.61	.95	.95	.04

Notes. Model 1: Multivariate model; Model 2: Three-level model with one random study effect; Model 3: Three-level model with separate random study effect; Model 4: Separate three-level models; Model 5: Two-level model (with the RVE method); RSEBs: Relative standard error bias; RVE: robust variance estimation method; RVc = robust variance correction; SE($\widehat{\gamma}_1$), SE($\widehat{\gamma}_2$), and SE($\widehat{\gamma}_3$) = standard error of the outcome Type_1, Type_2, and Type_3 estimates, respectively. CIs confidence intervals. Mean squared errors have been multiplied by 1000

reporting type of outcome 2 or type of outcome 3. In contrast, almost all standard errors showed acceptable RSEB when the number of primary studies was 40, except the standard error of $\widehat{\gamma}_1$, that was underestimated when types of outcomes were balanced and effect sizes were unbalanced across studies, and the standard error of $\widehat{\gamma}_3$, that was underestimated when the types of outcomes and the effect sizes across studies were unbalanced. The standard error of $\widehat{\gamma}_3$ was especially underestimated when the probability of reporting type of outcome 3 was smaller than the probability of reporting type of outcome 2 or type of outcome 1. The coverage proportion of the 95% CIs followed the same pattern as the RSEBs, being smaller than 95% in most conditions.

Model 2: Three-level model with one random study effect

The simulation factor conditions that predicted the bias of the standard errors estimated under this model were: the application of robust variance correction (η^2=.119, .370, and .065 for SE[$\widehat{\gamma}_1$], SE[$\widehat{\gamma}_2$], and SE[$\widehat{\gamma}_3$], respectively), the number of primary studies (η^2=.024, .148, and .092), whether effect sizes were (un)balanced across studies (η^2=.038, .188, and .038), and whether the variable types of outcome was (un)balanced across studies (η^2=.017, .027, and .086). Whether between-studies variances were the same or different for the three types of outcomes did not emerge as an influential factor condition, although theoretically it should have had an effect on the standard error estimates. Therefore, we have also included this factor in Table 4, where the RSEBs for each combination of conditions are reported. The other factors that did not result influential are fixed to intermediate levels.

As can be seen in Table 4, The RSEBs were out of the recommended bounds in almost all conditions when the robust variance correction was not applied. When the three between-studies variances were equal to 0.072, all standard errors were underestimated. When the between-studies variance of $\widehat{\gamma}_1$ was small, the standard error of $\widehat{\gamma}_1$ was overestimated, especially when the types of outcomes and effect sizes were balanced across studies. The same pattern was found for $\widehat{\gamma}_3$: its standard error was overestimated when its between-studies variance was small and the variable 'type of outcomes' and effect sizes were balanced across studies. The standard error of $\widehat{\gamma}_2$ was always underestimated.

The RSEBs substantially improved when the robust variance correction was applied. When the number of primary studies was 20, standard errors were still underestimated, especially the ones of $\widehat{\gamma}_3$, although in a lesser extent. When the number of studies was 40, all standard errors were appropriately estimated. Furthermore, the coverage proportion of the 95% CIs were very close to the nominal levels despite standard errors were sometimes underestimated. In contrast, when the robust variance correction was not applied, the coverage proportions were too small in the conditions where the RSEBs were outside the recommended bounds.

Model 3: Three-level model with separate random study effects

This model yielded almost identical results to the ones of the multivariate model (see Table 3). The factors that influenced the standard error bias in a larger extent were the number of studies (η^2= .352, .692, and .356 for SE[$\widehat{\gamma}_1$], SE[$\widehat{\gamma}_2$], and SE[$\widehat{\gamma}_3$], respectively), whether the variable 'types of outcome' was (un)balance across studies (η^2= .169, .014, and .270), and whether effect sizes were (un)balanced across studies (η^2= .124, .157, and .047). The RSEB were almost the same as that

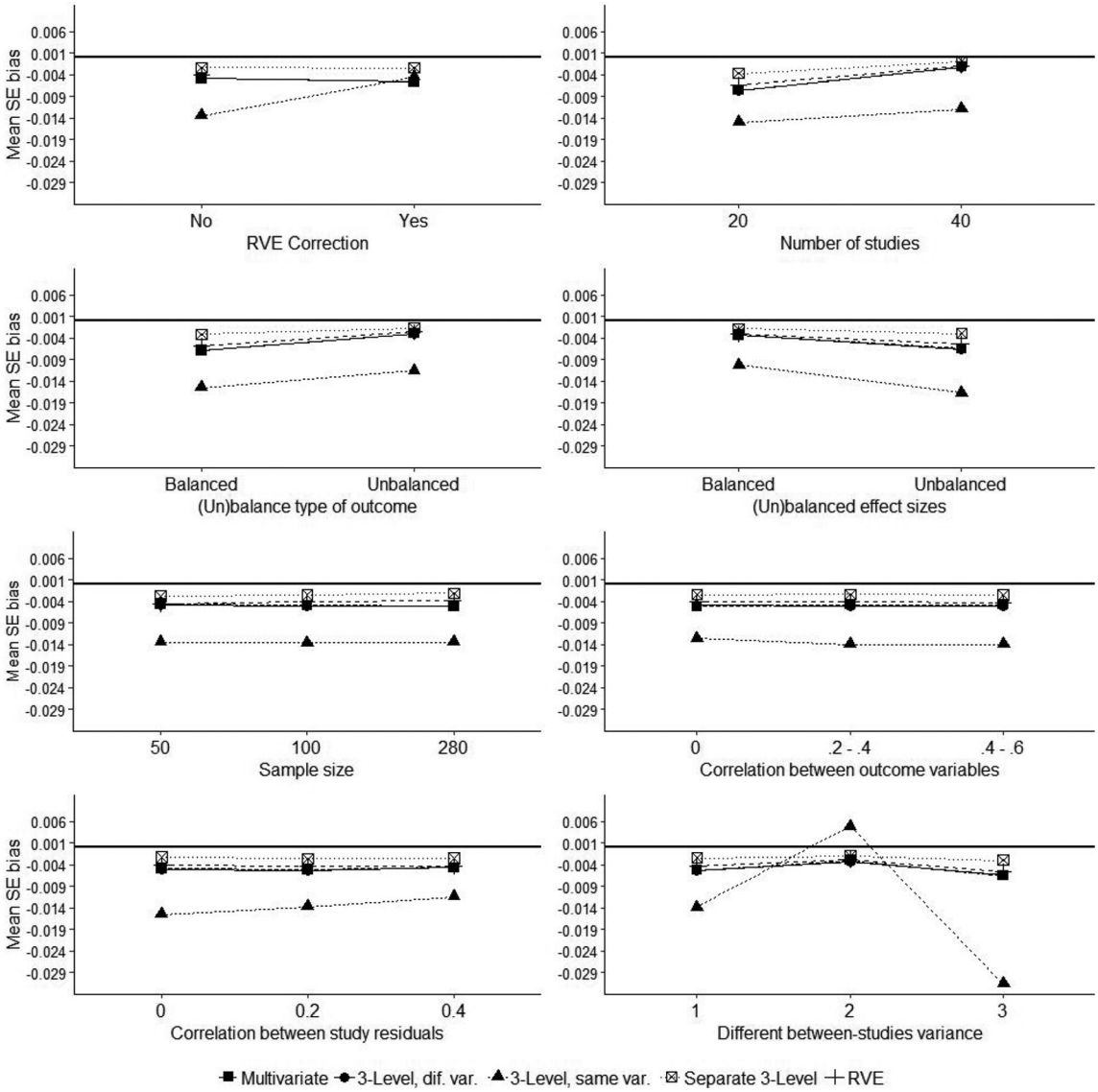

Fig. 1 Mean standard error bias of outcome Type_1 for each model and simulation factor condition when robust variance correction was not applied (except for the first graph). In the last graph, 1 refers to the condition where all between-studies variances are equal to 0.072; 2 = condition where $\sigma^2_{u_1} = 0.017$, $\sigma^2_{u_2} = 0.072$, and $\sigma^2_{u_3} = 0.153$; 3 = condition where $\sigma^2_{u_1} = 0.153$, $\sigma^2_{u_2} = 0.072$, and $\sigma^2_{u_3} = 0.017$. SE = standard error; RVE = robust variance estimation

observed for the multivariate approach and followed exactly the same pattern. The coverage proportion of the 95% CIs also followed the same pattern as the RSEBs: they were smaller than the nominal level, .95, especially if the number of studies was small.

Model 4: Separate three-level models

The simulation factors that predicted the standard error bias of these models were the number of primary studies (η^2=.264, .527, and .380 for SE[$\widehat{\gamma}_1$], SE[$\widehat{\gamma}_2$], and SE[$\widehat{\gamma}_3$], respectively), whether the moderator variable 'type of outcome' was (un)balanced across studies (η^2=.078, .019, and .233), and whether effect sizes were (un)balanced across studies (η^2=.045, .048, and .020). When the number of primary studies was 20, the standard error of $\widehat{\gamma}_3$ was underestimated, especially if the variable 'type of outcomes' was unbalanced (see Table 3). The standard error of $\widehat{\gamma}_2$ was underestimated when the variable 'type of outcomes' and the number of effect sizes across studies was unbalanced, and the standard error of $\widehat{\gamma}_1$ was adequately estimated. When the number of studies was larger (k = 40), all RSEBs were within the appropriate cutoffs. The results from the coverage proportion of the 95% CIs followed a similar trend than the results from the RSEB: they were smaller than .95, especially if the standard error very underestimated.

Table 3 Relative standard error bias and coverage proportion of the 95% confidence intervals (in italic) yielded by the multivariate model (Model 1), the three-level model that specifies outcome-specific study-effects (Model 3), and the three separate three-level models for each type of outcome (Model 4)

Model fitted	Type out.	ESs	$k = 20$			$k = 40$		
			SE($\widehat{\gamma}_1$)	SE($\widehat{\gamma}_2$)	SE($\widehat{\gamma}_3$)	SE($\widehat{\gamma}_1$)	SE($\widehat{\gamma}_2$)	SE($\widehat{\gamma}_3$)
Model 1	Unbal.	Unbal.	– 9.28 *.92*	**– 16.30** *.88*	**– 30.65** *.84*	– 6.47 *.92*	– 7.96 *.92*	**– 16.02** *.90*
		Bal.	– 5.38 *.93*	**– 12.00** *.90*	**– 26.19** *.86*	– 1.48 *.94*	4.46 *.96*	– 3.76 *.94*
	Bal.	Unbal.	**– 14.03** *.89*	**– 11.92** *.90*	**– 14.08** *.90*	**– 11.57** *.91*	– 8.43 *.91*	– 8.56 *.92*
		Bal.	**– 14.92** *.90*	**– 14.26** *.89*	**– 14.77** *.89*	– 2.88 *.93*	0.77 *.94*	– 3.86 *.93*
Model 3	Unbal.	Unbal.	– 9.19 *.91*	**– 16.61** *.88*	**– 31.90** *.83*	– 5.92 *.92*	– 7.83 *.91*	**– 15.56** *.90*
		Bal.	– 5.40 *.92*	**– 12.20** *.90*	**– 26.56** *.86*	– 1.11 *.94*	4.55 *.95*	– 2.97 *.94*
	Bal.	Unbal.	**– 14.01** *.89*	**– 11.75** *.90*	**– 15.12** *.89*	**– 11.44** *.91*	– 8.20 *.91*	– 9.03 *.93*
		Bal.	**– 14.77** *.89*	**– 13.89** *.88*	**– 14.58** *.89*	– 2.59 *.93*	1.52 *.94*	– 3.63 *.93*
Model 4	Unbal.	Unbal.	– 5.23 *.92*	**– 11.48** *.90*	**– 15.13** *.88*	– 4.06 *.93*	– 1.21 *.93*	– 2.54 *.93*
		Bal.	– 3.24 *.93*	– 5.61 *.92*	**– 15.30** *.88*	– 1.24 *.94*	4.99 *.95*	– 1.98 *.94*
	Bal.	Unbal.	– 7.01 *.91*	– 2.89 *.92*	– 7.02 *.91*	– 6.41 *.92*	– 0.66 *.93*	– 1.23 *.94*
		Bal.	– 9.22 *.91*	– 7.62 *.90*	**– 10.01** *.90*	– 2.18 *.93*	1.49 *.94*	– 3.63 *.93*

Notes. *Notes.* Model 1: Multivariate model; Model 3: Three-level model with separate random study effect; Model 4: Separate three-level models; SE = standard error; ESs = effect sizes; Type out. = Type of outcome; k = number of primary studies; Bal. = balanced; Unbal.= unbalanced; Values in bold indicate unacceptable relative bias values. The other simulation factors were fixed to intermediate levels: average sample size = 100, correlation between study-residuals = .2, correlation between variables = .2 - .4, all between-studies variances are 0.072, and the robust variance correction was not applied

Model 5: Two-level model (with RVE method)

The number of primary studies (η^2=.299, .558, and .281 for SE[$\widehat{\gamma}_1$], SE[$\widehat{\gamma}_2$], and SE[$\widehat{\gamma}_3$], respectively), whether the variable 'type of outcome' was (un)balanced (η^2=.146, .026, and .263), and whether effect sizes were balanced across studies (η^2=.077, .192, and .137) were the most influential factors in predicting the standard error bias of the random-effects model with RVE method. Surprisingly, whether the between-studies variances were equal or different for each type of outcome did not have an influential effect on the bias (η^2=.015, .00007, .016 for SE[$\widehat{\gamma}_1$], SE[$\widehat{\gamma}_2$], and SE[$\widehat{\gamma}_3$], respectively), although standard errors depend on the between-studies variance estimate. Because theoretically this factor condition must influence the standard error estimates, we have included it in Table 5, where the RSEBs of this combination of conditions is shown.

When the number of primary studies was 40, almost all standard errors were properly estimated. When the meta-analysis included 20 studies, the standard error of $\widehat{\gamma}_1$ was underestimated when the three types of outcomes were equally likely to be reported and the between-study variance was small (0.017) or medium (0.072). The standard error of $\widehat{\gamma}_2$ was sometimes underestimated when the other between-studies variances were different, and the standard error of $\widehat{\gamma}_3$ was underestimated in almost all conditions, except when the types of outcomes and the number of effect sizes across studies were unbalanced. Although we find some differences in the RSEBs across the different patterns of between-studies variances, these differences are not the ones expected theoretically. A last observation is that, although standard errors were underestimated under some conditions, the coverage proportions of the 95% CIs were always close to the nominal level.

Variance components

Table 6 shows the average between-studies variance estimates and their relative bias segregated by Model type and by their

Table 4 Relative standard error bias and coverage proportion of the 95% confidence intervals (in italic) of the three-level model that specified a common study-effect for all effect sizes (Model 2) for the conditions that resulted relevant in the ANOVA

	Variances	Type out.	ESs	$k = 20$			$k = 40$		
				$SE(\widehat{\gamma}_1)$	$SE(\widehat{\gamma}_2)$	$SE(\widehat{\gamma}_3)$	$SE(\widehat{\gamma}_1)$	$SE(\widehat{\gamma}_2)$	$SE(\widehat{\gamma}_3)$
No RV corr.	$\sigma^2_{u_1} = \sigma^2_{u_2} = \sigma^2_{u_3} = 0.072$	Unbal.	Unbal.	**– 22.36**	**– 25.88**	**– 29.39**	**– 22.38**	**– 26.71**	**– 25.02**
				.86	*.85*	*.85*	*.87*	*.84*	*.87*
			Bal.	**– 13.49**	**– 16.43**	**– 24.59**	**– 21.91**	**– 19.00**	**– 19.42**
				.91	*.89*	*.88*	*.88*	*.89*	*.88*
		Bal.	Unbal.	**– 24.71**	**– 21.82**	**– 23.09**	**– 24.75**	**– 22.96**	**– 23.30**
				.86	*.88*	*.87*	*.87*	*.87*	*.86*
			Bal.	**– 17.62**	**– 16.14**	**– 18.00**	**– 24.77**	**– 22.09**	**– 24.10**
				.89	*.89*	*.88*	*.84*	*.87*	*.86*
	$\sigma^2_{u_1} = 0.017$	Unbal.	Unbal.	– 0.64	**– 28.34**	**– 46.43**	– 0.04	**– 28.86**	**– 43.54**
	$\sigma^2_{u_2} = 0.072$			*.94*	*.84*	*.72*	*.96*	*.84*	*.72*
	$\sigma^2_{u_3} = 0.153$		Bal.	9.21	**– 25.75**	**– 42.68**	4.25	**– 34.76**	**– 49.25**
				.96	*.85*	*.75*	*.96*	*.79*	*.66*
		Bal.	Unbal.	4.72	**– 20.63**	**– 36.26**	**12.71**	**– 19.84**	**– 38.71**
				.97	*.86*	*.77*	*.97*	*.88*	*.77*
			Bal.	**22.62**	**– 13.79**	**– 35.12**	**21.00**	**– 19.75**	**– 38.62**
				.98	*.90*	*.80*	*.98*	*.87*	*.75*
	$\sigma^2_{u_1} = 0.153$	Unbal.	Unbal.	**– 30.70**	**– 21.61**	– 8.57	**– 28.06**	**– 19.40**	1.35
	$\sigma^2_{u_2} = 0.072$			*.83*	*.87*	*.94*	*.83*	*.90*	*.94*
	$\sigma^2_{u_3} = 0.017$		Bal.	**– 19.92**	– 6.23	0.38	**– 31.48**	– 8.61	**24.94**
				.88	*.93*	*.97*	*.82*	*.91*	*.97*
		Bal.	Unbal.	**– 38.42**	**– 19.44**	4.64	**– 37.67**	**– 22.46**	8.24
				.77	*.89*	*.95*	*.77*	*.86*	*.96*
			Bal.	**– 32.07**	**– 13.94**	**22.46**	**– 39.31**	**– 21.26**	**17.46**
				.81	*.90*	*.98*	*.77*	*.86*	*.97*
RV corr.	$\sigma^2_{u_1} = \sigma^2_{u_2} = \sigma^2_{u_3} = 0.072$	Unbal.	Unbal.	**– 10.10**	**– 13.00**	**– 20.70**	– 6.63	– 9.57	– 8.83
				.94	*.93*	*.94*	*.93*	*.94*	*.95*
			Bal.	– 4.45	– 7.09	**– 21.31**	– 2.09	1.34	– 0.80
				.94	*.94*	*.94*	*.94*	*.95*	*.94*
		Bal.	Unbal.	**– 13.31**	**– 10.00**	**– 12.35**	– 8.31	– 6.66	– 6.90
				.93	*.94*	*.94*	*.94*	*.93*	*.94*
			Bal.	**– 11.29**	– 8.55	**– 12.11**	– 3.15	0.37	– 3.19
				.93	*.94*	*.93*	*.95*	*.95*	*.94*
	$\sigma^2_{u_1} = 0.017$	Unbal.	Unbal.	– 7.16	**– 10.18**	**– 21.97**	– 4.10	– 6.08	– 7.25
	$\sigma^2_{u_2} = 0.072$			*.94*	*.94*	*.94*	*.95*	*.94*	*.94*
	$\sigma^2_{u_3} = 0.153$		Bal.	– 3.59	**– 10.78**	**– 15.79**	1.97	– 3.93	– 7.11
				.94	*.93*	*.94*	*.96*	*.95*	*.93*
		Bal.	Unbal.	**– 10.85**	**– 10.94**	– 7.87	– 1.09	– 5.50	– 5.95
				.95	*.94*	*.94*	*.96*	*.94*	*.94*
			Bal.	– 6.23	– 8.72	**– 11.40**	1.26	0.39	3.79
				.96	*.93*	*.93*	*.95*	*.96*	*.96*
	$\sigma^2_{u_1} = 0.153$	Unbal.	Unbal.	– 7.16	**– 10.18**	**– 21.97**	– 4.10	– 6.08	– 7.25
	$\sigma^2_{u_2} = 0.072$			*.94*	*.92*	*.95*	*.94*	*.94*	*.94*
	$\sigma^2_{u_3} = 0.017$		Bal.	– 3.59	**– 10.78**	**– 15.79**	1.97	– 3.93	– 7.11
				.96	*.93*	*.96*	*.94*	*.94*	*.96*
		Bal.	Unbal.	**– 10.85**	**– 10.94**	– 7.87	– 1.09	– 5.50	– 5.95
				.93	*.95*	*.94*	*.95*	*.93*	*.94*
			Bal.	– 6.23	– 8.72	**– 11.40**	1.26	0.39	3.79
				.94	*.94*	*.95*	*.96*	*.94*	*.94*

Notes. $\sigma^2_{u_1}$, $\sigma^2_{u_2}$, and $\sigma^2_{u_3}$ = between-studies variances of outcome Type_1, Type_2, and Type_3, respectively; SE = standard error; $\widehat{\gamma}_1$, $\widehat{\gamma}_2$, and $\widehat{\gamma}_3$ = estimated mean effect of outcome Type_1, Type_2, and Type_3 respectively; SE = standard error; ESs = effect sizes; k = number of primary studies; No RV corr. = robust variance correction was not applied. RV corr.: robust variance correction was applied after fitting the model; Bal. = balanced; Unbal.=unbalanced. Values in bold indicate unacceptable relative bias values. The other factors were fixed to intermediate values: average sample size = 100, correlation between study-residuals = .2, correlation between variables = .2 - .4

Table 5 Relative standard error bias and coverage proportion of the 95% confidence intervals (in italic) yielded by random-effects model (Model 5) estimated using RVE method

Variances	Type out.	ESs	$k = 20$			$k = 40$		
			$SE(\widehat{\gamma}_1)$	$SE(\widehat{\gamma}_2)$	$SE(\widehat{\gamma}_3)$	$SE(\widehat{\gamma}_1)$	$SE(\widehat{\gamma}_2)$	$SE(\widehat{\gamma}_3)$
$\sigma^2_{u_1} = \sigma^2_{u_2} = \sigma^2_{u_3} = 0.072$	Unbal.	Unbal.	– 7.27 *.94*	**– 19.54** *.92*	**– 26.96** *.94*	– 6.10 *.94*	– 5.18 *.94*	**– 13.13** *.94*
		Bal.	– 5.55 *.95*	– 7.68 *.94*	**– 19.46** *.95*	– 1.77 *.95*	1.93 *.96*	– 2.38 *.96*
	Bal.	Unbal.	**– 13.33** *.93*	– 7.18 *.95*	– 8.88 *.95*	– 8.25 *.94*	– 5.44 *.95*	– 4.83 *.94*
		Bal.	**– 10.82** *.94*	– 9.95 *.94*	**– 12.45** *.93*	– 1.46 *.95*	– 0.1 *.95*	– 2.8 *.94*
$\sigma^2_{u_1} = 0.017$ $\sigma^2_{u_2} = 0.072$ $\sigma^2_{u_3} = 0.153$	Unbal.	Unbal.	– 5.29 *.94*	– 9.73 *.95*	**– 27.03** *.95*	– 3.44 *.95*	– 4.84 *.94*	**– 12.81** *.96*
		Bal.	– 3.05 *.94*	**– 12.11** *.93*	**– 14.29** *.95*	0.53 *.96*	– 4.78 *.95*	– 5.48 *.94*
	Bal.	Unbal.	**– 10.78** *.94*	**– 10.36** *.94*	– 6.82 *.96*	– 2.73 *.95*	– 3.35 *.95*	– 3.06 *.96*
		Bal.	– 6.93 *.95*	– 8.78 *.93*	**– 12.21** *.93*	4.39 *.96*	– 0.98 *.95*	1.60 *.96*
$\sigma^2_{u_1} = 0.153$ $\sigma^2_{u_2} = 0.072$ $\sigma^2_{u_3} = 0.017$	Unbal.	Unbal.	– 7.17 *.94*	**– 12.44** *.94*	**– 26.53** *.95*	– 2.61 *.95*	– 5.43 *.94*	– 8.85 *.95*
		Bal.	– 2.29 *.95*	– 6.68 *.94*	**– 17.5** *.95*	– 2.58 *.94*	– 2.86 *.94*	– 2.7 *.95*
	Bal.	Unbal.	– 8.68 *.95*	– 9.07 *.95*	– 8.37 *.95*	– 4.5 *.94*	– 7.02 *.94*	– 6.08 *.93*
		Bal.	– 6.62 *.95*	**– 11.21** *.95*	– 7.04 *.96*	2.86 *.96*	– 2.56 *.94*	– 0.47 *.94*

Notes. $\sigma^2_{u_1}$, $\sigma^2_{u_2}$, and $\sigma^2_{u_3}$ = between-studies variances of outcome Type_1, Type_2, and Type_3, respectively; SE = standard error; $\widehat{\gamma}_1$, $\widehat{\gamma}_2$, and $\widehat{\gamma}_3$ = estimated mean effect of outcome Type_1, Type_2, and Type_3, respectively; k = number of primary studies; ESs: effect sizes; Type out.: type of outcome; Bal. = balanced; Unbal.=unbalanced. Values in bold indicate the relative standard error bias was out of the recommended thresholds. The other simulation factors were fixed to intermediate values: average sample size = 100, correlation between study-residuals = .2, correlation between variables = .2 - .4

real values. We have only included those models that estimate a different between-studies variance for each type of outcome, namely the Multivariate model (Model 1), the three-level model with separate random study effects (Model 3), and the three-level models applied separately for each type of outcome (Model 4). On average, when the variability of the effect sizes across studies was the same for all types of outcomes ($\sigma_{u_1} = \sigma_{u_2} = \sigma_{u_3} = 0.072$), the three models led to unbiased between-studies variance estimates. However, all models overestimated the between-studies variance when this variance was small ($\sigma_u = 0.017$). Generally speaking, between-studies variances estimated by Model 1 (multivariate approach) were mostly influenced by the number of studies and by the sample size of primary studies: if both the number of studies and the average sample size were small (k =20 and average n = 50), a relative bias of about – 15% was observed. In contrast, the between-studies variance estimates yielded by the multilevel approaches (Model 3 and Model 4) were mainly affected by the number of studies and by the correlation between outcome variables. When the number of studies was high (k =40), almost all between-studies variances were adequately estimated. However, when there were only 20 studies in the meta-analysis, the between-studies variances were often underestimated and, in some conditions, overestimated (when the true variance was small and when outcome variables were highly correlated). Even so, the relative bias observed in Model 4 (separate three-level models) was larger (about 25%) than the ones observed in Model 3 (three-level model with separate random study effects; about 15%). More information about between-variance estimates and their relative bias can be found in the supplementary material: https://osf.io/ywum6/.

Discussion

This study aimed to explore which method that addresses dependency between effect sizes performs better when the

Table 6 Average between-studies variance estimates and relative bias between parenthesis

	Original values of the between-studies variance								
	$\sigma^2_{u_1} = 0.017$, $\sigma^2_{u_1} = 0.072$, $\sigma^2_{u_3} = 0.153$			$\sigma^2_{u_1} = 0.072$, $\sigma^2_{u_1} = 0.072$, $\sigma^2_{u_3} = 0.072$			$\sigma^2_{u_1} = 0.153$, $\sigma^2_{u_1} = 0.072$, $\sigma^2_{u_3} = 0.017$		
	$\widehat{\sigma}^2_{u_1}$ (RB)	$\widehat{\sigma}^2_{u_2}$ (RB)	$\widehat{\sigma}^2_{u_3}$ (RB)	$\widehat{\sigma}^2_{u_1}$ (RB)	$\widehat{\sigma}^2_{u_2}$ (RB)	$\widehat{\sigma}^2_{u_3}$ (RB)	$\widehat{\sigma}^2_{u_1}$ (RB)	$\widehat{\sigma}^2_{u_2}$ (RB)	$\widehat{\sigma}^2_{u_3}$ (RB)
Model 1	0.018 (7.92%)	0.071 (– 1.78%)	0.150 (1.90%)	0.070 (– 2.35%)	0.071 (– 1.48%)	0.074 (2.22%)	0.148 (– 3.02%)	0.071 (– 2.04%)	**0.022 (31.02%)**
Model 3	**0.023 (37.62%)**	0.074 (3.44%)	0.152 (– 0.60%)	0.075 (4.22%)	0.075 (3.60%)	0.076 (6.23%)	0.152 (– 3.02%)	0.074 (– 2.04%)	**0.025 (31.02%)**
Model 4	**0.022 (27.00%)**	0.072 (– 0.67%)	0.140 (– 8.77%)	0.073 (1.86%)	0.071 (– 0.84%)	0.067 (– 7.04%)	0.150 (– 1.75%)	0.071 (– 1.05%)	**0.019 (12.97%)**

Notes. Model 1: Multivariate model; Model 3: Three-level model with separate random study effects; Model 4: Separate three-level models. RB= relative bias; $\sigma^2_{u_1}$ = between-studies variance of effect sizes belonging to Type of Outcome 1; $\sigma^2_{u_2}$ = between-studies variance of effect sizes belonging to Type of Outcome 2; $\sigma^2_{u_3}$ = between-studies variance of effect sizes belonging to Type of Outcome 3; RB= relative bias. Values in bold indicate the relative standard error bias was out of the recommended thresholds

interest is in the effects for different types of outcome (i.e., in their estimates, estimated standard errors and 95% CI coverage proportions). The manuscript extends the study of Park and Beretvas (2019) in three ways. First, it explores the performance of the multivariate approach, three three-level models (a model including one random study effect, a model with a random study effect for each type of outcome, and a separate three-level model for each type of outcome), and a two-level model. Three-level models with one random effect, or separate three-level models are the ones most commonly applied in practice, so learning about their performance is highly necessary. Second, in this study data have been generated following a more realistic model (i.e., multivariate two-level hierarchical model). Third, this study includes additional conditions, including the situation in which each category of the moderator variable (i.e., each outcome) have a different between-studies variance.

A first conclusion is that all approaches included in this study give accurate estimates of the effect sizes. Regarding the non-convergence rates, the multivariate two-level hierarchical model and the three-level model that specified different study-effects converged in more than the 99% of the analyses. This contrasts with the results of Park and Beretvas (2019), who found that the three-level model had convergence problems when the number of primary studies was small, sampling variance was larger compared to the within-study and between-studies variance, sample sizes were different across studies (in our simulation, sample sizes were always different across studies), and when the number of effect sizes and the number of outcomes were unbalanced across studies. In our simulation study, we did not have a condition where the three between-studies variances were small, so that might explain the differences between the results.

A second general result is that the robust variance correction did not lead to better standard error estimates when it was applied to the estimates yielded by the model that was used to generate the data (i.e., multivariate two-level); actually, in this case the standard errors were still underestimated. However, due to the small sample correction proposed by Tipton (2015), the coverage proportion of the 95% CIs was still adequate, meaning that the Type I error was under control even if standard errors were underestimated. In fact, the *a posteriori* RV correction led to correct 95% CIs in all models even if they were misspecified. Therefore, we do recommend the routine implementation of the robust variance correction after carrying out any of these models.

Taking a closer look at the performance of each model, we can conclude that the separate three-level meta-analysis for each type of outcome is the best option to get adequate estimates of the fixed effect standard errors as long as there are enough effect sizes within each type of outcome. However, with this approach it is not possible to statistically compare the overall outcome-specific effects with each other, and it should be also kept in mind that between-studies variances are often underestimated in this model. Moreover, compared to the use of a two-level model, this approach results in less precise (though still unbiased) estimates of the effect sizes. When the number of primary studies is large enough (around 40), the multivariate model, the three-level model that specifies different study-effects for each type of outcome, and the random-effects model with RVE method also give, in general, proper estimates of the standard errors and of the variance components if there are 40 studies or more. In addition, when the robust variance correction is applied using the small sample correction, the CIs are properly estimated in all models, being the false positive rate close to the optimal .05 level. Furthermore, these three models allow the statistical comparison of the overall outcome-specific effects, which is an additional advantage.

There are some results from each specific method that are worth mentioning. First, the multivariate approach and the three-level model that specified different study-effects for each type of outcome performed very similar, aligning with the results of Van den Noortgate et al. (2013). However, it is important to mention we have used the population correlation between outcomes variables within studies for applying the multivariate method, which in practice it is not known. Therefore, in practice the multivariate approach is expected to perform worse because most of the correlations will be approximations or estimates of the true one. Also, this was the method that led to the most accurate between-studies variance estimates.

Second, we observed that when the robust variance correction was not applied, the three-level model that assumed a common study-effect for all effect sizes led to underestimated standard errors. This result was expected, because this model estimates a common between-studies variance for the three outcomes (normally this estimate is close to the average of the three true between-studies variances). This estimated between-studies variance is the one used to calculate the weights of all effect sizes, and the sum of these weights is directly related to the standard error estimate of the overall effect. Under some conditions, the between-studies variances were different for the three outcomes (e.g., $\sigma^2_{u_1} = 0.017$, $\sigma^2_{u_2} = 0.072$, and $\sigma^2_{u_3} = 0.153$). Therefore, we expected that this misspecified three-level model overestimated the standard error of outcome Type_1 or of outcome Type_3 when their true between-studies variances were 0.017 (because the model assigned a larger between-studies variance, close to 0.081, to the effect sizes weights), and that it underestimated the standard error of outcome Type_1 or outcome Type_3 when their true between-studies variances were 0.153. This pattern is indeed reproduced in the results. However, this pattern disappeared once the robust variance correction was applied. With the robust variance correction, standard errors were still underestimated when the number of primary studies was 20, but when the number of studies increased to 40, all standard errors were properly estimated. Furthermore, coverage proportion of the 95% CIs were adequate when the robust variance correction was applied in all conditions. These results are very relevant, because this model is commonly applied in practice, and until this study, no research has indicated that the standard errors of moderator effects yielded by these models can be very biased. For future meta-analysis, we recommend researchers apply this model together with the robust variance correction to get robust standard error estimates and proper 95% CIs.

Third and last, regarding the two-level model that used RVE to correct standard error estimates, we see that when the number of studies was 40, almost all standard errors were properly estimated, despite the fact that this model also gives only one global estimate of the between-studies variance. More importantly, when the RVE method was applied using the small sample correction, the coverage proportion of the 95% CIs was always close to .95, meaning that the Type I error rate was always under control, even when the number of studies is small.

The present simulation study is not free of limitations. First, the conclusions extracted only apply to the conditions generated in the simulation. Although we have tried to simulate realistic data by taking characteristics of observed meta-analyses, it is impossible to account for all data structures that can be found in practice. Also, it should be kept in mind that in this simulation study, effect sizes have been directly generated from a normal distribution instead of calculated from simulated raw data. Therefore, results are only generalizable to situations where effect sizes are actually normally distributed. In addition, another limitation is that we have not explored the performance of other competing models, such as a multivariate model where the three between-studies variances are constrained to be the same, a model that would be more comparable to the random-effects with RVE method and to the three-level model that specified a common study-effect for all outcomes. In addition, regarding data generation, we have assumed that effect sizes that represent the same type of outcome are exactly the same. However, this might not be a realistic assumption, as it can be expected that there is some variation between effect sizes that refer to the same type of outcome. Future studies can extend this study by accounting for these alternative scenarios.

Acknowledgements This research has been supported by the Research Foundation – Flanders (FWO), through Grant G.0798.15N to the University of Leuven, Belgium. The opinion expressed are those of the authors and do not represent views of the FWO. For the simulations, we used the infrastructure of the VSC–Flemish Supercomputer Center, funded by the Hercules foundation and the Flemish Government–Department EWI.

Open Practices Statements The R code used to fitted the five different models, a simulated dataset to run the code, and the graphs of the standard error bias of type of outcome 2 and 3 are available at https://osf.io/ywum6/. This study was not preregistered.

References

Becker, B. J. (2000). Multivariate meta-analysis. In H. E. A. Tinsley & E. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 499–525). Orlando, FL: Academic Press.

Bell, R. M., & McCaffrey, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, *28*(2), 169-181.

Cheung, M. W.-L. (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). Hillsdale, NJ: Erlbaum.

Fernández-Castilla, B., Jamshidi, L., Declercq, L., Beretvas, S. N., Onghena, P., & Van Den Noortgate, W. (2020). *The application of meta-analytic models with multiple random effects: A systematic review*. Manuscript accepted for publication in *Behavior Research Methods.*

Fisher, Z., Tipton, E., & Zhipeng, H. (2017). *Robumeta: Robust Variance Meta-Regression*. R package version 2.0. https://CRAN.R-project.org/package=robumeta

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

Gleser, L. J., Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.). *The Handbook of Research Synthesis* (pp. 339–355). New York, NY: Russel Sage Foundation.

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.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling. An overview and a meta-analysis. *Sociological Methods & Research*, *26*, 329–367.

Kalaian, H. A., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods*, *1*, 227–235.

Lebuda, I., Zabelina, D. L., & Karwowski, M. (2016). Mind full of ideas: A meta-analysis of the mindfulness–creativity link. *Personality and Individual Differences*, *93*, 22–26.

Lee, S. (2014). *Within study dependence in meta-analysis: Comparison of GLS method and multilevel approaches* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database (UMI No. 3344745).

Moeyaert, M., Ugille, M., Beretvas, S. N., Ferron, J., Bunuan, R., & Van den Noortgate, W. (2017). Methods for dealing with multiple outcomes in meta-analysis: a comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *International Journal of Social Research Methodology*, *20*, 559–572.

Owen, K. B., Parker, P. D., Van Zanden, B., MacMillan, F., Astell-Burt, T., & Lonsdale, C. (2016). Physical activity and school engagement in youth: A systematic review and meta-analysis. *Educational Psychologist*, *51*, 129–145.

Park, S., & Beretvas, S. N. (2019). Synthesizing effects for multiple outcomes per study using robust variance estimation versus the three-level model. *Behavior Research Methods, 51*, 152–171.

Pustejovsky, J. (2018). *clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections*. R package version 0.3.2. https://CRAN.R-project.org/package=clubSandwich

Pustejovsky, J., Tipton, B., & Aloe, A. (2018, July). *Combining robust variance estimation with models for dependent effect sizes.* Paper presented at the 13th annual Society of Research Synthesis Methodology Conference, Bristol, UK.

Pustejovsky, J. E. & Tipton, E. (2017). Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business and Economic Statistics*, *36*(4), 672–683.

Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin*, *103*, 111–120.

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics*, *10*, 75–98.

Spruit, A., Assink, M., van Vugt, E., van der Put, C., & Stams, G. J. (2016). The effects of physical activity interventions on psychosocial outcomes in adolescents: A meta-analytic review. *Clinical Psychology Review*, *45*, 56–71.

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.

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). A history of meta-regression: Technical, conceptual, and practical developments between 1974 and 2018. *Research Synthesis Methods, 10,* 161–179.

Team, R. C. (2012). R: A language and environment for statistical computing. 2012. Vienna, Austria: R Foundation for Statistical Computing, 10.

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

Viechtbauer, W. (2010). Conducting Meta-Analyses in *R* with the *metafor* Package. *Journal of Statistical Software*, *36,* 1–48.