Synthesizing effects for multiple outcomes per study using robust variance estimation versus the three-level model

Sunyoung Park¹ · S. Natasha Beretvas²

Published online: 7 November 2018
© Psychonomic Society, Inc. 2018

Abstract

Primary studies increasingly report information that can be used to provide multiple effect sizes. Of interest in this study, primary studies might compare a treatment and a control group on multiple related outcomes that result in multiple dependent effect sizes to be synthesized. There are a number of ways to handle the resulting within-study “multiple-outcome” dependency. The present study focuses on use of the multilevel meta-analysis model (Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013) and robust variance estimation (Hedges, Tipton, & Johnson, 2010) for handling this dependency, as well as for estimating outcome-specific mean effect sizes. We assessed these two approaches under various conditions that differed from each other in within-study sample size; the number of effect sizes per outcome; the number of outcomes per study; the number of studies per meta-analysis; the ratio of variances at Levels 1, 2, and 3; and the true correlation between pairs of effect sizes at the between-study level. Limitations and directions for future research are discussed.

Keywords Multilevel meta-analysis · Robust variance estimation · Multiple-outcome dependency · Simulation study

Meta-analysis

Meta-analysis is used to synthesize the results from previous studies investigating a common research question that can be quantified using an effect size (Glass, 1976). Various types of effect sizes are categorized by the study design and the scale of the variables of interest—including, among others, the standardized mean difference, the odds ratio, and the correlation coefficient (Borenstein, Hedges, Higgins, & Rothstein, 2009; Hedges & Olkin, 1985). In the present study we focused on the standardized mean difference with small-sample-size correction (Hedges’s *g*; Hedges, 1981, p. 226).

After the effect size capturing the results for each study is calculated, typically the next step in a meta-analysis involves calculating a summary effect size that has been averaged across all of the studies. Two general models can be assumed when summarizing studies’ effect sizes: the fixed- and the

random-effect models. Under the *fixed-effect model*, it is assumed that effect size estimates are sampled from a population of estimates that share a single, true effect size, and thus that the resulting effect size estimates vary only as a function of sampling error. Under the more general *random-effect model*, on the other hand, the estimated effect size is assumed to be composed of the true effect size and the study-specific error and sampling error. If it is reasonable to assume that all of the sampled studies are sufficiently homogeneous, then the fixed-effect model can be used. However, that scenario is sufficiently rare that the random-effect model is more commonly encouraged. Methodological researchers have argued that the assumption of the fixed-effect model is often not practical. For that reason, many researchers (e.g., Hedges & Vevea, 1998; Hunter & Schmidt, 2004; Kisamore & Brannick, 2008; National Research Council, 1992; Raudenbush, 1994; Schmidt, Oh, & Hayes, 2009, among others) strongly recommend using the random-effect model. Thus, in the present study we used only the random-effect model.

If we can assume that each observed effect size is independent, then univariate meta-analysis models can be used for the synthesis and moderator analyses. However, when analyzing meta-analytic databases in which there are multiple effects sizes per study, it is recommended that meta-analysts handle this within-study dependency by using one of a number of options. A failure to handle within-study dependence can

✉ Sunyoung Park
sunyoungpark@callutheran.edu

¹ Graduate School of Psychology, California Lutheran University, 60 West Olsen Road #4250, Thousand Oaks, CA 91360, USA

² College of Education, University of Texas at Austin, 1912 Speedway, Stop D5800, Austin, TX 78712, USA

negatively impact the resulting inferences (e.g., Becker, 2000; Cheung, 2015; Cooper, Hedges, & Valentine, 2009; Matt & Cook, 2009).

Within-study dependence in effect size estimates There could be several reasons for within-study dependence. The first common reason is that multiple treatments are being compared with a single comparison group in a study, resulting in multiple effect sizes (one for each treatment). The second common reason for within-study dependence is that effect sizes for multiple related outcomes (even with a single treatment group and a single comparison group) are reported in a study. The present study focuses only on the second scenario, in which within-study dependence results from multiple outcomes per study.

As with analyses of primary data, the results of meta-analysis can be problematic when dependency in the effect sizes is ignored and estimates are treated as if they are independent. As the magnitude of the intercorrelations among pairs of effect sizes increases, the amount of independent information provided by each effect size decreases. The standard errors of the effect size estimates thus are likely to be underestimated, resulting in too-small confidence intervals and too high a proportion of incorrect rejections of the null hypotheses (Becker, 2000).

If only a very small proportion of studies report multiple outcomes in a large meta-analytic dataset, then ignoring the dependence among effect sizes will not have as strong an impact on the results as in scenarios with a higher proportion of dependent effect sizes. A sensitivity analysis can be conducted that compares the results of meta-analysis when ignoring versus not ignoring the dependence. If the results do not differ greatly, then choosing to ignore the dependence can be considered as an option. However, the standard error of the parameter estimates in the meta-analysis is likely to be underestimated if a meta-analyst ignores dependence, so this ad-hoc approach to handling (i.e., ignoring) within-study dependence is not encouraged (Becker, Hedges, & Pigott, 2004).

Another ad-hoc procedure for handling dependence involves excluding dependent effect sizes from the analysis. This is typically done by selecting one representative effect size per study, by averaging together dependent effect sizes within a study, or by separately synthesizing effect sizes for each outcome across studies. If only one effect size is chosen per study, then the chosen effect sizes can be considered independent of other studies' effect size estimates. However, this approach unnecessarily reduces the sample size, and thus the power of the meta-analysis is lowered. The generalizability of the meta-analysis's results is also restricted. In addition, interpretation of the results of the meta-analysis can be compromised if there is no valid criterion for selecting the representative effect size and choosing to exclude the other effect sizes. Moreover, if researchers are interested in

assessing potential differences between effect sizes as a function of the outcome used, then choosing one effect size per study also prohibits such analyses.

If the multiple outcomes reported in a study are measures of the same construct, it could be reasonable to use the average of the multiple effect sizes in a study (Borenstein et al., 2009). However, several methodological studies have indicated that the use of averaged effect sizes could be inappropriate (Raudenbush, Becker, & Kalaian, 1988) and could negatively impact the validity of results (Matt & Cook, 2009), because of overestimated standard errors (Moeyaert et al., 2017). Since many applied researchers are interested in synthesizing outcome-specific effect sizes (e.g., Sheridan, Kim, Beretvas, Smith, & Park, 2017), we focused on scenarios in which the multiple outcomes reported in a study are not measures of the same construct and in which the meta-analyst wants to synthesize outcome-specific effect sizes.

When handling dependence resulting from effect sizes being reported for multiple related outcomes, it is possible to conduct separate meta-analyses for each outcome's effect size. However, this approach will also result in a smaller number of effect sizes being analyzed in each outcome's meta-analysis, and this will obviously result in less power (Gleser & Olkin, 2009). In addition, when multiple univariate meta-analyses are performed (one for each outcome), it is not possible to test differences among the effect sizes as a function of the outcome type and of other moderator variables.

Instead of using any of these ad-hoc procedures, meta-analysts are encouraged to use one among a number of multivariate methods for handling within-study dependence (Raudenbush, Becker, & Kalaian, 1988). Each of these methods has several advantages (Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013). First, all data are analyzed when using one of the multivariate methods, and thus the approach does not reduce the size of the dataset being analyzed. Second, because these approaches are intended to handle dependency, the resulting inferences will typically be more valid than inferences based on models that ignore dependence. Third, these models allow for testing of more flexible hypotheses.

There are several methodological options for handling within-study dependence, including generalized least squares (GLS; see Becker, 2000), robust variance estimation (RVE; Hedges, Tipton, & Johnson, 2010), and the multilevel meta-analysis model (Van den Noortgate et al., 2013). The use of GLS involves an assumption that the covariances among the dependent effect sizes are known. This is not the case for the other two methods (RVE and the multilevel meta-analysis model). To calculate covariances among pairs of dependent effect sizes, the correlations among the outcome values are needed, which might be hard to collect in actual meta-analytic research, because of the lack of information reported in primary studies. Therefore, in the present study we focused

on RVE and multilevel meta-analysis as approaches to handle within-study dependence in effect sizes.

Robust variance estimation method The most recently suggested method for handling dependent effect sizes in meta-analysis is based on RVE (Hedges et al., 2010). Use of the RVE method for meta-analysis is intended to improve standard error estimation when pooling dependent effect sizes and estimating meta-regression models. One of the benefits is that the use of RVE does not require strict distributional assumptions. In addition, as we already noted, use of GLS involves an assumption that the covariances among the dependent effect sizes are known or can be estimated. This is not the case with RVE, for which the values of the covariances do not need to be exactly accurate.

The general fixed-effect meta-regression model can be written as follows:

$$\mathbf{d} = \hat{\beta}\mathbf{X} + \varepsilon \quad (1)$$

where $\mathbf{d}$ is the column vector of studies' effect size estimates, $\mathbf{X}$ is the design matrix, $\hat{\beta}$ is a vector of meta-regression model parameter estimates, and ε is the vector of residuals. To obtain the vector of parameter estimates, $\hat{\beta}$, RVE uses the following equation:

$$\hat{\beta} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}\mathbf{d}) \quad (2)$$

where $\mathbf{W}$ is the weight matrix, which is the inverse of the block-diagonal, across-studies covariance matrix ($\hat{\psi}$) containing blocks of the study-specific $\hat{\psi}_i$ matrices along its diagonal. Under RVE, the covariance matrix for the estimates of the vector of parameter estimates, $\hat{\beta}$, $V(\hat{\beta})_{RVE}$, uses the following formula:

$$V(\hat{\beta})_{RVE} = \left(\sum_{i=1}^k \mathbf{X}_i' \mathbf{W}_i \mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^k \mathbf{X}_i' \mathbf{W}_i \varepsilon_i \varepsilon_i' \mathbf{W}_i \mathbf{X}_i\right) \left(\sum_{i=1}^k \mathbf{X}_i' \mathbf{W}_i \mathbf{X}_i\right)^{-1}. \quad (3)$$

According to previous research (Tanner-Smith & Tipton, 2014), although the residuals' vector product, $\varepsilon_k \varepsilon_k'$, is a poor estimate of the observed effect sizes' covariance matrix, the resulting estimate of the parameter estimates' covariance matrix, $V(\hat{\beta})_{RVE}$, is unbiased for meta-analyses with a sufficiently large number of studies.

In addition to not requiring knowledge of the correlations between pairs of outcomes to calculate the covariances between effect size estimates and not requiring strict distributional assumptions, the RVE method does not involve strict requirements as to the choice of the weight matrix $\mathbf{W}$. Even though weights are used in the RVE method, the choice of weights is not as important to the method's functioning, in

contrast with GLS. With RVE, the choice of weights only impacts the efficiency of the parameter estimates, whereas with GLS, the choice of weights can result in biased values for the parameter estimates' covariance matrix, $V(\hat{\beta})_{GLS}$ (Tanner-Smith & Tipton, 2014). Thus, RVE provides a good replacement for GLS when handling dependence.

To examine the performance of RVE with respect to the number of studies and the number of effect sizes, several researchers have conducted simulation studies focused on meta-analysis of the standardized mean difference effect size (Hedges et al., 2010) and dichotomous outcome effect sizes (Tipton, 2013). These simulation studies have shown that RVE performed well when used for meta-analyzing the standardized mean difference, log odds ratio, log risk ratio, and risk difference effect sizes in scenarios involving conditions with varied numbers of studies, within-study sample sizes, numbers of effect sizes per study, within-study correlation, between-study variability, and population probability of dichotomous outcome effect sizes. Both of these studies suggested that at least ten studies are needed for calculating an average effect size, and at least 40 studies and five effect sizes per study (in total, 200 effect sizes) for meta-regression analyses.

To address these issues with small-sized meta-analyses, Tipton (2015) suggested that a small-sample-size correction be used with RVE. This small-sample-size correction involves a correction to the degrees of freedom used for the fixed-effect parameter estimates. The study suggested use of small sample size correction regardless of the number of studies involved in the meta-analysis and recommended that more stringent α levels be used if the adjusted degrees of freedom for a parameter estimate is less than 4.

Even though previous methodological research has investigated scenarios with multiple related effect sizes per study, most of these studies have examined the performance of RVE for estimating a single overall effect size across the multiple outcomes. However, there are scenarios in which applied researchers are interested in outcome-specific effect sizes when the related outcomes cannot be considered multiple measures of a single construct. In that case, parameter recovery for the necessary different parameterization of the fixed-effect parameters should be assessed in order to indicate the optimal conditions for stable results. Therefore, we conducted a simulation study to examine the performance of RVE (with small sample size correction) for estimating outcome-specific mean effect sizes under various conditions as follows:

$$\hat{\delta}_i = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e_i \quad (4)$$

However, because RVE is intended as an approach for adjusting standard error estimates, RVE only provides an overall measure of heterogeneity rather than specific random-effect variance component estimates (Tanner-Smith

& Tipton, 2014). Therefore, if the focus of research is not only on estimating fixed effects but also on estimating the separate random-effect variance estimates, then use of the multilevel modeling framework for meta-analysis should also be investigated.

Using the multilevel modeling framework for meta-analysis

Because meta-analysis involves the examination of data with a nested data structure, the multilevel model can also be used to handle dependent effect sizes in a meta-analysis (Van den Noortgate et al., 2013). The three-level meta-analysis model is an extension of the random-effect meta-analysis model (Borenstein et al., 2009). The three-level meta-analysis approach was proposed by Geeraert, Van den Noortgate, Grietens, and Onghena (2004) as a way of partitioning variability in effect sizes into the sampling variation for each effect size (Level 1), variation across multiple effect sizes within a study (Level 2), and variation across studies (Level 3).

At the first level of the multilevel model, sampling error is modeled for the effect size for outcome j from study k :

$$\hat{\delta}_{jk} = \beta_{jk} + e_{jk} \quad (5)$$

where e_{jk} is the sampling error residual, assumed to be normally distributed with a mean of zero and a known variance ($v_{\hat{\delta}_{jk}}$) (Van den Noortgate et al., 2013). Note that the indexing here matches that in related research (Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013, 2014).

At Level 2, the within-study dependence resulting from multiple effect sizes per study with dependence resulting from the measurement of related outcomes for the same sample is modeled as follows:

$$\beta_{jk} = \theta_{0k} + r_{jk} \quad (6)$$

where r_{jk} is the residual for outcome j within study k , typically assumed to follow a normal distribution with a mean of zero and variance (σ_r^2).

Finally, at the third level, the study-specific mean effect can be modeled as varying across studies around an overall mean effect size (γ_{00}), as is shown below:

$$\theta_{0k} = \gamma_{00} + u_{0k} \quad (7)$$

where u_{0k} is the residual for study k and it is typically assumed that residuals follow a normal distribution with a mean of zero and a variance (σ_u^2). In sum, the three-level model can be written as a single equation:

$$\hat{\delta}_{jk} = \gamma_{00} + u_{0k} + r_{jk} + e_{jk}. \quad (8)$$

The three-level meta-analysis model provides a single, summary effect size (γ_{00}) that is pooled across multiple

outcomes per study. Cheung (2015) explained that the three-level model is a special case of multivariate meta-analysis and requires an assumption that the degrees of heterogeneity for all effect sizes are similar. However, Van den Noortgate et al.'s (2013) study provided results that indicated that the three-level model works well even in scenarios in which this assumption is violated.

To use the model in Eq. 8, a researcher needs to have a valid rationale for pooling together the effect sizes for multiple related outcomes to form a single effect size. For example, in Abrami et al.' (2015) meta-analytic study, the authors examined the impact of interventions on the development of critical thinking skills, operationalized with three outcomes: achievement, content-specific critical thinking, and critical thinking dispositions. They collected 341 effect sizes from about 60 studies, which indicated that the primary studies reported multiple effect sizes for measuring an outcome, because many of the studies reported more than three effect sizes corresponding to the three outcomes. In their meta-analysis, Van der Kleij, Feskens, and Eggen (2015) examined the effectiveness of feedback in a computer-based learning environment. Their study focused on two types of outcomes: lower-order outcomes (e.g., measurements for recall and recognize contents) and higher-order outcomes (e.g., transfer). As a final example, Sheridan, Kim, Beretvas, Smith, and Park's (2017) meta-analysis gathered effect sizes for multiple outcomes that could be classified in three separate categories: academic achievement outcomes (that included effect sizes from measures of GPA, homework grades, and IQ scores), academic behaviors outcomes (that included effect sizes from measures of engagement scores, truancy scores, and time spent on homework), and social/behavior competence outcomes (which included effect sizes from measures of social competence, bullying, and peer relationships scores). Like the other meta-analysts we have mentioned, Sheridan et al. were interested in synthesizing the effects on these related but separate kinds of outcomes and did not want to consider them as effects on a single kind of outcome.

If researchers are interested in calculating outcome-specific effect sizes, this can be accomplished through the use of dummy-coded predictors (moderators). For example, if there are three related outcomes per study, the Level 1 model (see Eq. 5) will be the same. However, at the second level, each outcome's coefficient (β_{jk}) can be modeled as varying around each outcome-specific mean effect size, θ_{1k} , as follows:

$$\begin{aligned} \beta_{jk} = & \theta_{1k}(\text{Outcome})_{1k} + \theta_{2k}(\text{Outcome})_{2k} \\ & + \theta_{3k}(\text{Outcome})_{3k} + r_{jk} \end{aligned} \quad (9)$$

where $(\text{Outcome})_I$ is a dummy-coded variable indicating that the effect size is for outcome I (for $I = 1$ to 3, in this example), and r_{jk} is the between-outcome, within-study residual for the

effect size, which is typically assumed to follow a normal distribution with a mean of zero and a variance σ_r^2 . In the present study, we are assuming homoscedastic error variances within studies for the three outcomes' effect sizes. We recognize that this is a simplification, and future research can further extend this model to allow for within-study heterogeneity in effect sizes as a function of the outcome measure.

Finally, at the third (study) level, each outcome-specific coefficient (θ_{jk}) can be modeled as varying across studies around its outcome-specific overall mean effect size (r_{jk} , γ_{jk}), with the study-specific residual u_{jk} , as is shown below (for a three-outcome scenario):

$$\begin{aligned}\theta_{1k} &= \gamma_{10} + u_{1k} \\ \theta_{2k} &= \gamma_{20} + u_{2k} \\ \theta_{3k} &= \gamma_{30} + u_{3k}\end{aligned}\quad (10)$$

where u_{jk} is the study-specific residual for outcome j , and it is also assumed that the residuals follow a multivariate normal distribution with means of zero and the following variance–covariance matrix:

$$\begin{bmatrix} u_{1k} \\ u_{2k} \\ u_{3k} \end{bmatrix} \sim MVN \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} \sigma_{u11}^2 & \sigma_{u12} & \sigma_{u13} \\ \sigma_{u21} & \sigma_{u22}^2 & \sigma_{u21} \\ \sigma_{u31} & \sigma_{u32} & \sigma_{u33}^2 \end{pmatrix} \right]$$

If researchers (like Sheridan et al., 2017) want to estimate outcome-specific mean effect sizes, the variability of the effect sizes for each outcome, and the covariance among effects for each outcome type, then Eqs. 5, 9, and 10 need to be used. This three-level meta-analysis model can also be extended to a three-level mixed meta-analysis model by including predictors (moderator variables) at the respective levels to explain variability across outcomes and studies (see, e.g., Van den Noortgate et al., 2013).

In previous multilevel meta-analysis model studies (Van den Noortgate et al., 2013, 2014), the set of different outcomes have been assumed to be measures of a single, unidimensional construct. The present study will extend Van den Noortgate et al.'s (2013, 2014) research by assessing parameter recovery for the multilevel meta-analysis model when estimating outcome-specific effect sizes. Another difference between this and previous, related research is that in our parameterization, we are separating out the covariance parameters rather than (as in Van den Noortgate et al.'s, 2013, 2014, earlier parameterization) capturing the between-outcome within-study covariance as part of the study-level variance. In summary, under our parameterization in Eqs. 5, 9, and 10, including dummy-coded outcome indicators at the second level of the model allows for estimation of outcome-specific variances for the effect size for each outcome, as well as estimation of the study-level covariance between each pair of effect size residuals (σ_{uij*}).

To evaluate the three-level meta-analysis approach, Van den Noortgate et al. (2013) generated raw data using a two-

level multivariate model by varying a number of conditions. For each combination, each dataset was analyzed using a two-level multivariate model estimating only one effect size parameter synthesized across outcomes, a three-level model (see Eq. 8), and a traditional meta-analytic model ignoring effect sizes' within-study dependence. The resulting parameter estimates were compared across models. According to the results of their study, Van den Noortgate et al. (2013) found that use of the traditional meta-analytic model that ignores the dependence between effect sizes resulted in biased standard errors and problematic confidence interval coverage proportions. The traditional two-level random-effects model and two-level multivariate model were not found to work accurately, resulting in underestimated standard errors and overly narrow confidence interval coverage when the correlations among the outcomes were high. However, if there was dependence among outcomes, then the three-level meta-analysis model was found to work well for estimating the effect sizes' standard errors.

Building on the Van den Noortgate et al. (2013) study, another study was conducted that involved additional conditions (Van den Noortgate et al., 2014). The authors found corresponding results in this later study with a three-level model that recognized within-study dependence in effect sizes, resulting in unbiased standard error estimates and accurate confidence interval coverage for the effect size estimates. However, in this study the authors also assessed estimation of the total variance in the standardized mean difference effect size estimates (including between-study and between-outcome variance and the median per-outcome sampling error variance), and they found some underestimation bias. The authors suggested that this negative estimation bias could be explained by the relationships between the effect size estimate weights and the magnitude of the effect size estimates (Van den Noortgate et al., 2014).

A recent methodological study (Moeyaert et al., 2017) compared three approaches (averaging effect sizes per study with traditional two-level meta-analysis, vs. both three-level meta-analysis and RVE) for handling within-study dependence. The results indicated that the fixed effects' estimates and mean square errors were similar across all three approaches. However, use of averaging effect sizes per study and the traditional two-level meta-analysis model resulted in overly conservative test statistics, as compared with the RVE and three-level meta-analysis model results. In terms of the random-effect variance components, RVE estimates a total residuals' variance, whereas the three-level meta-analysis model estimates study-level and outcome-level variances. The overall variance estimated using RVE was not substantially biased. However, the level-three (between-study) residuals' variances were underestimated when using the three-level meta-analysis model in scenarios with fewer studies (here, 25 studies) in the meta-analysis. This recent study

compared the performance of RVE and of three-level meta-analysis only for a model that pooled estimates to provide a single, overall effect size across studies and outcomes.

This previous research focused on the use of multilevel meta-analysis model and RVE for handling within-study dependence, and it evaluated the two approaches across various conditions (Hedges et al., 2010; Moeyaert et al., 2017; Tanner-Smith & Tipton, 2014; Tipton, 2013, 2015; Van den Noortgate et al., 2013, 2014). However, some conditions remained to be considered. As we already mentioned, the studies all focused on synthesizing a single mean effect size across outcomes (see Eq. 8). However, in practice, if researchers are interested in calculating a summary effect size for each of the related outcomes (e.g., Sheridan et al., 2017), the equations that were suggested require a modification that distinguishes the effect sizes by using dummy-coded outcome indicators (see Eqs. 5, 9, and 10). This formulation also allows for the estimation of covariances between pairs of outcome-specific effect sizes. Even though the three-level meta-analysis model could provide more informative results than simply estimating a single, overall effect size, this parameterization increases the number of fixed-effect parameters and, more importantly, also increases the number of random-effect parameters that might require estimation. This more complex parameterization will likely impact the recommendations concerning the optimal sample size. Therefore, recovery of parameters under this more complex model should be assessed in order to help inform reasonable sample and study size guidelines for applied meta-analysts. Here, we first demonstrate the interpretation of parameters from the three-level meta-analysis and RVE models using data from an applied meta-analysis, described below. We also conducted a simulation study intended to assess parameter recovery using the three-level meta-analysis model versus RVE.

Review of quantitative meta-analysis characteristics

For the purposes of identifying the characteristics of quantitative meta-analysis, studies of meta-analysis published between 2010 and 2016 in the *Review of Educational Research* were examined. The information that was coded is presented in Table 1, which includes the number of primary studies (k) in each meta-analysis, the number of effect sizes for each primary study in the meta-analysis, and the sample size (n) for each primary study in the meta-analyses. Of the 110 studies published in the journal over this six-year period, 45 meta-analytic studies reported most of this information.

The mean number of studies (k) per meta-analysis was 30.3, the median was 26, and the mode was 30. The smallest 25% of the meta-analyses analyzed about 7 to 15 primary studies, whereas the smallest 50% of the meta-analyses

collected on average 26 primary studies. The 75th percentile for the number of studies per meta-analysis was approximately 34 primary studies. These findings support the idea that many meta-analyses collect a smaller number of primary studies than previous simulation studies have investigated (e.g., $k = 30$ and 60 in Van den Noortgate et al., 2013, and Van den Noortgate et al., 2014). Regarding the number of effect sizes, most of the primary studies in the reviewed meta-analyses reported one, two, or three. The median per-study sample size (n) was 176, with 25th and 75th percentile values of 66 and 608 participants per study, respectively. Thus, the within-study sample sizes were typically larger than those investigated ($n = 25$ or 50) in previous simulation studies (Van den Noortgate et al., 2013; Van den Noortgate et al., 2014). Most importantly, the majority of the applied meta-analyses (34 out of 45 studies, 75.6%) that we reviewed pooled effect sizes for each of several outcomes, because the outcomes could not meaningfully be considered measures of the same construct. In the meta-analyses, we found that some primary studies had reported multiple effect sizes that measured one single outcome. These applied meta-analysis characteristics helped inform the design of the simulation study that we conducted.

Real data study

Method

Dataset

Data from a meta-analysis of family–school interventions (Sheridan et al., 2017) were used to demonstrate results from using the three-level meta-analysis model versus RVE. The overall meta-analysis project gathered about 1,400 effect sizes from 223 studies published between 2001 and 2014 that focused on family–school partnership interventions (Sheridan et al., 2017). Even though the dataset includes various types of outcomes, the present study used data from three outcomes including measures of academic achievement, academic behaviors, and social/behavioral competence. This subset of the original meta-analytic database consisted of 276 effect sizes from 60 studies and 17 studies of 60 studies reported more than two outcomes. Academic achievement outcomes (165 effect sizes from 56 studies) included effect sizes on any of the following types of measures: grades, test scores, and academic competence scores. Academic behavior outcomes (34 effect sizes from 15 studies) encompassed effects on engagement, persistence, truancy, dropout, and time spent on homework. Finally, effect sizes capturing intervention effects on any of the following measures including social competence, bullying, peer relationship, behavioral regulation, externalizing concern, emotional regulation, and attention regulation were considered social/behavioral competence outcome effect sizes (77 effect sizes 29

Table 1 Characteristics of applied meta-analyses in the *Review of Educational Research* journal, 2010–2016

Characteristic	Statistic			
	Range (min; max)	Mean	Median	Percentile (25; 50; 75%)
Number of studies per meta-analysis (k)	7; 110	30.0	26	15; 26; 34
Number of effect sizes per primary study	1; 70	3.2	2	1; 2; 3
Sample size per primary study (n)	1; 2,925,005	20,721	176	66; 176; 608

min = minimum; max = maximum.

studies). Note that this real-data analysis is intended solely to demonstrate use of the three-level meta-analysis model and RVE model for synthesizing multiple effects on separate but related outcomes simultaneously. The results are not intended to be used for making broader substantive inferences about parent involvement interventions and the effects.

Estimating models

To handle the multiple outcomes and multiple effect sizes within a study, the three-level random-effect meta-analysis model (see Eqs. 5, 9, and 10) and RVE model (see Eq. 4) were used to provide three outcome-specific pooled estimates of the effect of parent involvement interventions on academic achievement, academic behavior and social/behavioral competence outcomes separately (while modeling the dependence in related outcomes' effect sizes). Meta-analysts interested in separate pooled effect by outcome would use this dummy-coding to capture each pooled effect while modeling the effect within-study dependence in these effect sizes. In addition, to mimic previous studies, we also estimated a single overall effect size across outcomes using the three-level meta-analysis model and RVE to demonstrate parameter interpretation and to compare the results. To demonstrate differences in the resulting parameters that are estimated, we have used both models (i.e., pooling a single effect and pooling three separate effects with one per outcome type). Note that in the simulation study, we are investigating how well the per outcome effect size model can be estimated. Others (e.g., Hedges et al., 2010; Moeyaert et al., 2017; Van den Noortgate et al., 2013, 2014) have already explored estimation of the single effect size model.

Results

Fixed-effect estimates

Table 2 compares the three-level meta-analysis model and RVE estimates using the parent involvement intervention effects on the academic achievement, academic behavior, and social/behavioral competence outcomes. For the three-level meta-analysis model, the pooled effect on academic achievement outcomes was statistically significant ($\hat{\delta} = 0.30$, $SE =$

0.06, $p < .05$), as was the pooled effect of parent involvement interventions on social/behavioral competence outcomes ($\hat{\delta} = 0.26$, $SE = 0.06$, $p < .05$). However, the pooled effect size estimate for academic behavior outcomes was based on a smaller number of effect size estimates and was not statistically significant ($\hat{\delta} = 0.29$, $SE = 0.15$, $p > .05$). The results of RVE had similar results, with statistically significant effects on academic achievement outcomes ($\hat{\delta} = 0.33$, $SE = 0.06$, $p < .05$) and social/behavioral competence outcomes ($\hat{\delta} = 0.25$, $SE = 0.08$, $p < .05$), although the effect on academic behavior outcomes was not statistically significant ($\hat{\delta} = 0.24$, $SE = 0.28$, $p > .05$). In models pooling across the three kinds of outcomes, the overall effect size was statistically significant under the three-level meta-analysis model ($\hat{\delta} = 0.29$, $SE = 0.05$, $p < .05$) and under RVE ($\hat{\delta} = 0.25$, $SE = 0.05$, $p < .05$). However, the overall effect size is less informative and less valid in this kind scenario, in which researchers wished to consider the outcomes as measuring different rather than a single unitary construct.

Variance and covariance component estimates

Table 2 also includes the estimates of Level 2 and Level 3 residuals' variance and covariance parameters under the three-level meta-analysis model. The variability between outcomes within a study (σ_r^2) and the variability across studies (σ_u^2) were estimated to range from 0.09 to 0.30 (see Table 2). The Level 3 between-study variability ranged in value from 0.07 to 0.30, with more variability being estimated between studies in effects on academic behavior ($\sigma_{u2}^2 = 0.30$) than for academic achievement outcomes ($\sigma_{u1}^2 = 0.10$) or social-behavioral competence outcomes ($\sigma_{u3}^2 = 0.07$). Also, the covariance between outcomes (σ_{uij*}) was estimated to range from 0.07 to 0.12 (see Table 2). On the basis of the variance and covariance values, the study-level correlations among effect sizes per outcome were calculated, with the following values resulting: between academic achievement and academic behavior, $r = .69$; between academic achievement and social/behavioral competence, $r = .96$; and between academic behavior and social/behavioral competence effect sizes, $r = .48$. As we mentioned above (Tanner-Smith & Tipton, 2014), because RVE focuses on estimating fixed effects and

Table 2 Parameter and standard error (*SE*) estimates for three-level and RVE models of family–school intervention effects for the multiple outcomes and overall mean effect size models

	Outcome-Specific Model				Overall Model			
	Three-Level		RVE		Three-Level		RVE	
Parameter Estimates	Est.	(<i>SE</i>)	Est.	(<i>SE</i>)	Est.	(<i>SE</i>)	Est.	(<i>SE</i>)
Fixed-Effect Estimates								
<i>AcadAch</i> (γ_{10})	0.30*	(0.06)	0.33*	(0.06)	0.29*	(0.05)	0.25*	(0.05)
<i>AcadBeh</i> (γ_{20})	0.29	(0.15)	0.24	(0.28)				
<i>SocBeh</i> (γ_{30})	0.26*	(0.06)	0.25*	(0.08)				
Random-Effect Variance			0.40				0.45	
Variance at Level 2 (σ_v^2)	0.09	(0.02)			0.09	(0.02)		
Variance at Level 3					0.13	(0.03)		
<i>AcadAch</i> ($\sigma_{u,11}^2$)	0.10	(0.04)						
<i>AcadBeh</i> ($\sigma_{u,22}^2$)	0.30	(0.19)						
<i>SocBeh</i> ($\sigma_{u,33}^2$)	0.07	(0.02)						
Covariance at Level 3								
<i>AcadAch</i> , <i>AcadBeh</i> ($\sigma_{u,12}^2$)	0.12	(0.14)						
<i>AcadAch</i> , <i>SocBeh</i> ($\sigma_{u,13}^2$)	0.08	(0.04)						
<i>AcadBeh</i> , <i>SocBeh</i> ($\sigma_{u,23}^2$)	0.07	(0.09)						

SE = standard error; *AcadAch* = academic achievement; *AcadBeh* = academic behavior; *SocBeh* = social/behavioral competence. * $p < .05$

adjusted standard errors, only between-study variability ($\tau^2 = 0.40$) was reported for the outcome-specific mean effect size model. Also, when estimating the overall mean effect size under RVE, the between-study variability value ($\tau^2 = 0.45$) was similar. When the three-level meta-analysis model was used to estimate the overall mean effect size, the value of the Level 2 variance ($\sigma_r^2 = 0.09$) was the same as the corresponding value when estimating the outcome-specific average effect sizes ($\sigma_r^2 = 0.09$). On the other hand, the value of the Level 3 variance estimate under the three-level meta-analysis model estimating a single, overall mean effect size ($\sigma_u^2 = 0.13$) was around the middle of the three variances (one for each outcome), as is shown in Table 2. Still, the variability of effect sizes at the outcome and study levels across outcomes could be hard to interpret practically.

In addition to demonstrating how model parameter estimates can be interpreted, the real-data analysis results also helped provide realistic values for generating parameters in the simulation study, described in the next section.

Simulation study

Method

Conditions

Study sample sizes According to our review of applied meta-analyses (see Table 1), the median per-primary-study sample

size was approximately 180, indicating that many primary studies had at least 180 participants. According to previous research, however, the three-level model and RVE worked well even with samples as small as 20 (Tipton, 2013) or 25 (Van den Noortgate et al., 2013). Therefore, in the present study we examined a reasonably small per-study sample size condition ($n = 25$). Because it is commonly the case that each study has a different sample size, however, it is important to evaluate parameter recovery with the three-level meta-analytic model when there are unbalanced per-study sample sizes. Also, according to the characteristics of the meta-analyses reviewed here, the distribution of sample sizes is slightly positively skewed. We thus generated meta-analytic data with unbalanced per-study sample sizes, with sample sizes randomly sampled from the following lognormal distribution $n \sim [6 + LnN(2.7, 0.7)]$, matching the distribution used in previous related research (Van den Noortgate et al., 2014).

Number of outcomes per study It is also common for each primary study to report effects for a different number of outcomes. Therefore, we investigated conditions with varying numbers of outcomes per study, to compare parameter recovery in scenarios with balanced data, in which each study reports all three outcomes' effect sizes, to scenarios with unbalanced data in terms of the outcomes each study reported. To generate data with unbalanced numbers of outcomes reported, we first generated data with three effect size estimates per study, and then we randomly deleted one or two effect sizes per study. We accomplished this using the following

procedure: For each study, we sampled a number from a uniform distribution ranging from 0.5 to 3.5. The resulting number was rounded to the closest integer, so that each study reported effect size estimates for one, two, or three outcomes, with two being the average. Also, if the study had two outcomes, the first (Outcome 1 in Eq. 10) and the second (Outcome 2 in Eq. 10) outcomes remained, and if the study had only one outcome, only the effect size for the first outcome was reported. Note that future research can extend the research we have conducted and explore other patterns of missing effect sizes.

Number of effect sizes per outcome It is also reasonable to expect that primary studies would assess intervention effects on more than one measure of an outcome and that the number of effect size estimates per outcome would not be the same across studies. Therefore, we examined both scenarios with a balanced number of effect size estimates per outcome and study (with consistently five effect size estimates per outcome and study) and scenarios with unbalanced data, with numbers of effect sizes per outcome (and study) ranging from one to five. Thus, a single primary study could report anywhere from one to 15 effect sizes, depending on the condition. When sampling the number of effect sizes per outcome and study, we sampled per outcome from a uniform distribution ranging from 0.5 to 5.5 and rounded the sampled number to the closest integer (to allow one to five effect sizes to be reported per outcome per study).

Number of studies Many methodological studies investigating the multilevel model (e.g., Maas & Hox, 2005; Van den Noortgate & Onghena, 2003; Viechtbauer, 2005) have indicated that analysis of data with a smaller number of Level 3 units might result in underestimated standard errors and inflated Type I error rates. Thus, we followed the recommendations of Kreft and De Leeuw (1998) by using more than 30 units at the highest (study) level. However, according to our review of applied meta-analytic studies, many meta-analyses do not collect more than 30 studies for their analysis. In simulation studies with RVE (Hedges et al., 2010; Moeyaert et al., 2017; Tipton, 2013), researchers have examined the performance of RVE using from one to 50 studies per meta-analysis and have concluded that RVE works well in scenarios with 40 studies and five effect sizes per study if the meta-regression model included moderators (Tanner-Smith & Tipton, 2014). Therefore, we also investigated how well multilevel meta-analysis parameter estimates are recovered with a smaller number of units at the highest level. The number of studies per meta-analysis was thus also a condition that we varied ($k = 20$ and 50), to evaluate the robustness of estimation of the multilevel meta-analysis model.

Ratio of variances for Sampling-error : Between-outcome within-study : Between-study variance values In their study, Van den Noortgate et al. (2014) found that 56% of the total variance in effect size estimates was associated with sampling error (Level 1), 33% was within-study (between-outcome) variance (Level 2), and 11% was between-study variance (Level 3). Therefore, in the present study we used a similar ratio, with 5:3:2 for the ratio of the sampling-error variance to between-outcome within-study variance to between-study variance, respectively. In addition, another three-level meta-analysis study using real data (Assink et al., 2015) reported 14 different results for three-level meta-analyses conducted for each of 14 outcomes. On the basis of the median values from the 14 ratios among the three variances, we calculated a corresponding ratio of approximately 2:3:5. Therefore, we examined conditions with 5:3:2 and with 2:3:5 as the true ratios of the three levels' variances.

Study-level covariance The value for the covariance between pairs of Level 3 effect size residuals was based on the results of the real-data analysis reported in Van den Noortgate et al.'s (2014) study. According to that study, the correlation between pairs of effect sizes was .11, which is relatively small. In one set of conditions in the present study, the correlations' values are relatively weaker, matching the actual values estimated in Van den Noortgate et al.'s (2014) study. In the second set of conditions, relatively stronger correlations among effect sizes were generated. Because, in reality, it is very unlikely that the correlation among the effect sizes for outcomes would be the same across all pairs, we also varied the values of the coefficients within the correlation matrix for each of these two conditions. The generating values for the correlation matrices for the two conditions were as follows:

$$\begin{bmatrix} 1 & & \\ .1 & 1 & \\ .3 & .5 & 1 \end{bmatrix} \text{ and } \begin{bmatrix} 1 & & \\ .3 & 1 & \\ .5 & .7 & 1 \end{bmatrix}.$$

Effect size The true effect size was unique for each of the three outcomes, with values of $\gamma_1 = 0.2$, $\gamma_2 = 0.4$, $\gamma_3 = 0.6$ (see Eq. 10), which were chosen to represent small to medium effect size values (Cohen, 1988).

In summary, we generated data to fit the combination of parameter and design condition values for each of 64 scenarios [Equality of Studies' Sample Sizes (2) $\times$ Equality of Counts of per-Outcome and Study Effect Sizes (2) $\times$ Equality of Counts of per-Study Outcomes (2) $\times$ Number of Studies (2) $\times$ Ratios of Variances (2) $\times$ Study-Level Correlation Between Outcomes' Effect Sizes (2)]. For each condition, we generated 1,000 datasets, resulting in a total of 64,000 datasets that were generated to fit the three-level meta-analysis model found in Eqs. 5, 9, and 10. For all analyses

using the three-level meta-analysis model, we used the SAS 9.3 software (SAS Institute, Cary, NC, 2011) with the restricted maximum likelihood (REML) estimation procedure that had been used in previous, related research (Van den Noortgate et al., 2013). For RVE, Tipton's (2015) small-sample bias correction was used to estimate the outcome-specific pooled effect sizes and their standard errors. The robumeta package (see Fisher & Tipton, 2014) was used for RVE. We tallied the percentages of nonconvergent solutions when estimating the correct three-level meta-analysis model, and when necessary we ran additional replications to ensure that we had 1,000 converged solutions for each condition.

Analyses

Relative parameter bias Relative parameter bias (Hoogland & Boomsma, 1998) was used to provide information about parameter recovery. The formula for relative parameter bias is shown below.

$$RPB(\hat{\theta}_j) = \frac{\bar{\hat{\theta}}_j - \theta_j}{\theta_j} \quad (1)$$

where θ_j represents the j th parameter's true value and $\hat{\theta}_j$ is the mean of the estimated parameters ($\bar{\hat{\theta}}_j$) averaged across 1,000 replications for each condition. Hoogland and Boomsma suggested that if a parameter's relative bias exceeds 0.05 in magnitude, then substantial bias can be inferred.

Relative standard error bias The relative standard error bias was calculated for the standard errors of each of the outcome-specific effect size estimates. The formula for relative standard error bias is shown below:

$$RSEB(s_{\hat{\theta}_j}) = \frac{\bar{\hat{S}}_{\hat{\theta}_j} - \hat{S}_{\hat{\theta}_j}}{\hat{S}_{\hat{\theta}_j}} \quad (2)$$

where $\bar{\hat{S}}_{\hat{\theta}_j}$ is the mean of the standard error estimates for estimated parameter $\hat{\theta}_j$ and $\hat{S}_{\hat{\theta}_j}$ is the standard deviation of $\hat{\theta}_j$. Because the present study involved generating a large number of datasets for each condition, it can be assumed that the observed standard deviation of the estimated parameters was a reasonable approximation of the standard error of the relevant sampling distribution for the parameter. According to Hoogland and Boomsma (1998), if the relative standard error bias exceeds 0.10, this indicates substantial bias.

Root mean squared error Finally, the root mean squared error was calculated for each of the outcome-specific effect size estimates. Because we considered deviance from parameter

to estimates $[\sqrt{(\sum_1^N (\hat{\theta}_j - \theta_j)^2)/N}]$, the root mean squared error was able to provide information about both efficiency and bias.

Confidence interval coverage rates The coverage proportions of the confidence intervals were calculated for fixed-effect parameters. Because the true parameter values for the fixed-effect parameters were not zero, the proportions of confidence intervals that contained the generated effect size (0.2, 0.4, or 0.6) were tallied across 1,000 replications per condition.

Results

Convergence rates

A 100% convergence rate was found when using RVE across all conditions. However, some serious convergence issues were encountered when estimating the three-level meta-analysis model under the conditions examined here (see Table 3). Several of the conditions that we manipulated were found to impact convergence rates, including within-study sample size, the number of outcomes per study, the number of effect size estimates per outcome and study, the number of studies per meta-analysis, and the ratio of the variances across levels. The results in Table 3 are summarized across the Level 3 covariance generating value conditions, because the value of the Level 3 covariance values did not impact the results. Not surprisingly, the size of the meta-analytic database (number of studies, whether studies reported effect sizes for all three outcomes, and whether each study reported five or fewer effect sizes per outcome) affected the convergence rates, such that the more effect sizes were being analyzed, the better the convergence rates. The within-study sample size, N , also affected the convergence rates, such that when every study used the same sample size, convergence rates were better (76.8%) than when the per-study sample sizes varied (70.5%). The Level 1 through Level 3 variance ratios also had an effect, although the effect interacted with the effect of the number of studies being meta-analyzed. Convergence was worst, sometimes with unacceptably low values, in conditions with only 20 (rather than 50) studies. When the meta-analysis involved the 5:3:2 variance ratios (for sampling-error : within-study/between-outcome : between-study variances), convergence rates were particularly low (47.4%), as compared to a still-too-low rate for the 2:3:5 conditions (with an average convergence rate of 78.3%). In summary, if the scenario included all of the worst-performing conditions (the number of studies was 20, the number of outcomes per study varied, and the number of effect sizes per outcome varied), then the lowest convergence rates (as low as 8.8% or 13.9%) resulted (see Table 3).

Table 3 Convergence rates by condition for the three-level meta-analysis model and RVE

Condition					Convergence Rate	
# of Studies	# Outcomes per Study	# Effect Sizes per Outcome ^a	Sample Size per Study	Variance Ratio	Three-Level	RVE
50	1–3	1–5	Different	2:3:5	84.7%	100.0%
				5:3:2	37.1%	100.0%
			Same	2:3:5	88.4%	100.0%
				5:3:2	50.4%	100.0%
		Five	Different	2:3:5	99.5%	100.0%
				5:3:2	84.2%	100.0%
			Same	2:3:5	100.0%	100.0%
				5:3:2	95.6%	100.0%
	Three	1–5	Different	2:3:5	99.0%	100.0%
				5:3:2	66.7%	100.0%
			Same	2:3:5	99.6%	100.0%
				5:3:2	84.5%	100.0%
		Five	Different	2:3:5	100.0%	100.0%
				5:3:2	99.5%	100.0%
			Same	2:3:5	100.0%	100.0%
				5:3:2	99.7%	100.0%
20	1–3	1–5	Different	2:3:5	38.6%	100.0%
				5:3:2	8.8%	100.0%
			Same	2:3:5	44.8%	100.0%
				5:3:2	13.9%	100.0%
		Five	Different	2:3:5	75.4%	100.0%
				5:3:2	36.1%	100.0%
			Same	2:3:5	90.7%	100.0%
				5:3:2	72.2%	100.0%
	Three	1–5	Different	2:3:5	86.9%	100.0%
				5:3:2	29.4%	100.0%
			Same	2:3:5	90.1%	100.0%
				5:3:2	44.0%	100.0%
		Five	Different	2:3:5	100.0%	100.0%
				5:3:2	82.4%	100.0%
			Same	2:3:5	100.0%	100.0%
				5:3:2	92.8%	100.0%

The results are summarized across the Level 3 covariance value conditions because this design factor did not impact the results. Variance Ratio = ratio of variances for sampling-error : between-outcome within-study : between-study variance values. ^a # Effect Sizes per Outcome = number of effect sizes per outcome and study.

Fixed-effect parameter and standard error estimates

The relative parameter bias was not substantial for estimates of the outcome-specific pooled effect size (ranging from – 4.6% to 4.4%) and the corresponding standard errors (ranging from – 10.0% to 4.1%) when using RVE across conditions. In terms of the three-level meta-analysis model estimates, the relative bias for fixed-effect parameter estimates was negative but was not substantial (i.e., the relative parameter bias was less than 5%), ranging from 0% to – 3.1%.

In terms of the relative standard error bias (*RSEB*) for the three fixed-effect parameter estimates, most of the multilevel model estimates did not exceed 10% in conditions with 50 studies per meta-analysis study (see Table 4). (Note that again, results in Table 4 are summarized across Level 3 covariance generating values, because this condition did not substantially impact the results). For each of the few conditions (six out of 16) in which substantial bias was found when $k = 50$, the variance ratio was 5:3:2. When the number of studies per meta-analysis was only 20, a larger proportion of the standard error estimates were

Table 4 Relative standard error bias for each of the three fixed-effect parameter estimates by condition using the three-level meta-analysis model and RVE

Condition					Three-Level Model			RVE		
# of Studies	# Outcomes per Study	# Effect Sizes per Outcome ^a	Sample Size per Study	Variance Ratio	γ_{10}	γ_{20}	γ_{30}	γ_{10}	γ_{20}	γ_{30}
50	1–3	1–5	Different	2:3:5	–.028	.017	–.003	–.061	–.032	–.071
				5:3:2	.023	.056	–.036	–.049	–.056	–.050
			Same	2:3:5	–.026	.038	–.011	–.056	–.052	–.057
				5:3:2	–.016	.040	–.060	–.058	–.053	–.065
		Five	Different	2:3:5	.001	–.011	–.063	–.022	–.034	–.025
				5:3:2	–.028	–.026	–.178	–.083	–.049	–.066
			Same	2:3:5	–.021	.011	–.077	–.035	–.015	–.044
				5:3:2	–.032	–.011	–.217	–.025	–.065	–.057
	Three	1–5	Different	2:3:5	.002	–.030	–.010	–.010	–.026	–.035
				5:3:2	–.012	–.001	.006	.000	–.003	–.012
			Same	2:3:5	–.019	–.035	–.007	.000	.007	–.017
				5:3:2	.023	.026	–.045	–.042	–.040	–.036
		Five	Different	2:3:5	–.010	–.046	–.010	–.014	.013	.011
				5:3:2	–.277	–.306	–.268	–.037	.000	–.020
			Same	2:3:5	.005	–.033	–.003	–.037	–.044	–.017
				5:3:2	–.041	–.126	–.111	–.004	.002	–.026
20	1–3	1–5	Different	2:3:5	–.092	–.059	–.090	.008	–.027	–.025
				5:3:2	–.415	–.410	–.295	–.039	–.020	–.018
			Same	2:3:5	–.138	–.028	–.072	–.025	.002	.004
				5:3:2	–.358	–.359	–.268	–.006	–.038	–.020
		Five	Different	2:3:5	–.048	–.038	–.128	–.017	–.022	–.045
				5:3:2	–.454	–.455	–.300	–.030	–.027	.004
			Same	2:3:5	–.031	–.024	–.124	–.050	–.005	–.019
				5:3:2	–.143	–.138	–.152	–.024	.010	–.023
	Three	1–5	Different	2:3:5	–.017	.018	–.024	–.027	–.004	–.020
				5:3:2	–.309	–.354	–.290	–.019	–.023	–.021
			Same	2:3:5	–.020	.025	–.003	–.003	–.016	–.014
				5:3:2	–.157	–.186	–.183	–.020	–.011	–.017
		Five	Different	2:3:5	.008	.032	.008	.033	–.025	–.032
				5:3:2	–.153	–.091	–.108	–.010	–.033	–.032
			Same	2:3:5	.002	.032	–.004	.005	–.006	.014
				5:3:2	–.046	–.015	–.042	.009	.007	.003

The results are summarized across the Level 3 covariance value conditions because this design factor did not impact the results. The relative standard error bias values are bolded when the values are larger than 0.10. Variance Ratio = ratio of variances for sampling-error : between-outcome within-study : between-study variance values. ^a # Effect Sizes per Outcome = number of effect sizes per outcome and study.

negatively biased, again predominantly when the variance ratio was 5:3:2. When $k = 20$ and the variance ratio was 5:3:2, the *RSEB* was - 23.7% across fixed-effect parameter estimates, and only - 3.4% for the 2:3:5 variance ratio conditions. The size of the meta-analytic database again impacted the estimation—here, of standard errors. When k was 20, the *RSEB* (- 7.8%) was acceptable when each study reported all effect size estimates for each of the three outcomes, as compared with the average *RSEB* of - 19.2% when studies reported effect size estimates for one to three of the

outcomes. Similarly, when each study provided five effect size estimates for each outcome reported, the *RSEB* average was better (- 10.1%) than when each study reported from one to five effect size estimates per outcome (- 17.0%).

Random-effect variance–covariance component estimates

The RVE provided a measure of variability that quantified the total variance across levels in the model (Tanner-Smith & Tipton, 2014). Therefore, the multilevel model estimates of

the between-outcome within-study (σ_r^2) and between-study (σ_u^2) variance parameters were summed in order to provide a total variance for the multilevel model, and the results were compared with the total variance reported in RVE. No substantial relative parameter bias (*RPB*) was found for RVE estimates of the total variance (*RPB* ranged from -3.4% to 2.5%) in all but 16 out of the 64 conditions. The *RPB* was substantially negatively biased (*RPB* = -5.4%) in the 16 conditions in which the true variance ratio was 2:3:5 and the true Level 3 correlations were large (.3, .5, and .7 for the three correlations).

Table 5 contains the *RPB* values for the Level 2 and Level 3 random-effect variance component estimates by conditions, estimated using the three-level meta-analysis model. Again, the Level 3 covariance values did not impact the results, which are summarized across values for this condition. As can be seen in Table 5, all of the Level 2 variance estimates (σ_r^2) were substantially negatively biased, with values exceeding .05. The bias was slightly worse when meta-analyzing 20 (*RPB* = -13.5%) versus 50 (*RPB* = -11.0%) studies. When all studies reported effect size estimates for three outcomes, this bias was slightly lower (*RPB* = -11.5%) than in scenarios in

Table 5 Relative parameter bias for Level 2 and Level 3 variance estimates by condition using the three-level meta-analysis model

Condition					Variance Estimated			
# of Studies	# Outcomes per Study	# Effect Sizes per Outcome ^a	Sample Size per Study	Variance Ratio	σ_r^2	$\sigma_{\mu l}^2$	σ_{u2}^2	σ_{u3}^2
50	1–3	1–5	Different	2:3:5	–.130	–.061	–.054	–.033
				5:3:2	–.098	–.087	–.025	.180
			Same	2:3:5	–.092	–.035	–.028	.009
				5:3:2	–.147	–.008	.025	.228
		Five	Different	2:3:5	–.130	–.060	–.053	–.039
				5:3:2	–.059	–.121	–.102	–.007
			Same	2:3:5	–.097	–.039	–.034	–.019
				5:3:2	–.136	–.042	–.034	.031
	Three	1–5	Different	2:3:5	–.135	–.066	–.058	–.062
				5:3:2	–.073	–.080	–.093	–.072
			Same	2:3:5	–.098	–.038	–.032	–.041
				5:3:2	–.150	.001	.008	.013
		Five	Different	2:3:5	–.130	–.059	–.053	–.059
				5:3:2	–.055	–.123	–.110	–.108
			Same	2:3:5	–.096	–.039	–.034	–.041
				5:3:2	–.136	–.045	–.039	–.037
20	1–3	1–5	Different	2:3:5	–.158	–.072	–.011	.140
				5:3:2	–.188	.153	.250	.909
			Same	2:3:5	–.107	–.049	.015	.180
				5:3:2	–.237	.205	.330	.436
		Five	Different	2:3:5	–.138	–.062	–.038	.119
				5:3:2	–.100	–.052	.007	.411
			Same	2:3:5	–.097	–.042	–.032	.183
				5:3:2	–.158	–.001	.026	.348
	Three	1–5	Different	2:3:5	–.142	–.063	–.053	–.060
				5:3:2	–.096	.062	.088	.095
			Same	2:3:5	–.096	–.033	–.039	–.040
				5:3:2	–.180	.114	.136	.117
		Five	Different	2:3:5	–.136	–.061	–.059	–.078
				5:3:2	–.083	–.075	–.065	–.092
			Same	2:3:5	–.096	–.039	–.042	–.060
				5:3:2	–.145	–.003	–.027	–.031

The results are summarized across the Level 3 covariance value conditions because this design factor did not impact the results. The relative parameter bias values are bolded when the values are larger than 0.05. Variance Ratio = ratio of variances for sampling-error : between-outcome within-study : between-study variance values. ^a # of Effect Sizes per Outcome = number of effect sizes per outcome and study.

which studies reported effect sizes for one to three outcomes ($RPB = -13.3\%$). When studies consistently reported five effect size estimates per outcome, the RPB was again slightly better ($RPB = -11.2\%$) than when studies reported anywhere from one to five effect size estimates per outcome ($RPB = -13.3\%$). The RPB differences as a function of the degree of balance in studies' sample sizes and the variance ratio values were only just over 1%. These condition effects were only slightly greater when the number of studies per meta-analysis was 20 versus 50.

In terms of the Level 3 residuals' variance components, the RPB was consistently worse when 20 studies were being meta-analyzed than for the $k = 50$ conditions. Within the $k = 20$ conditions, in scenarios in which at least one among the number of outcomes, effect sizes, or sample size per study was unbalanced, then the RPB for the third-outcome residuals' variance estimates was typically worse than those for the first and second outcome residuals' variances (see Table 5). Although estimates of the Level 3 residuals' variance components for Outcome 1 were not substantially biased, estimates of the residuals' variances for Outcomes 2 (σ_{u2}^2) and 3 (σ_{u3}^2) were biased. The degree of the bias was less for the second than for the third outcome, with very substantial bias found when only 20 studies were being meta-analyzed. With a little variability in the $k = 20$ conditions, the direction of most of the bias in the σ_{u3}^2 estimates was positive.

For the $k = 20$ conditions and the third outcome measure, the RPB for estimates of σ_{u3}^2 was predominantly positive when effect size estimates were reported for one to three outcomes per study ($RPB = 34.1\%$), with lower bias that was even sometimes negative ($RPB = -1.9\%$) when studies reported effect size estimates for all three outcomes. Similarly, when studies consistently reported five effect sizes per outcome, the RPB for the σ_{u3}^2 estimates was better ($RPB = 10.0\%$) than when studies reported different numbers of effect sizes (one to five) per outcome ($RPB = 22.2\%$). Also within the $k = 20$ conditions, the variance ratio value again impacted the bias of σ_{u3}^2 estimates. When the variance ratio was 5:3:2, a severe positive bias ($RPB = 27.4\%$) was found, as compared with scenarios in which the variance ratio was 2:3:5 ($RPB = 4.8\%$). That is, in the $k = 20$ conditions, σ_{u3}^2 estimates were particularly biased when the higher-level true variance ratio was small (i.e., in the 5:3:2 variance ratio conditions; $RPB = 27.4\%$), when studies reported different numbers of effect sizes (one to five) per outcome ($RPB = 22.2\%$), and when effect size estimates were reported for one to three outcomes per study ($RPB = 34.1\%$),

Finally, in terms of the Level 3 random-effect covariance estimates, the RPB was mostly substantially negative, especially the covariance between residuals for the

first and second outcomes (σ_{u12}) (see Table 6). When only 20 rather than 50 studies' results were included in the simulated meta-analysis, the RPB results differed more as a function of the same conditions that impacted the other results. The negative bias was worse when the ratio of variances was 5:3:2 versus 2:3:5 and when the number of effect sizes per outcome and the number of outcomes per study were not the same across studies.

Regarding the root mean squared error ($RMSE$) for three fixed-effect parameter estimates whereas the mean $RMSE$ using the multilevel model was 0.027, the mean $RMSE$ using RVE was 0.020 across three fixed effects and across conditions. Mostly, the $RMSE$ using the multilevel model was larger than the $RMSE$ using RVE, but—only within the $k = 50$ conditions and when the number of effect sizes per outcome and the number of outcomes per study were the same across studies—the $RMSE$ values using multilevel model were smaller than the values using RVE.

Confidence interval coverage rates

Under RVE, the coverage rates for confidence intervals was ranged from 93.8% to 96.2%. The results when using the three-level meta-analysis model were affected by the number of studies, the number of outcomes per primary study, and the number of effect sizes per outcome. For example, when the number of primary studies in meta-analyses was 20, with effect sizes for one to three outcomes and one to five effect size estimates per outcome and study, the confidence interval coverage rates ranged from 67.8% to 89.7%, indicating poor accuracy. On the other hand, in scenarios with more primary studies ($k = 50$), all three outcomes reported, and all five effect sizes per outcome reported, the confidence interval coverage rates ranged from 93.7% to 94.8%.

Discussion

In the present study we investigated use of the multilevel meta-analysis model and the RVE model to handle within-study effect size dependence, with a focus on estimating outcome-specific effect sizes, rather than overall effect sizes summarized across related outcomes. RVE provides robust sampling variance estimates by using cross product of the residuals instead of covariance matrix among effect sizes within each study (Hedges et al., 2010). On the other hand, the three-level meta-analysis model that was used handles the sampling variation for each effect size at Level 1, variation in effect sizes for multiple outcomes within studies at Level 2, and variation across studies at Level 3. Although previous studies using the multilevel meta-analysis model (Moeyaert et al., 2017; Van den Noortgate et al., 2013, 2014) and RVE

Table 6 Relative parameter bias for each of the three Level 3 covariance parameter estimates by condition using the three-level meta-analysis model

Condition					Covariance Estimated		
# of Studies	# Outcomes per Study	# Effect Sizes per Outcome ^a	Sample Size per Study	Variance Ratio	$\sigma^2_{u.12}$	$\sigma^2_{u.13}$	$\sigma^2_{u.23}$
50	1–3	1–5	Different	2:3:5	–.173	–.090	–.085
				5:3:2	–.304	–.203	–.211
			Same	2:3:5	–.088	–.045	–.050
				5:3:2	–.091	–.120	–.090
	Three	Five	Different	2:3:5	–.111	–.059	–.055
				5:3:2	–.196	–.051	–.073
			Same	2:3:5	–.106	–.044	–.045
				5:3:2	–.169	–.025	–.039
		1–5	Different	2:3:5	–.030	–.051	–.054
				5:3:2	–.067	–.095	–.139
			Same	2:3:5	–.008	–.036	–.042
				5:3:2	.078	.010	–.008
		Five	Different	2:3:5	–.045	–.039	–.051
				5:3:2	–.082	–.068	–.082
			Same	2:3:5	–.036	–.030	–.044
				5:3:2	–.060	–.028	–.049
20	1–3	1–5	Different	2:3:5	–.106	–.160	–.198
				5:3:2	–.530	–.436	–.144
			Same	2:3:5	–.062	–.101	–.153
				5:3:2	–.258	–.235	–.083
	Three	Five	Different	2:3:5	–.007	.025	–.050
				5:3:2	.123	–.016	–.025
			Same	2:3:5	–.066	.078	–.024
				5:3:2	–.057	.189	.108
		1–5	Different	2:3:5	–.091	–.052	–.081
				5:3:2	–.171	–.014	–.106
			Same	2:3:5	–.072	–.017	–.068
				5:3:2	–.012	.022	–.004
		Five	Different	2:3:5	–.074	–.023	–.066
				5:3:2	–.110	–.025	–.050
			Same	2:3:5	–.061	–.011	–.060
				5:3:2	–.033	.047	–.022

The results are summarized across the Level 3 covariance value conditions because this design factor did not impact the results. The relative parameter bias values are bolded when the values are larger than 0.05. Variance Ratio = ratio of variances for sampling-error : between-outcome within-study : between-study variance values. ^a # Effect Sizes per Outcome = number of effect sizes per outcome and study.

(Hedges et al., 2010; Tipton, 2013) have focused on estimating a single average effect size across all outcomes, the present study includes dummy-coded indicator variables to distinguish each of three outcomes for which effect sizes are reported. We first demonstrated use and interpretation of the three-level meta-analysis model and RVE with a subset of real data taken from an actual meta-analysis.

In the real dataset there was a lack of balance, with unequal numbers of effect sizes within studies and per outcome within and across studies. We matched this authentic lack of balance in some of the conditions of our simulation. In the Results

section of the real-data analysis, we explained how parameter estimates can be interpreted for the three-level meta-analysis model and RVE by estimating outcome-specific effect sizes. As a final note about the real dataset, the overall size of the meta-analytic dataset (276 effect sizes from 100 studies) was sufficiently large that we did not encounter convergence issues for the outcome-specific mean effect sizes model that we estimated and that corresponded with the model we estimated in the simulation study.

In the simulation study, RVE did not have any convergence issues or biases in its estimates of the fixed-effect parameters

and their standard errors across conditions. However, under the three-level meta-analysis model, the convergence rate results were very problematic under the conditions examined here when the meta-analyses consisted of only 20 studies. Convergence rates were generally impacted by the size of the meta-analytic database (including the number of studies, the number of effect sizes per outcome, and the number of outcomes per study). In addition, when the ratio of the three levels' residuals' variances was 5:3:2, convergence rates were much worse than when this ratio was 2:3:5.

These convergence rate results match what would be expected on the basis of previous methodological studies that have focused on estimating multilevel models (see, e.g., Hox, & Maas, 2001; McNeish & Stapleton, 2016a, 2016b). This previous work has found that when too few clusters (which correspond to studies in meta-analysis) are included in the dataset being analyzed, when data are unbalanced in terms of the within-cluster sample size (in our study, the number of outcomes per study and the number of effect sizes per outcome) and when the true values for the variance of residuals at the higher levels of the model are close to zero then estimation will be challenging. As a recommendation founded in this previous work and in our own findings, if a meta-analyst wishes to synthesize results for a dataset with fewer than 20 studies and multiple outcomes of interest, use of the three-level meta-analysis model and REML estimation does not work well. Therefore, use of robust variance estimation (Hedges et al., 2010) for handling within-study dependence in effect sizes is recommended for stable convergence estimation.

Future methodological research could evaluate robustness of the use of a multilevel meta-analysis model that does not include as many random-effect terms. In the three-level meta-analysis model that we explored in this study, a three-by-three covariance matrix for Level 3 random effects was estimated (matching the generating model). However, while the true model was being estimated, estimation of a random-effect covariance matrix is already challenging, and as we have already noted, especially so when the true values of the covariance matrix elements are close to zero. For a meta-analysis, the scale of the outcome (the effect size) is small and the associated variances will be even smaller in scale. The true value of the Level-3 residuals' variances used in this study's simulation study was sometimes as low as .032, which makes it even harder to estimate with unbalanced data and few clusters (studies). Estimation of a three-level meta-analysis model that constrains (study) Level 3 residuals' variances to be equal across outcomes and only estimates a single random-effect variance component for Level 3 residuals will likely lead to better convergence rates. And even if the random-effect structure is misspecified, fixed-effect parameter estimates will likely be unbiased (Moeyaert et al., 2017). However, it is unclear whether standard error estimates, associated test statistics and Type I error rates will also be unaffected.

For converged solutions under the three-level meta-analysis model, we found no substantial bias in estimates of the fixed effects across all conditions. In the intercepts-only model that we estimated, this implies that no bias would be found in pooled estimates of the three outcomes' effect size estimates. In Van den Noortgate et al.'s (2013, 2014) multivariate multilevel model studies, the authors also found no bias in fixed-effect estimates when pooling estimates across as few as 30 studies. This then implies that even if the number of studies in a meta-analysis is as small as 20, and if the model's estimation converges, then the three-level meta-analysis model works well in terms of recovering true effect size values.

In terms of the estimation of standard error estimates for fixed-effect parameters, when the database consisted of 50 studies, there was rarely substantial standard error bias. In each of the $k = 50$ conditions in which substantial bias was found, the standard errors were under-estimated and the conditions were those with the smallest Level 3 residuals' variance values (the 5:3:2 variance ratio conditions). With too few clusters (like $k = 20$), it is not unexpected that fixed-effect parameters' standard errors would be underestimated. Use of the Kenward–Roger (1997; K-R) correction can improve the performance of associated statistical tests (McNeish & Wentzel, 2017), with smaller numbers of clusters at the highest level. As additional analyses, we also conducted a simulation study testing the K-R correction under the same conditions. Even though the K-R correction improved the standard error estimations slightly, we still found similar amounts and patterns of both convergence rates and bias in the standard errors estimated for fixed-effect parameters in scenarios with few studies being meta-analyzed. This preliminary investigation of the use of the K-R correction could be extended by exploring the performance of the K-R correction for models that incorrectly constrain elements in the random-effect covariance matrix to be zero.

The results from analyses of replication datasets that did not converge, were not included in the assessment of relative parameter and standard error bias. Given that the lack of convergence resulted from Level 3 variance component values that were estimated to be lower than zero, the converged solutions were expected to include positively biased estimates of Level 3 residuals variances. The results generally matched this expectation of positive bias for Level 3 variance component estimates. For the most part, Level 3 residuals' variance components were not biased in the $k = 50$ conditions. However, in the subset of conditions with the worst convergence problems ($k = 20$), substantial bias was found. And this bias was typically positive with overestimated Level 3 variances especially for the third outcome's effect size. The third outcome was the one for which the fewest datasets were available in unbalanced data conditions. Thus, it makes sense for estimation to be most challenging for parameters for the third outcome. And this was again reflected in the worst bias being found when the

number of outcomes reported per study and the number of effect sizes (per outcome and study) were unbalanced as opposed to scenarios when all three outcomes and five effect sizes per outcome were reported in each study in which much less bias was found (when $k = 20$).

Interestingly, the Level 3 residuals' covariance estimates tended to have negative bias even though the variances were positively biased when $k = 20$. Unlike the results for the corresponding variance components, the bias patterns were very similar whether the conditions included 20 or 50 studies. Again, for the most part the degree of balance in the dataset impacted estimation bias such that the more unbalanced the dataset, the more bias was found. And the variance ratios impacted bias with worse bias found in the 5:3:2 conditions than in the 2:3:5 conditions again linking a smaller true value for the Level 3 variance component with poor random-effect covariance matrix recovery. It is unclear why the covariances are negatively biased. Future research should explore the source of this bias and might consider whether the positive bias in the variance estimates helps mask the degree of bias in the standardized covariance (i.e., correlation) between pairs of random effects at Level 3.

In terms of confidence interval coverage rates for the fixed effects, the three-level meta-analysis model estimates were less accurate in scenarios with fewer studies, an unbalanced number of outcomes, and number of effect sizes per outcome and study. With fewer ($k = 20$) studies, varying numbers of outcomes and effect sizes per outcome, accuracy was lower than when the meta-analysis entailed more studies. These results match what would be expected in terms of the relationship between accuracy and the number of effects sizes in a meta-analytic database.

In conclusion, whereas RVE estimates outcome-specific pooled effect sizes and standard errors for fixed effects accurately, the random-effect variance component was underestimated under some conditions. On the other hand, estimation with the three-level meta-analysis model worked as well as RVE did when the number of studies was at least 50 for both fixed- and random-effect parameter estimation. However, problems were encountered with the multilevel meta-analysis model estimates when the variance ratio (for sampling-error : within-study between-outcome : between-study variances) was 5:3:2.

Even though in the present study we examined scenarios with the same number of studies per meta-analysis as had been used in Moeyaert et al.'s (2017) study, the results differed. In Moeyaert et al.'s study, the authors found that using the multilevel meta-analysis model to estimate a single overall effect size worked well. However, in our study we found that the multilevel meta-analysis model did not perform as well when estimating outcome-specific effect sizes. Therefore, if applied researchers intend to estimate pooled outcome-specific effect sizes but have a relatively small number of studies and

unbalanced studies' data, then the use of RVE is recommended. On the other hand, if enough studies and effect sizes are in the dataset and the applied meta-analyst is also interested in the variance and covariance component values, then use of the three-level meta-analysis with outcome-specific mean effect sizes works well, too. In terms of efficiency of estimates, if there are enough studies, the performance of three-level meta-analysis with outcome-specific mean effect sizes works better than RVE.

As with most research studies, some limitations should be delineated, which typically provides useful avenues for future research. The first limitation is that the results from this study are based on a restricted set of conditions. The results might have differed somewhat if we had used a different pattern of correlations matrix values with the effect sizes. We also only investigated one type of effect sizes—standardized mean differences. We also only evaluated estimation of the very simple, intercepts-only model, without including moderator variables in the model. In addition, we generated data assuming a normal distribution and using only restricted maximum likelihood estimation. Thus, we did not consider the robustness of the three-level meta-analysis model with data that violated that model's distributional assumptions and when the correct model was estimated. Also, when we generated an unbalanced number of outcomes in the data, we randomly deleted the effect size for each outcome with a probability of .5. However, in practical terms, the effect size will likely not be missing at random.

In addition, our three-level meta-analysis model was different from the one developed in previous three-level meta-analysis studies (Van den Noortgate et al., 2013, 2014), because we directly modeled the covariance between effect sizes for each outcome variables (Outcomes 1, 2, and 3) at the third level, by including dummy variables at the second level. Because the estimates might be different depending on the levels at which the outcome variables were distinguished, there is also a need to check whether estimation differs depending on the type of multilevel meta-analysis model estimated. Also, we compared two approaches (the three-level meta-analysis model and RVE) that both use a multivariate meta-analysis modeling framework within which separate but related effect sizes were pooled for each of multiple outcomes. Given many applied researchers' use of univariate meta-analysis models and pooling of effects separately by outcome, future research could compare the use of univariate versus multivariate meta-analysis models, to assess the robustness of using multiple univariate models rather than one multivariate model for estimating pooled effects for multiple outcomes.

Therefore, future research could consider more extensive conditions that were not included in this study, such as a larger number of outcomes per study, different

patterns of correlations among the effect sizes for outcomes, other types of effect sizes, and other types of models, including models with and without moderator variables. Also, to evaluate the robustness of three-level meta-analysis and RVE when estimating outcome-specific mean effect sizes, scenarios in which assumptions are violated and effect sizes are missing not completely at random could also be considered. It would also be very helpful to test whether estimation with the three-level meta-analysis model works well with fewer than 50 (but more than 20) studies, to help provide more specific guidelines for applied meta-analysts. Finally, a comparison among the various types of meta-analysis models for handling within-study dependence (e.g., separately modeling each outcome approach, GLS model, or three-level meta-analysis model estimating a single overall effect size) could provide a useful basis to understand differences in performance that depend on the choice of approaches for handling within-study dependence.

Appendix: SAS (Proc Mixed) and R package (robumeta) codes for the example dataset

Data set format

The dataset to be analyzed using RVE and the multilevel model includes one row for each observed effect size. The data as shown below called DATA1, contains the following variables:

ESID: an effect size indicator with values from 1 to 276,

StudyID: a study indicator with values from 1 to 60,

g : the observed effect size,

v : the estimated sampling variance for each observed effect size, and

O1, O2, and O3: three dummy-coded indicator variables, one for each of the three outcome-specific effect sizes (O1 refers to academic achievement, O2 is for academic behavior, and O3 refers to social/behavioral competence).

Nine of the rows for this data example (DATA1) are given below:

ESID	StudyID	g	v	O1	O2	O3
1	1	−1.507	.318	0	1	0
2	2	0.143	.219	0	0	1
3	3	0.643	.249	1	0	0
4	3	0.304	.247	1	0	0
5	3	0.611	.249	1	0	0
6	3	0.513	.248	1	0	0
...						
274	59	1.229	.408	1	0	0
275	60	1.195	.415	0	0	1
276	60	0.675	.394	0	0	1

For the random-effect three-level analysis estimating pooled effect sizes across three outcomes, the following SAS code is run that is revised on the basis of Hox, Moerbeek, and van de Schoot's (2018) example code:

```
PROC mixed DATA = DATA1 COVTEST ASYCORR
METHOD = REML;
```

```
CLASS StudyID ESID;
```

```
MODEL g = /SOLUTION DDFM = BETWITHIN
NOTEST;
```

```
RANDOM INTERCEPT/SUB = StudyID;
```

```
RANDOM INTERCEPT/ SUB = ESID (StudyID);
```

```
REPEATED/GROUP = ESID (StudyID);
```

```
PARMS (1) (1) (.318) (.219) (.249) (.247) (.249) (.238)
```

```
... (.408) (.415) (.394)/ EQCONS = 3 TO 278;
```

The SAS code for the three-level model analysis estimating pooled effect sizes across outcomes as shown above is run with the following commands:

Proc Mixed: mixed procedure for multilevel models,

Data: dataset is defined (here, DATA1),

METHOD: estimation method is defined (here, we use restricted maximum likelihood estimation, REML),

CLASS: statement to define classes in the dataset [here, effect size level (ESID) and study level (StudyID) are defined],

MODEL: defines the model (here, g is our dependent variable and moderator variables were not included because a single intercept is estimated by default),

SOLUTION: statement to request the parameter regression coefficient estimates and test in the output,

DDFM: statement specifies the estimation method for the denominator degrees of freedom (here, the "BETWITHIN" option is used),

WEIGHT: specifies the weight variable used,

RANDOM: specifies the random effects [here, between studies (SUB = StudyID) and between outcomes within studies, SUB = ESID(StudyID), were estimated],

REPEATED: specifies separate residual variances per study

PARMS . . . /EQCONS: can be used to fix variance and covariances to known values (here, the variances corresponding to each effect size estimate).

For using RVE, the “robumeta” package (Fisher & Tipton, 2014) can be used:

```
library(robumeta)
robu(formula = g ~ 1, data = DATA1, studynum = StudyID, var.eff.size = v, small = TRUE).
```

The R code for RVE in robumeta package as shown above is run with the following commands:

```
library(robumeta):Calls the already installed robumeta package,
robu: calls the robumeta function for RVE analysis,
formula : defines the statistical model (here,  $g \sim 1$  indicates that we are estimating the intercept-only model with dependent variable,  $g$ ),
data: identifies the dataset (here, DATA1),
studynum: identifies study level indicator (here, StudyID),
var.eff.size: identifies the sampling variance variable (here,  $v$ ), and
small = TRUE: use small-sample-size correction (Tipton, 2015).
```

For the random-effect three-level analysis estimating outcome-specific effect sizes, the following SAS code is run:

```
PROC mixed DATA = DATA1 COVTEST ASYCORR
METHOD = REML;
CLASS StudyID ESID;
MODEL g = O1 O2 O3/SOLUTION NOINT DDFM =
BETWITHIN NOTEST;
RANDOM O1 O2 O3/SUB = StudyID;
RANDOM INTERCEPT/ SUB = ESID(StudyID);
REPEATED/GROUP = ESID(StudyID);
PARMS (1) (1) (1) (1) (.318) (.219) (.249) (.249)
(.238)
. . . (.408) (.415) (.394)/ EQCONS = 5 TO 280;
RUN;
```

The SAS code for the three-level model analysis that estimates outcome-specific effect sizes as shown above is run. The primary difference is inclusion of the three dummy-coded indicators and specification of NOINT (no intercept).

To estimate this same model using RVE, the robu command can be adapted as follows:

```
robu(formula = g ~ -1 + O1 + O2 + O3, data = DATA1,
studynum = StudyID, var.eff.size = v, small = TRUE).
```

Note that as with the SAS code, we include the three dummy-coded indicators and “-1” to indicate no intercept should be estimated in this model.

References

- Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. *Review of Educational Research*, 85, 275–314.
- Assink, M., van der Put, C. E., Hoeve, M., de Vries, S. L. A., Stams, G. J. M., & Oort, F. J. (2015). Risk factors for persistent delinquent behavior among juveniles: A meta-analytic review. *Clinical Psychology Review*, 42, 47–61.
- 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.
- Becker, B. J., Hedges, L. V., & Pigott, T. D. (2004). *Campbell Collaboration statistical analysis policy brief* (Campbell Collaboration resource document). Available from http://www.campbellcollaboration.org/ECG/policy_stat.asp
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Hoboken, NJ: Wiley.
- Cheung, M. W.-L. (2015). *Meta-analysis: A structural equation modeling approach*. Chichester, UK: Wiley.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis*. New York, NY: Russell Sage Foundation.
- Fisher, Z., & Tipton, E. (2014). robumeta: An R-package for robust variance estimation in meta-analysis. Working paper, Department of Statistics, Northwestern University, Evanston, IL.
- Geeraert, L., Van den Noortgate, W., Grietens, H., & Onghena, P. (2004). The effects of early prevention programs for families with young children at risk for physical child abuse and neglect: A meta-analysis. *Child Maltreatment*, 9, 277–291.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5, 3–8.
- Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 357–376). New York, NY: Russell Sage Foundation.
- Hedges, L. V. (1981). Distribution theory for Glass’s estimator of effect size and related estimators. *Journal of Educational Statistics*, 6, 107–128.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation of meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1, 39–65.
- Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, 3, 486–504. doi:<https://doi.org/10.1037/1082-989X.3.4.486>
- Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling. *Sociological Methods and Research*, 26, 329–367.
- Hox, J. J., & Maas, C. J. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling*, 8, 157–174.
- Hox, J. J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications* (3rd ed.). New York, NY: Routledge.
- Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage.

- Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53, 983–997.
- Kisamore, J. L., & Brannick, M. T. (2008). An illustration of the consequences of meta-analysis model choice. *Organizational Research Methods*, 11, 35–53.
- Kreft, I. G. G., & De Leeuw, J. (1998). Introducing multilevel modeling. Newbury Park, CA: Sage.
- Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, 1, 86–92. doi:<https://doi.org/10.1027/1614-2241.1.3.86>
- Matt, G. E., & Cook, T. D. (2009). Threats to the validity of generalized inferences. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 538–557). New York, NY: Sage.
- McNeish, D. M., & Stapleton, L. M. (2016a). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, 28.
- McNeish, D. M., & Stapleton, L. M. (2016b). Modeling clustered data with very few clusters. *Multivariate Behavioral Research*, 51, 495–518.
- McNeish, D., & Wentzel, K. R. (2017). Accommodating small sample sizes in three-level models when the third level is incidental. *Multivariate Behavioral Research*, 52, 200–215.
- 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.
- National Research Council. (1992). *Combining information: Statistical issues and opportunities for research*. Washington, DC: National Academy of Science Press.
- Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 301–322). New York, NY: Russell Sage Foundation.
- Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin*, 103, 111–120. doi:<https://doi.org/10.1037/0033-2909.103.1.111>
- Schmidt, F. L., Oh, I., & Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology*, 62, 97–128.
- Sheridan, S. M., Kim, E. M., Beretvas, S. N., Smith, T., & Park, S. Y. (2017). *Family-school interventions and academic and social behavioral outcomes: What matters?* Paper presented at the annual meeting of American Psychological Association, Washington, DC.
- Tanner-Smith, E. E., & 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. doi:<https://doi.org/10.1002/jrsm.1091>
- Tipton, E. (2013). Robust variance estimation in meta-regression for binary dependent outcomes. *Research Synthesis Methods*, 4(2), 169–187. <https://doi.org/10.1002/jrsm.1070>
- Tipton, E. (2015). Small-sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, 20(3), 375–393. doi:<https://doi.org/10.1037/met0000011>
- Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analyses of dependent effect sizes. *Behavior Research Methods*, 45, 576–594. doi:<https://doi.org/10.3758/s13428-012-0261-6>
- Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2014). 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, 765–79.
- Van der Kleij, F. M., Feskens, R. C., & Eggen, T. J. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research*, 85, 475–511.
- Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, 30, 261–293.